Audiovisual Processing For Sports-Video Summarisation Technology

by

David A. Sadlier B.E., M.E.

Submitted in fulfilment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Supervised by

Dr Noel O'Connor

School of Electronic Engineering
Dublin City University
Dublin, Ireland

February 2006

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed:..... Date: 17/2/06

David Anthony Sadlier

ID No: 96410442

Tiomnú

Ba mhaith liom an trachtas seo a thiomnú do mo thuismitheoiri, George agus Stephanie Sadlier, agus do mo bhean cheile Muirne, ceadsearc mo shaoil, murach iad siud ni fheadfainn an saothar seo a thabhairt chun criche

Acknowledgements

I would like to address special thanks to my supervisor Dr Noel O'Connor for his extensive guidance and commitment to this project Thanks also to Prof Alan Smeaton, Dr Sean Marlow, and Dr Noel Murphy, for their input and direction, and to all my friends/colleagues who make the Centre for Digital Video Processing the inspiring work environment it is

Table of Publications

The following are the journal/conference articles previously published by this author in support of this work

1 D A Sadlier, N E O'Connor, "Event Detection in Field Sports Video Using Audio-Visual Features and a Support Vector Machine" in ***IEEE Transactions on Circuits and Systems for Video Technology***, (eds F Pereira, P van Beek, A C Kot, J Ostermann), pp 1225-1233, Volume 15, Number 10, October 2005

2 D A Sadlier, N E O'Connor, "Event Detection Based On Generic Characteristics Of Field Sports," proc ***IEEE International Conference on Multimedia and Expo (ICME 2005)***, pp 759-762, Amsterdam, The Netherlands, 6-8 July 2005

3 D A Sadlier, N O'Connor, N Murphy, S Marlow, "A Framework for Event Detection in Field-Sports Video Broadcasts Based On SVM Generated Audio-Visual Feature Model Case-Study Soccer Video," proc ***11th International Workshop on Systems, Signals and Image Processing (IWSSIP'04)***, pp 243-246, Poznan, Poland, 13-15 September 2004

4 D A Sadlier, N O'Connor, S Marlow, N Murphy, "A Combined Audio-Visual Contribution to Event Detection in Field Sports Broadcast Video Case study Gaelic Football," proc ***3rd IEEE International Symposium on Signal Processing and Information Technology (ISSPIT'03)***, pp 552-555, Darmstadt, Germany, December 14-17, 2003

5 S Marlow, D Sadlier, N O'Connor, N Murphy, "Voice Processing For Automatic TV Sports Program Highlights Detection," proc ***8th International Symposium on Social Communication***, Santiago de Cuba, Cuba, 20 -24 January 2003

6 D A Sadlier, S Marlow, N O'Connor and N Murphy, "MPEG Audio Bitstream Processing Towards the Automatic Generation of Sports Programme Summaries," proc ***IEEE International Conference on Multimedia and Expo (ICME'02)***, pp 77-80, Lausanne, Switzerland, 2002

7 S Marlow, D Sadlier, N O'Connor, N Murphy, "Audio Processing For Automatic TV Sports Program Highlights Detection," proc ***Irish Signals and Systems Conference (ISSC 2002)***, Cork, Ireland, 25-26 June 2002

Abstract

In this thesis a novel audiovisual feature-based scheme is proposed for the automatic summarization of sports-video content The scope of operability of the scheme is designed to encompass the wide variety of sports genres that come under the description 'field-sports' Given the assumption that, in terms of conveying the narrative of a field-sports-video, score-update events constitute the most significant moments, it is proposed that their detection should thus yield a favourable summarisation solution To this end, a generic methodology is proposed for the automatic identification of score-update events in field-sports-video content The scheme is based on the development of robust extractors for a set of critical features, which are shown to reliably indicate their locations The evidence gathered by the feature extractors is combined and analysed using a Support Vector Machine (SVM), which performs the event detection process An SVM is chosen on the basis that its underlying technology represents an implementation of the latest generation of machine learning algorithms, based on the recent advances in statistical learning Effectively, an SVM offers a solution to optimising the classification performance of a decision hypothesis, inferred from a given set of training data Via a learning phase that utilizes a 90-hour field-sports-video training-corpus, the SVM infers a score-update event model by observing patterns in the extracted feature evidence Using a similar but distinct 90-hour evaluation corpus, the effectiveness of this model is then tested generically across multiple genres of field-sports-video including soccer, rugby, field hockey, hurling, and Gaelic football The results suggest that in terms of the summarization task, both high event retrieval and content rejection statistics are achievable

Table of Contents

PREFACES:

CHAPTERS:

APPENDICES:

List of Figures

List of Tables

Table of Acronyms

AC	Alternating Current (used to indicate a non-zero frequency signal)
ADPCM	Adaptive Differential Pulse Code Modulation
API	Application Programming Interface
B(b)	Bi-directionally predicted
CF(s)	Critical Feature(s)
CIC	Crowd Image Confidence
CIF	Common Intermediate Format
CRR	Content Rejection Ratio
CSM	Cosine Similarity Measure
CuC	Close-Up Confidence
DACSM	Dissimilarity Analogue of the Cosine Similarity Measure
DC	Direct Current (used to indicate a zero-frequency signal)
DCT	Discrete Cosine Transform
DHPR	Dominant Hue Pixel Ratio
DPCM	Differential Pulse Code Modulation
DVE(s)	Digital Video Effect(s)
EPR(s)	Edge Pixel Ratio(s)
ERM	Empirical Risk Minimization
FOS	Figure of Significance
FPC(s)	Field-Pixel Candidate(s)
FSV	Field-Sports-Video
GOP	Group of Pictures
HT	Hough Transform
HLT	Hough Line Transform

I(i)	Intra
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
JPEG	Joint Picture Experts Group (typically used to denote the ISO/IEC image compression standard developed by this group)
KKTC	Karush-Kuhn-Tucker Conditions
KNN	K- Nearest Neighbour
LM	Learning Machine
LPC	Linear Predictive Coding
MFCC(s)	Mel Frequency Cepstral Coefficient(s)
MPEG	Motion Picture Experts Group (typically used to denote the ISO/IEC video compression standards developed by this group)
MV(s)	Motion Vector(s)
MVM	Mode-Variance Measure
NFVAM(s)	Near-Field Visual Activity Measure(s)
NN	Neural Network
NTP(s)	Negative Training Point(s)
NZMVC	Non-Zero Motion Vector Count
OCR	Optical Character Recogniser
P(p)	Predicted
PC	Personal Computer
PSB(s)	Potential Scoreboard Block(s)
PTP(s)	Positive Training Point(s)
RFPC(s)	Refined Field Pixel Candidate(s)
RLE	Run Length Encoding
ROE	Regions of Expectancy
ROI	Regions of Interest

RP	Reaction-Phase
RPSW	Reaction-Phase Seek Window
RSB(s)	Recognized Scoreboard Block(s)
RSBP(s)	Recognised Scoreboard Block Pixel(s)
SBAL(s)	Speech-Band Audio Level(s)
SEB	Shot-End Boundary
SHPR	Skin Hue Pixel Ratio
SRM	Structural Risk Minimization
SUE(s)	Score-Update Episode(s)
SUERR	Score-Update Episode Retention Ratio
SVM	Support Vector Machine
TP(s)	Training Point(s)
TV	Television
WWW	World Wide Web

Chapter 1

Introduction

Entitled *Audiovisual Processing for Sports-Video Summarization Technology*, this thesis describes a generic methodology for the automatic summarisation of field-sports video content, based on the detection of the most significant events constituting such Prefaced by a foreword concerning relevant background information, this chapter presents an introduction to the topic of video summarization, with particular emphasis on that related to sports-video Following this, a detailed description of the research objectives targeted in this thesis is provided, which is supplemented by an overview of the proposed realisation approach

1.1. Background

1.1.1. The Digital Video Era

Increasingly, more and more personal video material is being captured, shared and archived worldwide [1] Ostensibly, the catalysts for such developments include (i) the manifestation of video in the digital domain, (ii) the emergence of the World Wide Web (WWW), and (iii) the accessibility of relatively inexpensive, sizeable data storage hardware to the consumer However, it is evident that, in addition, the recent advances in digital video compression technologies (e g the suite of ISO/IEC MPEG video coding standards) have also played a major role in driving the acceleration of video-related activity

1 1.2. Video Modelling

The challenge of video modeling corresponds to developing mathematical representations of video structure and/or semantic concepts The recent escalation in video-related activity serves to greatly illuminate the deficiency in video motleling solutions In response to this, there currently exists an abundance of research projects worldwide that aim to provide solutions to many of the aspects of the video-modeling question For instance, much attention has recently been paid to the task of making the WWW more searchable for multimedia content, exemplified by the ISO/IEC standard MPEG-7 The MPEG-7 standard aims to tackle the issue by offering a comprehensive set of low-level audiovisual description tools in creating descriptors, which form the basis for search, filtering and browsing tasks However, experiential evidence suggests that users of content collections prefer to query video content at the conceptual or semantic level rather than at a feature level [2] - hence the issue of the 'semantic gap' in video processing [3] The ***semantic gap*** is a multimedia retrieval-based concept that relates to the virtual gap between the rich meaning that a user desires, and the shallowness of the multi-modal description features that may be automatically extracted from the content It is a commonly held principle among the research community, that one of the most pressing aspects of the video modeling question concerns this issue of extending the nature of user interaction with multimedia content towards real semantics That is, bridging the semantic gap may be seen as the fundamental challenge to be overcome in the development of most real-world video modeling applications

1.2. Video Summarization

1.2.1. Overview

Video summarization corresponds to the process whereby given a quantity of video, its magnitude and/or its playback duration time is reduced, such that its underlying narrative (*i e the unfolding of events*) may be conveyed in an abstracted form Providing for the downscaling of video content in this way effectively corresponds to a trade-off between the level of abstraction desired, and the coherency of the narrative to be maintained

Clearly, the challenge of unsupervised (automated) video summarization is an instance of a video modeling application where the issue of the semantic gap is pertinent That is, given a quantity of video for summarization, the relevant question is,

by what hypothesis can the narrative be synopsized, given a feature-based description of the content

1.2.2. Motivation

Driven by the perceived consumer marketability of projected applications, the development of automated video summarization technology is currently a hot topic in the area of digital video analysis and is presently receiving a lot of attention from the research community The popularity of this topic may be considered to be in direct response to user-driven demands that, as alluded to above, have stemmed from the escalation in video-related activity characterizing the digital video era

As discussed, modern developments in digital video compression technologies have paved the way for extensive archiving of video content However, this increase in content availability has not necessarily resulted in an increase in the ease of user accessibility That is, there are certain practical factors that impede the development of many user-orientated video content distribution applications These issues are most apparent when considering a mobile (wireless) delivery scenario, i e where a hand-held device is required to receive an encoded video-bitstream transmission (e g an MPEG bitstream), decode it, and then display (playback) the content Such an application is largely hampered by two main factors (i) the limited bandwidth of the transmission channel, and (ii) the limited battery life of the device Although higher wireless bandwidth standards such as '3G' are imminent, live streaming of a complete video may still be impracticable or even unaffordable to a given user That is, in the mobile domain, bandwidth remains a valuable and costly commodity both for the service provider and for the consumer Elsewhere, research continues in the field of power efficient hardware solutions for video applications Many of these works concentrate on designing new methods of implementing the power hungry algorithms that tend to be required by many video-bitstream decoders For example, Kinane *et al* [4] propose an energy efficient hardware architecture approach for the Discrete Cosine Transform (a key component in MPEG video encoding) A general discourse on the overall topic of power efficient hardware solutions for video applications may be found in [5] The substantial activity in this research field serves to suggest that the significant shortcomings in the power capabilities of hand-held devices in relation to video applications have yet to be resolved to satisfactory levels Overall, these two particular constraints render impractical the extended high-quality decoding and playback of

videos in their entirety in the mobile scenario, and hence suggest an increasingly crucial role for the accelerated presentation or summarization of content for the development of such applications

Given the accessibility of wired broadband connectivity and on-demand power, fixed-line environment applications (e g TV and/or desktop PC) tend not to be so restricted when dealing with video content delivery/playback However, the era of satellite television broadcasting has served to substantially increase the number of video events being broadcast or made available Thus, it is often not possible for even the most avidly engaged user to watch more than a small fraction of the available coverage of complete events Therefore, even in the favourable circumstances of fixed-line platforms, automatic summarisation of content should still play a vital role in improving the efficiency of video browsing, thus reducing the time consumed, and hence cost involved, in viewing the ever-increasing proliferation of available content

There exist a variety of proposed methodologies for video summarization technology in the literature These may be broadly classed into two categories, i e those corresponding to the accelerated presentation of content (basic summarization), and those involving the detection of critical events (highlighting)

1 2 3. Accelerated Presentation (Basic Summarization)

The accelerated presentation of video content is concerned with representing the video narrative in a more succinct form, by varying (i e increasing) the traversal speed via which the content may be viewed This is also known as 'video skimming' and represents a well-known basic approach to the task of summarizing the contents of a video There are a variety of ways in which this task may be tackled, the most basic of which concerns sampling the video stream at regular fixed intervals A more sophisticated approach, such as that advocated in [6], concerns the detection of representative 'key-frames', the presentation of which tends to be more accurate and reliable in conveying the narrative of a video However, while such approaches typically contribute positively in terms of their respective tasks, the methodology of basic accelerated presentation falls short in constituting an optimal solution for the extraction of the narrative from video content since, while their generated outputs do correspond to a more terse representation of the input video, they typically still convey information that may be considered redundant Overall, in terms of generating sufficiently condensed output, unless the development of highly sophisticated skimming methods

are targeted, such as that developed by Chang *et al* [7] (in which audiovisual methods are proposed for determining localized temporal content significance and skimming on that basis), we are motivated towards a more sophisticated highlight-detection orientated approach to the problem

1.2.4 Event Detection (Highlighting)

The event detection-based video summarization methodology concerns the development of hypotheses for automatically determining which phases of the content are most critical to the narrative (highlights), and by the same token, which may be considered redundant If the most significant events may be reliably detected, they may be then extracted, concatenated, and packaged in chronological order, such that a narrative-only summarized version of the input video is generated Furthermore, if desired, an event-only summary could be then presented in an accelerated manner using one or more of the methods described in the previous section However, based on an observation of the relevant literature, it is clear that the detection of narrative-critical events in video sequences is considered a challenging task One of the troublesome aspects is that in many scenarios, the events are subjective, i e their interpretation varies from user to user On the other hand, it seems to be commonly accepted that this difficulty may be alleviated somewhat if the nature of the content is limited to a specified domain That is, in circumstances where the nature of the content is known (e g sports, news, movies, etc), the narrative-critical events may become more objectively defined Furthermore, given a set of specified events for a constrained scenario, the features intrinsically characterizing the particular domain may be exploited, thereby aiding the event detection process Given this, the approaches to event detection-based summarization may be classified into two broad categories, (i) general approaches for situations where the nature of the content is unknown (generic video scenario), and (ii) more specific approaches for when the nature of the content is constrained (restricted domain scenario)

1 2 4 1 Generic Video Scenario

In the generic video scenario, no assumptions may be made about the exact nature of the content, and therefore the events of interest may not be specified in advance Furthermore, there tends to be no scope for the exploitation of domain particular characteristics Nonetheless, it was observed from the various approaches found in the

literature, that the typical approach for the task of generic event detection-based video summarization, is to model the significant events as those constituting the most conspicuously effervescent moments For example, this is the methodology undertaken by Hanjalic in [8] and by Lienhart in [1], whereby it is proposed that given a quantity of generic video, the narrative-critical events may be implied from the content on the basis of modelling video excitement In general, it is acknowledged that the exploitation of the following criteria is useful,

(i) accelerated motion activity
(ii) video luminance dynamics
(iii) increased audio energy
(iv) high shot-cut rate

The prospect that a summary generated from these criteria will convey a reliable account of the narrative is clearly rooted in the nature of the content, i e on the extent of the correlation between the narrative-critical events and the excitation in the audiovisual signals Nonetheless, it has been shown in the works mentioned, that event detection via the excitement modeling approach provides for a reasonable contribution to the realization of the task of video abstraction in circumstances where the content domain is unknown

1 2 4 2 Restricted Domain Scenario

In the restricted domain scenario, the narrative-critical events have the potential of becoming more objectively defined Given this, a more specific event hypothesis may be invoked, compared to that of the generic case described above Furthermore, the limitation of domain scope has additional benefits in relation to the actual event detection task That is, since each distinct video domain exhibits particular structural and broadcast rules, given a well-defined event concept, the domain specific characteristics may be exploited in developing robust event identification heuristics

Sports, news, and movie-video are examples of restricted domain scenarios that typically exhibit the domain-constrained advantages as described Hence the profusion of related works in the literature For example, in both [9] and [10], methodologies are proposed for the extraction of events from news-video content Therein, the authors advocate a story-based event detection solution, realised by exploiting the intrinsic domain-particular characteristics of such content Furthermore,

towards synopsising movie-video content, Lehane *et al* propose techniques for both dialogue-event detection [11] and action-event detection [12], which are based on observed film syntax conventions Likewise, numerous approaches for sports-video summarisation solutions have been observed in the literature The details of these will be expounded in a subsequent chapter However, on the surface, it is apparent that as instances of restricted domains, sports-videos arguably represent the most conducive context for event detection-based summarization This is explained further in the following section

1.3. Sports-Video Summarization

1.3.1. Amenability of Sports-Video to Summarization

The popularity of sports-video as a summarization domain stems from the anticipation of successful outcomes This expectation is primarily due to the fact that, as instances of restricted domains, sports-videos tend to be of substantial duration with few exciting moments That is, as a rule, the general structure of sports-video may be considered as a dynamic interleaving of inconsequential periods and significant episodes, where the former tend to constitute the greater part Furthermore, in such content, the majority of the significant episodes are typically well defined within their particular genres, e g (i) score-update events in soccer games, (ii) start/finish and overtake manoeuvres in athletics races, (iii) start/finish, overtakes, and crashes in motor races, (iv) knock-down and 'on the ropes' moments in boxing matches, etc It is arguable that such episodes alone constitute the moments that are most significant to the narratives of their respective games (i e the narrative-critical events), and these examples illustrate how relatively objective the concepts can be for sports-video content Given this, it is a commonly held argument that in their capacity as restricted domains, sports-videos tend to be innately conducive to event detection-based summarization

It is also recognized that every sports genre is characterized by a strict set of rules that apply to its underlying game A consequence of this is that the broadcast conventions in sport-video tend to be constrained to a larger degree than in other restricted domain scenarios, such as news or movies This phenomenon renders sports-video exceptionally conducive to heuristic orientated modeling That is, given a particular event concept pertaining to one or more sports genres, the unusually constrained broadcast formats serve to aid the prospect of the accurate detection of

such within the content

1.3.2. Approach Methodologies For Sports-Video Summarization

From the literature, it is evident that the existing approaches to sports-video summarization can be broadly classified into two distinct categories, i e genre-specific and genre-independent methodologies An explanation for this, and a description of the underlying principles of each approach follows

1 3 2 1 Genre-Specific Methodologies

Due to the dramatic variances in broadcast styles for different sports genres, and given the advantages offered by maximizing the domain constraints, many of the existing approaches to sports-video summarization adopt a genre-specific methodology The particulars of these works will be expounded in *Chapter 2*, however, it is observed that overall, given their objectives, many report accurate and reliable performances via this approach However, given that they are orientated towards a specific domain, central to most schemes are typically non-recyclable algorithms based on intrinsically characteristic critical features that are peculiar to the sports genre in question That is, generality tends to be sacrificed for the sake of optimized performance accuracy Hence, the drawback of these schemes is that blanket execution of the obtained solutions across multiple sports genres is generally not viable This shortcoming serves to somewhat lessen their impact in the field Recognizing this as a significant disadvantage, the research community has recently been led to focus on more generic methodologies to the summarization task

1 3 2 2 Genre-Independent Methodologies

As will be shown in detail in *Chapter 2*, the recent shift towards more genre-independent approaches to sports-video summarization is reflected in the more contemporary research literature output, where the challenge is to attempt to overcome the multi-genre inapplicability limitation in a more genre-independent approach to event detection-based summarization in sports-video The realization of such a task thus relies on the development of hypotheses that can reveal the common structures of multiple events across multiple sports genres While many generic schemes do exist in the literature (see *Chapter 2*), most are only evaluated across a narrow genre scope Furthermore, as will be discussed in *Chapter 2*, in many cases the solutions have been

developed such that they only work across a small set of (ostensibly hand-picked) sports genres, the link between which would not necessarily be made in another context Clearly, the ultimate generic solution would be that which has the potential to provide consistently reliable results given any input sports-video genre However, it is recognized that the pursuit of a 'one-size-fits-all' solution in developing a generic approach is impractical This can be explained as follows, consider a tennis-video scenario Within the scope of this restricted domain, it is arguable that the narrative-critical events correspond to those episodes associated with scoreboard updates Thus, in terms of the summarization task, the event concept may be defined accordingly However, considering another sports genre, e g boxing video, the former event concept (i e a scoreboard update episode) does not hold As a result, the event concept breaks down, and therefore cannot be applied generically across both genres Therein lies the crux of the problem for the development of a genre-independent approach to sports-video summarization - a conflict exists between the definition of the event concept, and the required provision for generic applicability It is concluded that for the development of a practical genre-independent solution to sports-video summarization, this conflict must be somehow addressed in order that robust generically functional solutions might be attained

1.4. A Proposed Compromise Methodology

As explained above, the principle difficulty pertaining to the development of a genre-independent solution to sports-video summarization concerns the conflict that exists between the event concept definition, and the required provision for generic applicability The above example used to illustrate this is an extreme case involving two sports genres that differ vastly in game format and video structure characteristics Nevertheless, it serves to highlight the fact that it is unfeasible to suggest that there exists a unique solution for the event detection-based summarization task that will operate successfully across all genres of sports-video However, conceding this, it may be argued that a subsequent problem is deserving of investigation, i e how feasible is it to propose that certain sports genres do in fact exhibit similar characteristics and therefore, in the context of the summarization task, may be grouped together and treated as one entity? That is, is it possible that sports-video subsets may be delineated, throughout which, the definition of event concepts may be robustly sustained? Given

the concession that no unique 'one-size-fits-all' solution exists in terms of developing a generic approach for the task, it is considered desirable to ascertain whether or not this compromised approach may be shown to realize successful outcomes

Such compromised generality concerns the assumption that there exists a solution that can reveal the common structures of multiple events across multiple sports genres, the characteristics of which are consistent to those indicative of a predetermined sports-video subset This author proposes that such an assumption holds, and that this represents a feasible solution to the problem of developing a generic methodology for sports-video summarization Given this, an approach for the realization of this hypothesis is proposed, based on setting a meaningful boundary on the generality attribute, via the introduction of the concept of the sports-video supergenre

A ***supergenre*** is defined as a limited collective of characteristically similar sports genres in a single class Given the aforementioned arguments, in characterizing a supergenre, it is desired to limit the domain scope to the extent that similar genres may be automatically summarized en masse, while simultaneously avoiding a situation where the heuristics become excessively biased towards one genre in particular Hence, in terms of the event detection-based summarization task, it is aimed to push the multi-genre operability envelope, while simultaneously maintaining robustness in the definition of event concepts A listing of suggested supergenres and their constituents is presented in **Table 1 1** Examples include racquet-sports, motor-sports, field-sports, etc It is proposed that if supergenre solutions may be generated, which operate with consistent performance across each of their respective sports genre constituents, this represents a valuable quasi-generic solution to the problem of genre-independent sports-video summarization

1.5. Research Objective & Realization Approach

In this section the research objective of this thesis is explicitly stated This is then followed by a description of the proposed approach to be undertaken, via which it is anticipated this objective may be realised

1.5.1. Target Case Study: Field-Sports-Video

The requirement of a genre-independent solution to the problem of sports-video summarization represents the primary motivation for the work undertaken in this thesis

Table 1 1 Proposed supergenres and their constituents

Supergenre	Constituent Genres
Racquet Sports	Tennis, Badminton, Table Tennis, Squash
Motor Sports	Formula-1, Superbikes, Speedway
Target Sports	Archery, Darts, Rifling
Ring Sports	Boxing, Wrestling, Martial Arts
Arena Sports	Baseball, Cricket, Rounders
Court Sports	Basketball, Volleyball, Netball, Handball
Field Sports	Soccer, Rugby, Hurling, American Football, Australian Rules Football, Gaelic Football, Field Hockey

To this end, the relevant issues of motivation, background, and potential difficulties, have been thus far described Initially, the real world issues providing the motivation for the development of automatic video summarization technology were outlined It was then explained why event detection-based approaches yield the most favourable solutions Following this, the advantages offered by constrained domain scenarios to the summarization task were described Above all, it was outlined how sports-videos, as instances of restricted domains, are particularly suitable Next, it was described why genre-independent approaches to sports-video summarization have recently become more preferable to those concerning specifically targeted genres

In light of the obstacles discussed that challenge the development of generic solutions, a compromised approach was proposed in *Section 1 4*, which involves outlining subsets of the overall sports-video domain (i e supergenres) throughout which both the event concepts and general aspects of the games might be said to be consistent For a given supergenre, the challenge is to develop a generic solution that can yield consistent performances across its constituent genres The ideal target solution is that which yields accuracy comparable to that offered by the individual genre-targeted approaches

On the basis of this proposed approach, the specific research objective of this thesis concerns addressing this challenge for a chosen supergenre, i e ***field-sports-video*** **(FSV)** That is, the specific task is to develop a generic solution for event detection-based summarization in field-sports video, whereby the attained solution

provides consistent performances across the various sports genres that constitute this supergenre (see Table 1 1) Furthermore, the performances should exhibit accuracy that rivals that of the genre-specific equivalent solutions Emphasis is focused on this supergenre in particular on the basis that it is ostensibly the most populated, and its constituents represent some of the most conventionally popular sports genres Assuming such a solution could be arrived at, it would represent a significant improvement on the existing state of the art since it would put a meaningful boundary on a generic solution to sports-video summarisation

1 5.2. The Proposed Realisation Approach

Given the research objective as described above, this section aims to outline the proposed approach to be undertaken, via which such might be successfully realised

1 5 2 1 Field-Sports-Video Data Corpus

Towards providing a platform from which observations and suppositions in regards to the solution development may be drawn and tested, over 180-hours of FSV content was captured from broadcast television, comprised of genres including rugby, field hockey, hurling, soccer, and Gaelic football To ensure generality, the content was obtained from a wide variety of TV network sources Video images were captured at CIF resolution (352 pixels wide * 288 pixels high), at a framerate of 25 frames per second, and audio data was captured in 128kbits/sec stereo, with sampling frequency of 44100 samples per second, per channel The entire corpus was compressed and archived according to the MPEG-1 digital video standard **Fig 1 1** illustrates the relative proportions of each genre within the overall corpus **Table 1 2** provides details of the average broadcast durations of each genre (note each captured broadcast included a halftime interval and typically some quantity of added time) While, no American football or Australian Rules football content was captured (see Table 1 1), it was recognized that the five genres represented nonetheless provide a good diversity of field-sport games

1 5 2 2 Field-Sports Supergenre Characterisation

Given this data corpus, the next requirement should be to determine and specify exactly what is meant by the field sport description Once finalized, the solution developed should be then applicable to any sport that satisfactorily fits this description To this end, it is proposed that the five genres constituting the data corpus be analyzed towards

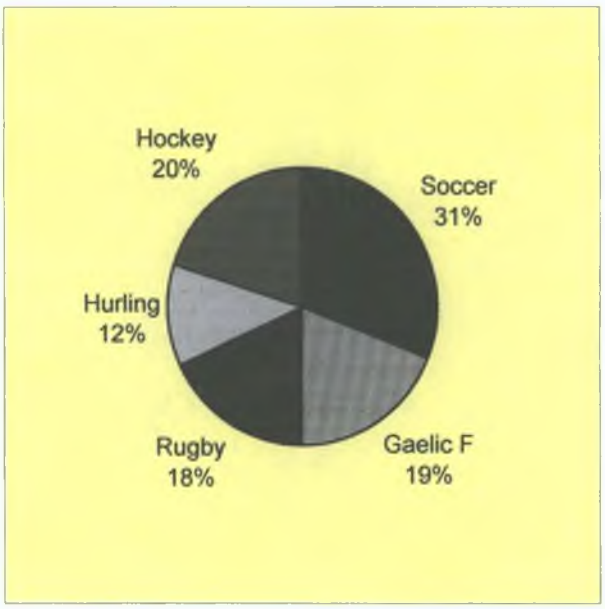


Fig. 1.1. The relative proportions of the individual sports genres constituting the FSV experimental corpus.

Table. 1.2. Average broadcast durations of the sports genres constituting the FSV experimental corpus.

FSV Genre	Average Broadcast Durations
Rugby	100 minutes
Field Hockey	87 minutes
Hurling	88 minutes
Soccer	109 minutes
Gaelic Football	91 minutes

determining exactly what the specific characteristics are that link them under the banner 'field-sports'. In terms of realising the research objective of *Section 1.5.1*, it is proposed that, once finalized, these common characteristics should then form the necessary criteria for defining the bounds of operability of the solution. That is, they should define the bounds of the supergenre, within which the solution should work with consistency across any sports-genre that exhibits them.

1.5.2.3. Narrative-Critical Events

It is recognized that the score tally is an aspect that is fundamental to the concept of all field-sports. In fact, it is arguable that above all, the dynamics of score count represent to a large extent the most interesting developments (i.e. the narrative-critical events) of

the underlying games This argument is founded on the basis that within field-sport games, it is the accumulation of scores (e g *goals* in soccer or hockey, *tries* and/or *conversions* in rugby, *points* in Gaelic football, etc) that dictates the overall outcome (i e the winners and losers of the contest) On this basis it is proposed that in terms of the summarization task for the field-sport supergenre as a whole, the detection of these ***score-update episode*** **(SUE)** highlights should provide for game summaries that have a satisfactory level of extracted narrative That is, while it is acknowledged that it is not uncommon for other non-SUE episodes to occur that may also exhibit a user interest level (e g near misses, controversy, important player substitutions, etc), these are not specifically proposed for targeting in terms of the summarization, on the basis that their level of interest tends to be more subjective compared to that of SUEs, which are recognized as being more objectively critical to the narrative

1 5 2 4 Score-Update Episode Characterization

Towards modeling SUEs, it is proposed that multiple incidences of such be surveyed from the data corpus towards determining what features (if any) may be said to consistently characterize them across the five FSV genres to hand If such a set of critical features could be found, it is anticipated that a quantification of the prevalence/intensity of these within appropriate temporal boundaries should then provide a reliable basis for SUE detection

1 5 2 5 Supervised Learning Approach

To preserve the scientific integrity of any experimental analysis, it is always desirable to base the system development on one set of data, and then evaluate the learned hypothesis on another distinct dataset Hence, the overall 180hr corpus was divided into two 90hr sub-corpuses, one for hypothesis development (i e the ***training-corpus***) and another for use in the experimental phase (i e the ***test-corpus***) Note that the relative proportions of the five test-corpus genre were preserved in the division procedure, and following a manual investigation, the SUE distributions within the two separate corpuses were determined (presented in **Tables 1 3** and **1 4** respectively) From these tables it is evident that, in terms of SUE occurrence, the two distinct datasets are reasonably balanced (i e the training-corpus contains 883 SUEs and the test-corpus contains 850 SUEs)

Given this, towards the development of a SUE-shot prototype, it is proposed

Table 1.3. Breakdown of training-corpus SUEs.

Training Corpus Genre	# SUEs	Description
Soccer	67	Goals
Hurling	227	goals, points
Rugby	169	tries, placed kicks, goals
Gaelic Football	365	goals, points
Hockey	55	Goals
Total	883	

Table 1.4. Breakdown of test-corpus SUEs.

Test Corpus Genre	# SUEs	Description
Soccer	56	Goals
Hurling	245	goals, points
Rugby	167	tries, placed kicks, goals
Gaelic Football	334	goals, points
Hockey	48	Goals
Total	850	

that a supervised learning approach be undertaken, i.e. train and learn from the training-corpus, then using the learned model, evaluate on the test-corpus. Given the substantial size of the dataset (i.e. 180hrs, which is at least 4 times that of the largest prior art training set found), and assuming the investigation into SUE features suggests a well-defined critical feature characterisation (i.e. a well-defined target function), it is proposed that this decision is justified. That is, it is well known that for supervised learning to be reliable, the dataset from which the knowledge is drawn must be sufficiently comprehensive such that almost every reliable and relevant representation of the concept that you are trying to model is observed and learned from. While there is no precise way of knowing when this point is reached, it is recognised that it can be asymptotically reached quite reliably by having a very large dataset. On this basis, and given the extensive dataset to be used, it is proposed that the adoption of a supervised learning approach as described is valid.

1.5.2.6. Proposed Evaluation Format

Assuming that an SUE model may be successfully learned from the training data, the effectiveness of this model in detecting (extracting) test-corpus SUEs towards

summarising the content will be gauged Specifically, the effectiveness of the scheme will be presented in terms of the accuracy to which the SUEs may be detected (retained in the summary), and the extent to which the remaining content may be rejected Such a representation is preferable over basic precision/recall hit-rate statistics since, as well as indicating SUE recall (generally recognised as the most important performance quantifier), they also provide the user with an indication of the level to which the duration of the input video has been compressed (summarised)

Given these statistics for the test-corpus content as a whole, it is proposed that a comparison of the relative responses of the individual test-corpus genres be performed towards ascertaining whether or not a consistency of performance is realised across each sport Assuming this is realised, it will then be determined to what level this performance accuracy is comparable with that of the genre-specific equivalent schemes, which represents a very important aspect of the overall performance quantification

1.6. Organisation Of Thesis

The organisation of this thesis reflects the proposed approach to the realisation of the research objective as described in *Section 1 5 2*, and may be summarised as follows

Chapter 1, the current chapter, provided an introduction to the topic of video summarization in general, and to the topic in relation to sports-video content in particular Given this, the specifics of the research objectives to be targeted in this thesis were then formally introduced

In *Chapter 2* an overview of the current state-of-the-art of sports-video analysis technology is provided The literature is presented chronologically, and is categorized according to the modality and methodology of the approaches undertaken The chapter concludes with a discussion on the limitations of the existing schemes

In *Chapter 3*, background knowledge pertaining to the principles of digital video is introduced, with special emphasis on the MPEG-1 video encoding standard (which is the audiovisual representation relevant to this work) This overview is provided such that the subsequent video analysis techniques may be comprehended without difficulty

In *Chapter 4*, the hypothesis for the proposed solution to the problem of developing generic field-sports video summarisation is presented Firstly, the boundaries of the field-sport supergenre are specified in terms of a set of qualities that are said to innately characterise such sports Given this, a generic hypothesis for the automatic

summarisation of field-sports-video is proposed and justified, which is based on the recognition of score update episodes, via the detection and analysis of a set of critical features that are shown to be indicative of them

In *Chapter 5*, the implementation of the hypothesis proposed in *Chapter 4* is described The implementation approach reflects the nature of the content representation used, i e MPEG-1

In *Chapter 6*, in terms of researching a supervised decision making process, the motivation for a machine-learning approach is discussed, coupled with a comprehensive overview of the topic Following this, the justification for employing a Support Vector Machine is presented

In *Chapter 7*, a comprehensive description of the experiments performed is provided, which is supplemented with a detailed evaluation of the results obtained, including a comparison to related work

In *Chapter 8*, the final chapter, a synopsis of the thesis is presented Next, an account of the conclusions drawn following the results evaluation is provided This is then followed by a discussion on potential future work aspects with regards to both the scheme developed herein, and the overall field of sports-video analysis in general

In *Appendix A*, a general introduction to the topic of shot boundary detection is provided, which is supplemented by a comprehensive description and appraisal of the particular shot boundary detection tool used in this work

In *Appendix B*, methodologies are introduced describing how the audiovisual content of an encoded MPEG-1 video may be mined for signal-level data, which is fundamental to the implementation of the hypothesis proposed in *Chapter 5*

In *Appendix C*, the concept of pixel erosion is introduced, a technique that is utilised in the implementation stages of this work

In *Appendix D*, an overview introduction to the technology underpinning Support Vector Machines is presented, which represents the chosen pattern classification (decision making) methodology of this work

In *Appendix E*, the specific Support Vector Machine implementation chosen to realise the pattern classification process is introduced

In *Appendix F*, an analysis into the speed response of the developed system is presented

In *Appendix G*, potential avenues for improving the speed response of the system are discussed

1.7. Chapter Summary

In this chapter, the motivation for video summarization was introduced, coupled with an overview of the two broad approach methodologies typically used to realize such technology, i e accelerated presentation (basic summarisation) and event detection (highlighting) Next, the more specialized area of sports-video summarisation was discussed, with particular reference to the amenability of such content towards event detection-based summarization Also outlined was the dichotomy in approach methodologies for sports-video analysis, i e those of a genre-specific orientation, and those geared towards genre-independent solutions Given the arguments for a more generic methodology, the obstacles challenging the development of such were discussed Towards overcoming these challenges, an approach was proposed based on the division of the sports-video domain into subgroups consisting of characteristically similar genres, i e supergenres Given the supergenre concept, it was then described how the research objective for the work undertaken in this thesis corresponds to targeting a specific case study of this approach, i e the development of a generic, event detection-based, summarisation solution for the field-sports-video supergenre Next, the proposed realisation approach was outlined, and the chapter then concluded with a description of the organization of the thesis

Chapter 2

Sports-Video Analysis

In this chapter a comprehensive overview of the current state-of-the-art of sports-video analysis technology is provided The literature is presented chronologically, however, it is also categorized according to the generality of the approach methodology and/or degree of signal modality of the underlying techniques employed Following this a discussion is presented in which the limitations of the existing schemes are described

2.1. Overview

Given the large television audience figures recorded, it is clear that sports-events broadcasts exhibit substantial public appeal In response, extensive research activity is currently in progress, the aim of which is to adequately model the subject from a video processing perspective Given its amenability to event-based highlighting described in *Chapter 1*, much of this research is concerned with finding robust solutions to the problem of automatic summarization of such content As explained, if this problem may be satisfactorily addressed, it will function as a catalyst in driving the development of more comprehensive sports-video browsing/streaming applications

As mentioned in *Chapter 1*, the schemes constituting the sports-video analysis literature are numerous, but may be broadly classified into two distinct categories, i e genre-specific and genre-independent (generic) methodologies However, as will become evident during the forthcoming discourse, the large majority of these adopt the former methodology As described, this inclination is due primarily to the combination of (i) the dramatic variances in broadcast styles observed for different sports genres, and (ii) the accuracy/performance attainable by maximizing the domain constraints

2.2. Genre-Specific Approaches

This section aims to provide an overview of the current state-of-the-art of genre-specific approaches to sports-video analysis In such works, the solutions derived pertain to unique genres That is, multi-sports genre applicability tends to be forfeited for the sake of increased performance accuracy in the target genre The genre-specific schemes listed are organized according to the degree of signal modality of their underlying processing techniques

2 2.1. Uni-Modal Techniques

Uni-modal schemes correspond to those whose processing techniques are rooted in the analysis of a particular signal domain only Categorized on the basis of modality type (i e video/audio), the following is an overview of uni-modal genre-specific approaches to sports-video analysis

2 2 1 1 Video-Based Techniques

In 1995, Yow *et al* published a study entitled "*Analysis and Presentation of Soccer Highlights from Digital Video*" [13] Therein, the authors present a methodology for the automatic extraction of the effervescent moments (highlights) from soccer-video using purely visual-based analysis metrics The algorithms utilized exploit prominent features of the soccer game, such as ball tracking, goal post detection, and camera movement compensation In addition, the issue of user presentation is investigated, whereby the authors show how camera motion parameters may be used in generating image mosaics for visual browsing Specifically, they propose the construction of panoramic views, arguing that presentation of the highlights via the panoramic construction allows a clearer view of the field and a more accurate depiction of motion paths

In 1997, Choi *et al* published a discourse entitled "*Where are the Ball and Players? Soccer Game Analysis with Color-Based Tracking and Image Mosaick*" [14] In this paper, the authors suggest an approach towards the detection and tracking of soccer objects again towards soccer-video mosaicing In this instance, the objects of interest are the soccer ball and individual players Initially, the scheme is concerned with the precise identification of said objects, and then subsequently it attempts to accurately trace their trajectories throughout the game The techniques are based purely in the visual domain

and they are rooted in metrics pertaining to dominant colour detection and template matching

Also in 1997, Saur *et al* published a paper entitled "*Automated Analysis and Annotation of Basketball Video*" [15] In this work the authors propose an approach for the automatic indexing of basketball video utilizing purely visual-based analysis techniques Specifically, low-level visual feature data is extracted from the content, which is coupled with advanced knowledge of basketball video structure On this basis a high-level segmentation of the content into pre-defined categories is achieved The categories are chosen on an empirical basis and include close-up views, wide-angle views, fast-breaks, and steals The authors maintain that classification of segments into these categories is sufficient such that basketball video annotation may be achieved to satisfactory levels

In 1998, Kawashima *et al* published a paper entitled "*Indexing of Baseball Telecast for Content-Based Video Retrieval*" [16] In this work, an approach is proposed which addresses the challenge of automatic indexing in a baseball-video context, which is based solely in visual analysis techniques In this work, the authors argue that baseball video is inherently cyclic, so that shot-types exhibit explicit periodicity This shot-type periodicity is coupled with some camera view constraints, and together both features are exploited in the reasoning of the annotation hypothesis To perform shot-type classification, colour templates are extracted for each shot-type, such that a set of shot-type templates is generated Subsequently, the colour features of a given frame are compared with those corresponding to each of the set of preconceived shot-type templates Additionally, on-screen graphical text is detected and recognised via a conventional optical character recogniser This feature is then exploited towards providing a further cue for the overall indexing task

Also in 1998, Sudhir, Lee, and Jain published a paper entitled "*Automatic Classification of Tennis Video for High-Level Content-Based Retrieval*" [17] In this work the authors suggest an approach towards the automatic indexing of tennis video, towards realizing an efficient retrieval solution in the context of the domain The approach, which utilizes visual analysis techniques exclusively, is based upon the generation of an image model for the tennis court lines The method exploits knowledge of tennis court dimensions, line connectivity and typical camera perspectives for the genre Furthermore, the tennis court surface type (clay, grass, cement, carpet) is estimated based on colour information Subsequently, player tracking is performed utilizing a template-matching algorithm Armed with these features, the authors propose that data

pertaining to court line location and player positioning may be integrated such that the recognition of high-level semantic events may be realized

In 2000, Zhou, Vellaikal, and Kuo published a paper entitled "*Rule-Based Video Classification System for Basketball Video Indexing*" [18] In this study, the authors propose a video classification methodology for basketball content, based on a feature-orientated supervised heuristic scheme Specifically, the system aims to automatically segment, classify and cluster basketball video scenes into a finite number of semantic categories relative to the nature of the game The rules for the classification process are determined using an inductive decision-tree learning approach, applied to multiple low-level visual image features The specific visual features utilized in the analysis include colour, edge detection, and motion direction estimation

Also in 2000, a paper entitled "*Soccer Video Mosaicing using Self-Calibration and Line Tracking*" [19] was published by Kim and Hong Therein, the authors propose a visual-based scheme that attempts to automatically generate mosaics from soccer-video The methodology is rooted in the detection and tracking of playing field lines, which the authors maintain provide a reliable basis for mosaic construction To this end, an algorithm is designed and employed in estimating the field line locations Once such are located, camera motion parameters are exploited towards self-calibrating the line-tracking algorithm It is maintained that, given the self-calibration aspect, the scheme should reliably handle rotating and zooming camera angles

In 2001, Xu *et al* published a paper entitled "*Algorithms and System for Segmentation and Structure Analysis in Soccer Video*" [20] In this work the authors propose an approach to a high-level segmentation task for soccer-video content Specifically, the basic objective of the scheme is to provide an indication of whether the ball is in play or not – a task commonly known as play-break segmentation The authors argue that this information should provide a good platform for a more sophisticated analysis to be performed at a later stage The approach uses visual analysis metrics exclusively, and is based on algorithms performing both dominant colour detection, and shot-type classification into well-defined categories such as global, zoom-in, and close-up

Also in 2001, Tovinkere and Qian published a paper entitled "*Detecting Semantic Events in Soccer Games Towards a Complete Solution*" [21] Therein, the authors present a methodology designed to detect a wide range of semantic events that may occur in soccer matches The event detection scheme is rooted in the exploitation of player/ball positional knowledge, and is based on the development of a set of heuristic rules

representing prior knowledge of such events However, the scheme is entirely dependent on the availability and accuracy of this object position knowledge The authors suggest that this may either be inferred from the processing of video sequences, or from a tracking system interpreting signals emitted by transponders attached to the players and ball during the game

In 2002, Utsumi *et al* published a study entitled "*An Object Detection Method for Describing Soccer Games from Video*" [22] In this work the authors propose a scheme for automated indexing of soccer-video It is argued that the tracking of soccer video objects, such as players and field-lines, is critical to the task of describing the contents of a game In the proposed scheme, playing field regions are initially extracted based on an *a priori* assumption of field colour Next, super-imposed graphics are detected by exploiting the edge density of video-text, and such regions are then excluded from the subsequent analysis Following this, an algorithm for player detection is then proposed, based on colour rarity and local edge properties The authors argue that because players follow erratic movements, template matching becomes the natural choice for the robust tracking of players On this basis, once detected, a tracking algorithm for players is proposed using a colour-based pattern matching technique

Additionally in 2002, Assfalg *et al* published a paper entitled "*Soccer Highlights Detection and Recognition using HMMs*" [23] In this study the authors propose purely visual-based analysis techniques, in an approach for automatic highlight detection within the framework of soccer-video Specifically, the scheme is based on the detection of event-characteristic patterns of (i) particular object locations, and (ii) temporal evolutions of camera motion On the basis of these features, the system aims to detect distinct soccer-video events such as free kicks, corner kicks and penalties The classification is performed using Hidden Markov Models in a statistical modeling procedure

Also in 2002, Xie and Divakaran, published a paper entitled "*Structure Analysis of Soccer video with Hidden Markov Models*" [24] This work utilizes visual-based analysis techniques in an attempt to provide a high-level temporal segmentation of soccer-video Specifically, the task is play-break detection, which corresponds to the challenge of segmenting the content into two mutually exclusive states i e ball-in-play and ball-out-of-play The techniques involved exploit metrics pertaining to dominant colour ratio and visual motion intensity Given these features, it is shown how each distinct state of the

game may be represented and subsequently classified using a set of hidden Markov models

Also in 2002, Chang, Han, and Gong, published an article entitled "*Extract Highlights from Baseball Game Video with Hidden Markov Models*" [25] In this work, the authors utilize visual-based analysis techniques in an approach to automatic highlight detection in the context of baseball-video The principal argument of the scheme is that ostensibly most highlights in baseball games are composed of certain types of camera shots Furthermore, it is postulated that for highlight scenes, such shot-types exhibit a special transition context in time It is argued that the recognition of these highlight-indicative shot-type transitions should provide for reliable highlight detection within the context of the genre Camera motion parameters, colour features, and edge features, are exploited in a shot-type classification procedure Following this, the highlight-indicative shot-type transitions are inferred via a statistical learning method based on hidden Markov models

Also in 2002, Lazarescu, Venkatesh, and West, published a paper entitled "*On the Automatic Indexing of Cricket using Camera Motion Parameters*" [26] In this work the authors propose a visual-based method that addresses the challenge of automatic video annotation applied to cricket-video Based on an estimation of camera motion activity towards shot-type categorization, visual analysis metrics are designed in order to compute shot-level features such as dominant camera motion, average dominant motion, angle of camera movement, and shot length Shot-type classification is then performed via a fusion of the data corresponding to these feature extractors On this basis, a video index is then inferred from knowledge of shot-types ascertained

In 2003, Ekin, Tekalp, and Mehrotra published an article entitled "*Automatic Soccer Video Analysis and Summarization*" [27] In this work the authors propose a comprehensive approach to the challenge of event detection-based summarization of soccer-video Specifically, the scheme is rooted in visual-based algorithms that perform a variety of low-to-mid-level feature extractions The mid-level features extracted include dominant colour region detection, shot-type classification (into long, medium, and short categories), referee tracking, line tracking, and penalty box detection Based upon a heuristically driven fusion of evidence pertaining to these extracted features, it is shown how higher-level semantic knowledge (i e highlights, including goals) may be inferred from the content

Additionally in 2003 Kijak, Oisel, and Gros, published a discourse entitled "*Temporal Structure Analysis of Broadcast Tennis Video using Hidden Markov Models*" [28] Therein, the authors propose a visual-based analysis approach for video structure analysis in a tennis video context Specifically, colour and motion attributes of detected camera shots are used to perform shot-type classification into two distinct categories (i) global view, and (ii) other It is argued that from this knowledge of tennis video, a temporal segmentation of the overall game into play/playbreak scenes may be inferred Following this, a trained hidden Markov model is used to analyse the temporal interleaving of shot-types, towards revealing the identification of higher-level semantic events within the content

Also in 2003, Assfalg *et al* published a paper entitled "*Automatic Interpretation of Soccer Video for Highlights Extraction and Annotation*" [29] Therein, a visual-based approach is proposed for the detection of significant events in soccer-video Based on temporal logic, methodologies are proposed for the detection of four distinct highlight events These 'basic' episodes correspond to forward launches, shoots on goal, possession turnovers, and placed kicks The features exploited in the development of the event modelling schemes correspond to (i) the recognition of play-field zones in the frames, (ii) the analysis of camera motion parameters for inferring ball movement, and (iii) estimations of player presence density within critical field regions

2 2 1 2 Audio-Based Techniques

In 2000, Rui, Gupta, and Acero published a paper entitled "*Automatically Extracting Highlights for TV Baseball Programs*" [30] In this work, the authors propose a purely audio-based scheme for automatic highlight detection in baseball video, arguing that the exploitation of visual domain features is typically overly computationally expensive In this analysis, the authors maintain that, within the domain context limitations, audio segments that exhibit both substantial energy and high pitch level, typically correspond to those of enthusiastic human speech On this premise, the authors propose a scheme that attempts to segment the audio track into speech and non-speech segments, utilizing a metric based on the first derivative of Mel Frequency Cepstral Coefficients (MFCC) and band energy Furthermore, it is postulated that the majority of the exciting segments in baseball games occur immediately after the incidence of a 'pitch-and-hit' event Hence the development of an audio-based baseball hit detection scheme Armed with such evidence, it is then proposed that highlight detection may be achieved via a system of

data incorporation, which methodically fuses the results from the two distinct feature analyses

In 2001 Zhang and Ellis published a technical report entitled "*Detecting Sound Events in Basketball Video Archive*" [31] This paper reports on a proposed audio-based scheme for automatic highlight detection in basketball-video The primary argument of the approach is that there is a substantial correlation between event significance and the phenomenon of spectator cheering To this end, low-level audio features are extracted from the audio track These include MFCCs, LPC entropy, and normalized energy This feature evidence is utilized in a Neural Network based learning process, which, following a training phase, infers models for the classification of both enthused crowd noise and human speech Furthermore, it is proposed that other basketball events, such as ball dribbling, exhibit specific aural characteristics, and are therefore conducive to an aural-based classification using template matching methodology

2 2 2 Multi-Modal Techniques

Multi-modal schemes correspond to those whose processing techniques are rooted in the fusion of data extracted from more than one signal domain The following is an overview of multi-modal genre-specific approaches to sports-video analysis

In 2001 Nepal, Srinivasan, and Reynolds, published a study entitled "*Automatic Detection of Goal Segments in Basketball Videos*" [2] In this work, the authors propose audiovisual analysis techniques in addressing the issue of delimiting score events within basketball-video content The approach is based on feature detection used in combination with heuristic rules inferred from a manual observation of basketball content Specifically, the authors argue that goal segments are flagged by key events such as crowd cheer, scoreboard display, and a change in direction of player orientation Feature extractors pertaining to these characteristics are thus designed using techniques including volume envelope estimation, graphical text detection, and motion vector field analysis Data obtained from these feature extractors is then fused according to heuristic rules in ascertaining the locations of score segments

In 2002 Cabasson and Divakaran, published a dissertation entitled "*Automatic Extraction of Soccer Video Highlights using a Combination of Motion and Audio Features*" [32] In this work, the authors propose audiovisual analysis techniques in an approach to the challenge of automatically highlighting soccer-video Specifically, it is observed that within such content, any important event (e g a goal) leads to a temporary interruption

of the underlying game On this basis, it is argued that the intensity of motion should be indicative of event importance To this end, a motion activity descriptor metric is designed based on mean motion vector magnitude of video frames Furthermore, based on the observation that significant events in soccer-video are typically associated with short-term audio energy surges (resulting from crowd noise and/or human speech), a method for tracking audio energy levels is developed Given the two extracted features, the temporal patterns of motion activity surrounding detected audio peaks are used in inferring events of interest from within the content

Additionally in 2002, Petkovic *et al* published a discourse entitled "*Multi-Modal Extraction of Highlights from TV Formula 1 Programs*" [33] Therein, the authors propose an approach for the automatic detection of highlights in broadcast Formula-1 video, based on the fusion of data from audio, visual, and textual information sources Initially it is postulated, that when an important event occurs within Formula-1, the announcer raises his/her voice in excitement Such incidents are detected using algorithms for speech end-point detection, followed by excited speech detection This audio evidence is then combined with that gathered by visual analysis metrics relating to colour, shape and motion The multi-modal evidence is then exploited towards modelling events such as over-take, race-start, and fly-out Furthermore, the authors propose that within this specific genre, superimposed text tends to be event descriptive On this premise, they propose an event-based query-and-retrieval model, which is centred on the recognition and interpretation of this video-text

Also in 2002, Li and Sezan published an article entitled "*Event Detection and Summarization in American Football Broadcast Video*" [34] In this study the authors propose a framework for automatically highlighting American football content Therein, it is argued that the issue of play/playbreak detection is fundamental to the summarization procedure, and to this end, approaches for the detection of the play/playbreak segments are proposed based on visual characteristics such as dominant colour detection, playing field detection, and global view detection It is proposed that once the play segments are delimited, they may be extracted and subsequently concatenated, thus generating a compact, time-compressed summary of the original video It is argued that such a summary is comprehensive, in that it encapsulates all of the important moments of the underlying game Additionally, it is proposed that this provides a superior platform for more sophisticated highlighting procedures, compared to the original content Finally, it is argued that audio energy level is reliably indicative of event significance Thus it is

proposed that, following the summarization procedure, audio level evidence should be exploited in generating a significance hierarchy of the events constituting the generated summary

In 2003, Dayhot, Kokaram, and Rea, published a paper entitled "*Joint Audio-Visual Retrieval for Tennis Broadcasts*" [35] The authors suggest audiovisual analysis techniques in an approach towards the automatic extraction of the basic semantic episodes within tennis-video Specifically, the authors argue that segments that constitute a continuous passage of play represent the fundamental elements of such content It is argued that these episodes exhibit both a global court view, and a specific audio characteristic corresponding to the noise of the ball hitting the racquets On the basis of these features it is proposed that these segments may be detected and hence extracted To this end, global court views are detected using Hough transform analysis, coupled with advanced knowledge of scene geometry In detecting ball hits, the power spectrum of the audio signal is windowed into 40ms segments, and Principle Component Analysis is used to identify the distinct sound of the ball hitting the racquet Evidence pertaining to these features is then probabilistically fused in detecting and extracting the required segments

Also in 2003, Chen *et al* published a paper entitled "*Detection of Soccer Goal Shots Using Joint Multimedia Features and Classification Rules*" [36] In this work, the authors propose a multi-modal data-mining framework for the identification of goal events in soccer-video Initially, methodologies are proposed for the shot-level extraction of low-level descriptors to characterize the dynamics in critical soccer-video features such as grass ratio in the visual domain, and audio energy It is then proposed that goal shot candidates may be inferred from specific patterns exhibited in these shot-level descriptors, based on a set of rules inferred from an exploitation of domain specific knowledge of soccer-video The scheme is tested across a small soccer-video dataset

2.3. Generic Approaches

This section aims to provide an overview of the current state-of-the-art technology for approaches to sports-video analysis that aim to be more generic in terms of multi-genre operability Again, the schemes listed are organized according to the degree of signal modality of their underlying processing techniques

2 3.1 Uni-Modal Techniques

2 3 1 1 Video-Based Techniques

In 2001, Pan, van Beek, and Sezan, published a paper entitled "*Detection of Slow-Motion Segments in Sports Video for Highlights Generation*" [37] In this work, the authors propose a visual-based methodology for the genre-independent generation of sports-video highlights inferred from the detection of slow-motion episodes It is argued that the mechanisms that facilitate the variance in playback speed in slow-motion segments, are based on video-frame repetition and/or video-frame dropping Furthermore, it is postulated that such frame repetitions/drops cause large fluctuations in colour intensity between neighbouring frames In exploiting this indicative characteristic, several feature discriminators are employed, which are based on the mean-square-difference of the RGB colour intensity of successive frames These include zero-crossing rate, absolute minima, and absolute difference Following this, a hidden Markov model assumes the feature evidence and calculates the probability of each slow-motion candidate

Also in 2001, Zhong and Chang published a work entitled "*Structure Analysis of Sports Video using Domain Models*" [38] In this investigation, the authors propose a framework for scene detection towards structure analysis in both tennis and baseball-video contexts Specifically, the authors argue that sports-videos exhibit consistencies, which may be exploited in their analyses For example, (i) they usually occur in a specific playground, (ii) they have a fixed number of camera views, (iii) they contain abundant motion information, and (iv) they exhibit well defined content structures In the analysis of tennis and baseball content, the temporal structure of the video is automatically segmented, by detecting the re-occurring event boundaries for each genre, i e the serve in tennis and the pitch in baseball The underlying techniques for these tasks involves the detection of the camera views fundamental to the respective events This is achieved via visual metrics based on colour filtering, object segmentation, and edge detection The approach is illustrated independently for both tennis and baseball-video, and the authors argue that once detected, these events indicate the boundaries of higher-level semantic structures

In 2002, Wu *et al* published a discourse entitled "*Events Recognition by Semantic Inference for Sports Video*" [39] Therein, the authors propose a visual-based semantic inference scheme for generic event recognition within integrated athletics-video broadcasts Specifically, it is argued that when a semantic concept changes within a sports-video, it is typically accompanied by an abrupt change in the velocity of the

global motion characteristic On this basis, a global motion estimation (GME) algorithm is utilized in segmenting athletics-video sequences, according to changes in its velocity levels Following this, for a segmented event clip, it is proposed that knowledge pertaining to background-type, foreground objects, and motion velocity, should contribute effectively towards the event recognition procedure In developing this hypothesis, GME is used in separating foreground from background layers in the video images Subsequently, low-level features such as colour and texture are used in characterising the background/foreground features of the clip Lastly, GME is used again in characterising the local motion of the clip This low-level clip evidence is then mapped to a set of mid-level semantic concepts, which describe the nature of the clip The event-specific pattern of semantic concepts is input to a trained finite-state machine, which ultimately provides the event-type decisions

Additionally in 2002, Assfalg *et al* published an article entitled "*Semantic Annotation of Sports Videos*" [40] In this work, the authors propose a visual-based approach to sports genre identification, in the context of integrated Olympic Games-video broadcasts At the outset, it is argued that discrimination between studio and live-action content may be achieved by exploiting the well-defined syntax of studio scenes That is, it is argued that such scenes exhibit consistent characteristics, such as a limited number of camera views, and a repeating pattern of shot content Following this segmentation, visual analysis techniques are proposed for content knowledge acquisition concerning the live-action segments Specifically, colour, edge, shape, and luminance feature metrics are extracted These are then employed in a shot-type classification process, which classifies according to global, close-up, graphical and crowd view categories Furthermore, it is argued that the most relevant distinguishing feature of global (playing-field) views corresponds to colour Thus, following a playing field segmentation procedure (based on dominant colour), a colour feature metric is coupled with a field-line orientation distribution analysis Based on this feature evidence, individual sports genres are automatically distinguished within the overall broadcasts

Also in 2002, Pan, Li, and Sezan published an article entitled "*Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions*" [41] In this paper (essentially an extension to their previous work on slow-motion detection towards sports video highlighting [37]) an algorithm is proposed for the detection of all replay segments in sports-video, i e capturing even those that do not exhibit slow-motion playback The method exploits the typical use of graphical effects

in scene transitions that are typically used to delimit the replay segments Specifically, a colour histogram-based template matching technique is used to detect the broadcaster logo, which is purportedly prevalent during such transitions The templates are generated dynamically by sifting through the content corresponding to slow-motion segments, which are detected automatically using the technique previously developed The authors maintain that logo locations correspond exclusively to the start and end points of replay segments, and that the detection of such provides for reliable replay segmentation

2.3.1.2 Audio-Based Techniques

In 2003 Xiong *et al* published a work entitled "*Audio Events Detection Based Highlights Extraction from Baseball, Golf, and Soccer Games in a Unified Framework*" [42] In this study, the authors propose an audio-based approach to automatic sports highlights detection, which aims to be generically applicable across baseball, golf and soccer-video genres The principal argument of the scheme is that within these sports genres, the spectators typically show appreciation for exciting or interesting play by loudly applauding and/or cheering On the basis of this correlation, it is argued that reliable identification of such phenomena within the audio content should contribute effectively towards the automatic highlighting task In developing this hypothesis, frequency-spectrum based MPEG-7 audio features are extracted from the audio track Based on indicative feature patterns of this data, hidden Markov models are employed for the classification of the critical audio segments This process is also augmented by some pre/post-processing techniques for the filtering of false positives from commercials, etc

2.3.2. Multi-Modal Techniques

In 2002 Peker, Cabasson, and Divakaran published an article entitled "*Rapid Generation of Sports Video High-Lights using the MPEG-7 Motion Activity Descriptor*" [43] In this work, the authors propose an audiovisual-based methodology for automatic highlights detection, which is applicable to multiple genres of sports-video The principal argument of the scheme is that temporal patterns of motion activity are intrinsically related to the grammar of sports content Specifically, it is thus proposed that highlights may be detected by falling/rising edges of a motion activity characteristic, and therefore the detection of such enables the skipping of uninteresting events To this end, the MPEG-7 motion activity descriptor is employed to represent the temporal patterns of this

characteristic Furthermore, it is proposed that other compressed domain features may be used to further improve the accuracy of the scheme, i e it is maintained that interesting events in sports-video are typically accompanied by high-energy audio segments, resulting from crowd noise and/or enthused human speech On this basis, it is proposed that energy peaks in the audio signal be detected, and hence utilized in refining the initial analysis

Additionally in 2002, Babaguchi, Kawai and Kitashi, published an article entitled "*Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration*" [44] In this paper the authors propose a visual-textual based approach to sports-video indexing via the automatic recognition of semantic events The techniques employed exploit the temporal correlation between aspects of visual events and the syntax in an associated closed-caption stream Assuming that the structure of sports-video is well defined, given a particular sports-video, it is proposed that a structure tree may be derived that models the chain of events for the underlying game On the basis of this advanced knowledge, the sports genre structure tree is analysed, such that each target event is characterised in terms of a set of appropriate keywords The closed-caption stream is then probed, and the detection of a specific event keyword activates the particular analysis conventions for that event These include the selection of a temporal interval, the expected frequency of the keyword within that interval, and the definition of correlated keywords It is argued that such keyword ontologies and structure trees may be constructed for any sports genre and hence, the method is ostensibly transferable across multiple genres, which exhibit closed-caption textual streams The scheme is demonstrated for American football video

Also in 2002, Duan *et al* published a study entitled "*A Unified Framework for Semantic Shot Classification in Sports Videos*" [45] In this work, the authors present an approach towards the automatic cataloguing of generic sports video shots into semantic categories It is argued that for sports-video in general, a finite number of predefined semantic shot categories are sufficient to represent the majority of scenarios that constitute such content Proposed categories include field view, court view, goal view, zoom-in, close-up, audience view, etc Furthermore, it is proposed that a specific sports genre may be represented wholly by just a subset of these categories In practice it is required that advanced knowledge of the sports genre in question be known such that an appropriate subset may be instigated Low-level features such as colour, texture, and motion vectors, are extracted from the content Evidence from these sources is mapped

to mid-level semantic features such as dominant object motion (e g a player), camera motion patterns, and homogeneous regions (e g court shape) Such mid-level features are then appropriately fused so that they map to high-level semantic shot attributes These shot-level attributes are then used in a genre-specific heuristic process, such that each shot is classified into one of the predefined categories of the prescribed genre subset The proposed method is demonstrated across tennis, basketball, and soccer video

In 2003, Li *et al* published a paper entitled "*Bridging the Semantic Gap in Sports*" [46] In this study, the authors describe how sports-video modeling towards event detection contributes to the reduction of the semantic gap by providing rudimentary semantic information to the user, obtained through media analysis Specifically, a general framework for indexing sports broadcast programmes is proposed The framework is based on a high-level model of sports-video, which utilizes the concept of an event, defined according to genre-specific knowledge The event detection algorithms are developed via pattern recognition analyses in both the visual and aural signals However, in practice, advanced knowledge of sports genre is required such that the framework is suitably configured, and appropriate event models chosen Furthermore, it is explained how the solution may be further advanced by exploiting the availability of an independently generated source of rich textual metadata The overall scheme is demonstrated for American football, baseball, Japanese sumo wrestling, and soccer video

Additionally in 2003, in progressing their previous works, Xiong, Radhakrishan, and Divakaran, published a paper entitled "*Generation of Sports Highlights using Motion Activity in Combination with a Common Audio Feature Extraction Framework*" [47] Therein, the authors propose a combination of their earlier techniques, which concerned the exploitation of camera motion [43], and audio characteristics [42], respectively This combined multi-modal approach aims to tackle a similar challenge to that addressed previously, i e that of developing a generic solution for the automatic highlighting of soccer, golf, and baseball-video It is shown that the combined fusion of aural and visual features in this multi-modal approach achieves increased performance accuracy for the task

Also in 2003, Hanjalic published a paper entitled "*Generic Approach to Highlights Extraction from a Sport Video*" [48] In this work the author proposes an audiovisual-based approach to genre-independent automatic sports-video highlighting The principal

argument of the scheme is that for sports-video in general, exciting moments are typically correlated with indicative feature characteristics, including intense motion activity, high shot-cut density, and surges of audio energy In exploiting this correlation, low-level data extractors are developed for the mining of sports-video content towards the characterization of these features Specifically, motion estimation is quantified via a standard block-based algorithm, and metrics for both audio energy envelope detection, and shot-cut rate tracking are developed Upon the extraction of this low-level feature evidence, it is justified how an overall temporal excitement curve may be generated, based on a weighted average of all three components It is then shown how a video abstract may be inferred on the basis of its excitement distribution The method is illustrated in a soccer-video context

Finally in 2004, Jianyun *et al* published a paper entitled "*A Unified Framework for Semantic Content Analysis in Sports Video*" [49] In this work, the authors propose an audiovisual-based approach, towards the generation of a genre-independent framework for the syntactical segmentation of sports video The approach aims to model sports-video as a three-tiered hierarchy of basic semantic units (BSUs), which increase in scene granularity from top to bottom However, the work is primarily concerned with content segmentation at the level of the first and second tiers of such At these levels, the BSUs correspond to live-action/advertisement discrimination, and play/play-break discrimination, respectively In addressing these tasks, it is argued that all sports-video programmes consist of regular domain rules and video editing grammar Given this, appropriate low-level feature metrics are developed and employed to mine the content accordingly These correspond to shot duration, audio classification, colour analysis, and camera view classification This feature evidence is then heuristically combined with knowledge of structure consistency, such that the required segmentations may be realised The scheme is illustrated in a soccer-video context

2.4. Discussion

Clearly, the scope of the listed works is extensive, hence, for clarity, an overview is provided in **Fig 2 1** Given these schemes, and focusing on those of a genre-independent orientation in particular, it is required that their limitations are fully expounded towards discerning what is currently lacking, and thus towards enabling an assessment of to what extent any generic solution derived in this work may be

considered as a valuable contribution to the field

2.4 1. Limitations Of The State-Of-The-Art

Overall, it is evident that the major portion of the literature concerns approaches of a genre-specific methodology Furthermore, within this approach domain, analyses specific to soccer-video saturate the field Clearly the abundance of soccer-video schemes reflects the fact that it is the only truly global sport, whereas, the motivation for genre-specific solutions in general has been explained as stemming from the variances in broadcast styles of each genre, as well as exploiting the benefits generated by maximizing the domain limitations However, while many report accurate and reliable performances via this approach (e g the soccer-video solutions in [13], [23], [27], the tennis-video solutions in [17], [28], [35], the baseball-video solutions in [16], [25], [30], etc), as explained earlier, given that they are orientated towards a specific domain, central to most schemes are typically non-recyclable algorithms based on intrinsically characteristic critical features that are peculiar to the sports genre in question That is, towards optimizing performance accuracy for the domain in question, multi-genre operability tends to be sacrificed This inflexibility is a significant shortcoming, and to target solving the overall problem of sports-video summarization by means of developing multiple solutions on a genre-by-genre basis is undesirable from a complexity and an efficiency point of view

In recognition of the drawbacks of the genre-specific approaches, the more recent literature has begun targeting the development of more flexible, widely applicable solutions Of the generic schemes mentioned, while none propose an ultimate 'one-size-fits-all' solution that claims to operate robustly across all potential sport genres, many propose generic frameworks in which sports-genres that are linked by a common event model may be analysed together For example, in [42] and [47] a generic solution is proposed for the automatic highlighting of soccer, baseball and golf, using a common event model based on exploiting spectator cheering and motion dynamics In [38] the authors propose a generic approach for the combined analysis of tennis and baseball video However, many of these schemes, while generic in outlook, have only been evaluated on a narrow genre scope For example, the multi-genre solution of [43] has only been shown to operate on golf-video, while that of [48] has only been tested on soccer-video Furthermore, most schemes do not specify what the limits of their generality are That is, it is typically quite easy to think of genres for which the solutions

would be challenged For example, the genre targeted in [33] (i e Formula-1 motor-racing) is, in general, characterized by constant high levels of both motion and audio noise, which would thus surely have consequences for the multi-genre solution developed in e g [47] So this begs the following questions By what reasoning were the sports-genres chosen to constitute the test bed? How is the uncertainty explained for other sports genres? In short, although the generic schemes mentioned are shown to perform well on the test genres used, there is a lack of specification on the limits of the generality of the solutions

2.4.2. General Observations

It is evident that the vast majority of the conventional approaches to sports-video analysis, whether genre-specific or generic, tend to be uni-modal in nature While many of the uni-modal techniques have been shown to yield reasonable performances in their respective tasks, the results obtained via multi-modal techniques, reported in some of the more contemporary works, suggest that enhanced performances are obtained by means of fusing evidence obtained from multiple signal domains

Overall, the visual-mode features that are most commonly exploited correspond to the pixel-level tracking of colour, luminance, edge histogram, etc, and/or block-level motion estimation and tracking Commonly used audio-mode features include time-domain tracking of short-term energy, zero-crossing rate, etc, and power spectral density (PSD), pitch estimation, MFCCs, etc in the frequency domain Other relevant features that have been shown to be constructively exploitable include those of a text-based orientation, such as superimposed video-text, and closed-captions in the metadata domain

2.5. Chapter Summary

In this chapter a synopsis of the current state-of-the-art technology for sports-video analysis was provided The listed works, spanning a 10-year timeframe, were categorized according to date, approach methodology, and degree/nature of signal modality Following this overview the limitations of the current schemes were described, towards discerning what is lacking in the current state of the art, and thus providing a basis for an assessment of to what extent any generic solution derived in this work may be considered a contribution to the field Some general observations were then discussed

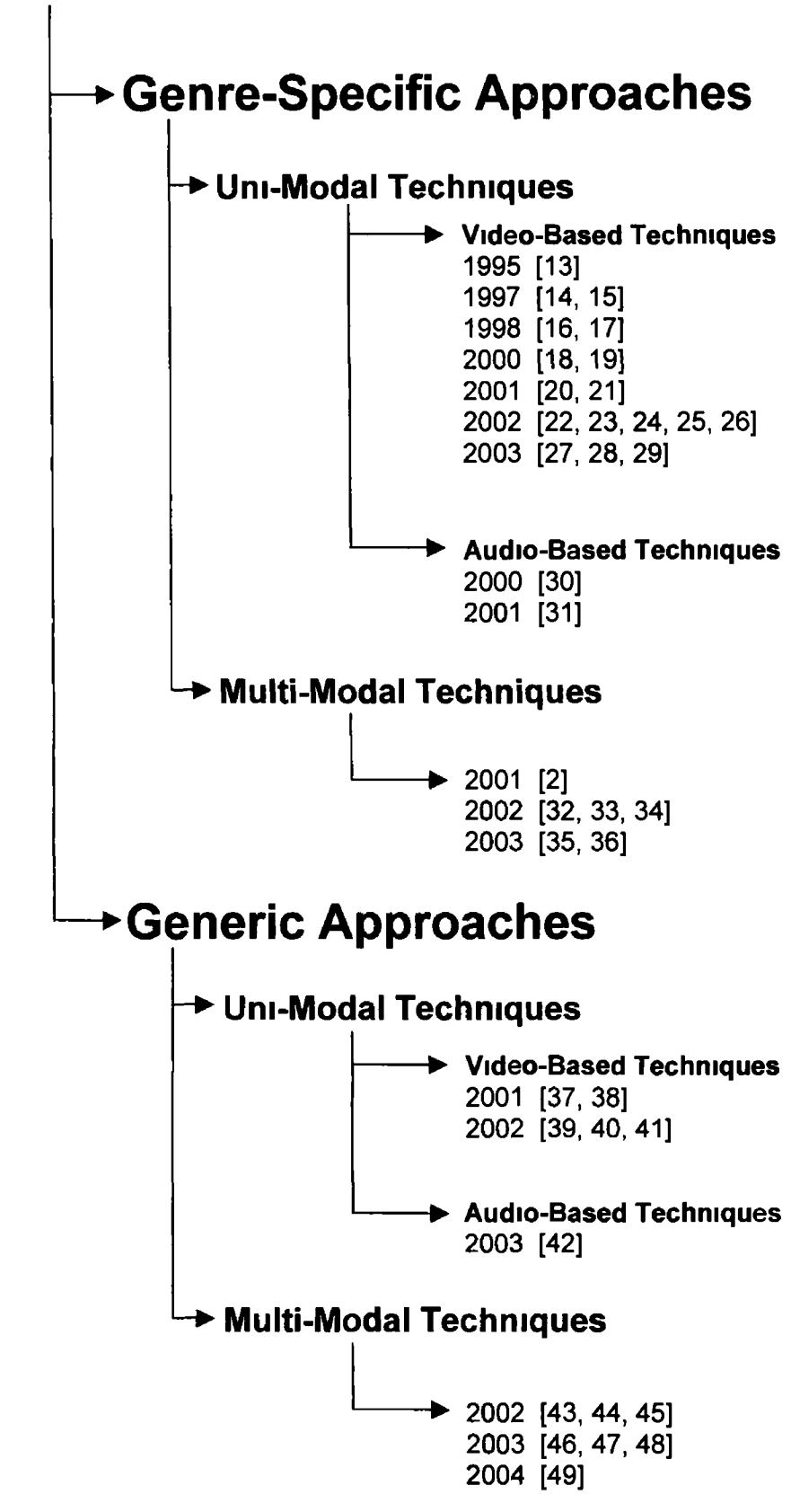


Fig 2 1 An overview of the sports-video analysis literature listed

Chapter 3

Digital Video Principles

As introduced in *Chapter 1*, this thesis is concerned with developing a solution for the automatic summarisation of field-sports-video The procedure will involve the analysis and processing in the digital domain of both the audio and video signals that constitute such content As a starting point for the uninitiated reader, the purpose of this chapter is to provide an introduction to the basic principles of digital video so that the analysis procedures, described in subsequent chapters, may be fully understood Given that the representation used is the MPEG-1 digital video standard (see *Section 1 5 2 1*), an overview of this particular standard is also provided The chapter begins by introducing the concepts of digital video, colour-space models, and video structure modelling This is then followed by an introduction to the topic of data coding and compression, which then leads to a description of both the audio and visual aspects of the MPEG-1 digital video standard

3.1. Digital Video

In recent times there has been a hugely increased interest in multimedia communications from both personal and commercial perspectives This has served to stimulate significant developments in the field of digital video encoding An analogue image signal is generated when a camera scans a 2-D scene and converts the data to an electrical signal In digitising such an image, the signal is sampled, and the samples are then quantised, whereby each sample corresponds to an image pixel Since the pixels are individually embodied as discrete entities, digital images tend to exhibit significant advantages over conventional analogue representations These primarily relate to

efficiency, quality, and conduciveness to analysis and processing For instance, digitised video may be exploited in the development of clever redundancy reduction techniques, which aim to represent the content in a compressed format To this end, many international digital multimedia encoding standards have been established Of specific relevance to the work of this thesis is the MPEG-1 standard video compression and hence a complete description of such is required Prior to this however, the basic concepts appropriate to its underlying technology are introduced, i e colour-space models, video structure, and the approaches to data coding/compression

3.2. Colour-Space Models

The tint, chroma, and brightness attributes of a given colour are directly dependent upon the combined intensities of the fundamental components that constitute the colour-space concerned For example, when particular intensities of the basic primary colours of light are combined, they together comprise a progeny colour, which exhibits unique attributes in accordance to those abovementioned Many colour-space schemes exist in the literature, however, it is the formats that are most relevant to digital video representation that are discussed in this section

3.2.1 RGB Colour-Space Format

The red, green, and blue (RGB) 3-D colour format is the basic colour-space from which all other standard formats may be derived It is the most popular choice for computer graphic applications, since cathode ray tubes (CRTs) utilize red, green, and blue phosphors in creating colour [50] In the RGB scheme, it is the relative intensities of the individual red, green, and blue components, which define the overall progeny attributes of colour, brightness, and saturation To offset the typically non-linear transfer functions of most CRTs, RGB signals are generally put through a process of ***gamma-correction***, which effectively compensates for this non-linearity by inversely warping the RGB values accordingly [51] However, since the human eye is more sensitive to variations in luminance relative to chrominance [52], RGB space is generally not the most efficient representative scheme Hence the development of more effective formats

3.2.2. Luminance-Independent Colour-Space Formats

To exploit the luminance-dominant sensitivity of the human visual system, many TV broadcast schemes, and image-coding standards alike, utilize independent luminance and colour-difference signals to represent visual images One such format is the ***YUV*** colour space, which is the scheme employed in the NTSC, PAL, and SECAM broadcasting standards In this scheme, Y corresponds to the luminance component, and U and V, the colour information YUV signals may be derived from gamma-corrected RGB space as shown below in **(3 1)** [51]

$$\begin{aligned} Y &= (0\ 299 * R) + (0\ 587 * G) + (0\ 114 * B) \\ U &= (-0\ 147 * R) + (-0\ 289 * G) + (0\ 436 * B) \\ V &= (0\ 615 * R) + (-0\ 515 * G) + (-0\ 100 * B) \end{aligned} \qquad \textbf{(3 1)}$$

The main advantage of the YUV colour scheme is that the chrominance information may be subsampled or quantized independently of the luminance information, so that the chrominance bandwidth is reduced compared to that of the luminance component This results in a more efficient overall representation A further advantage of the YUV colour scheme is that it allows for colour television broadcasts to be backward compatible with the prototypical 'black-and-white' TV receivers That is, they are able to receive and interpret the luminance component, while disregarding the colour information

The $\mathbf{YC_bC_r}$ scheme is a similar, but scaled offset version of the YUV format, where Y is defined to have a nominal range of [16-235], and C_b & C_r are defined to have a range [16-240], with zero signal corresponding to level-128 Most of the standard video coding schemes adopt this format as an input image signal YC_bC_r signals may be derived from gamma-corrected RGB space as shown in **(3 2)** [51]

$$\begin{aligned} Y &= (0\ 299 * R) + (0\ 587 * G) + (0\ 114 * B) \\ C_b &= (-0\ 169 * R) + (-0\ 331 * G) + (0\ 500 * B) \\ C_r &= (0\ 500 * R) + (-0\ 419 * G) + (-0\ 081 * B) \end{aligned} \qquad \textbf{(3 2)}$$

Fig 3 1 illustrates an RGB colour image and its equivalent YC_bC_r components Within these the lower spatial sampling rate of the colour difference components is observable as being less sharp (or blurry) compared to the luminance component

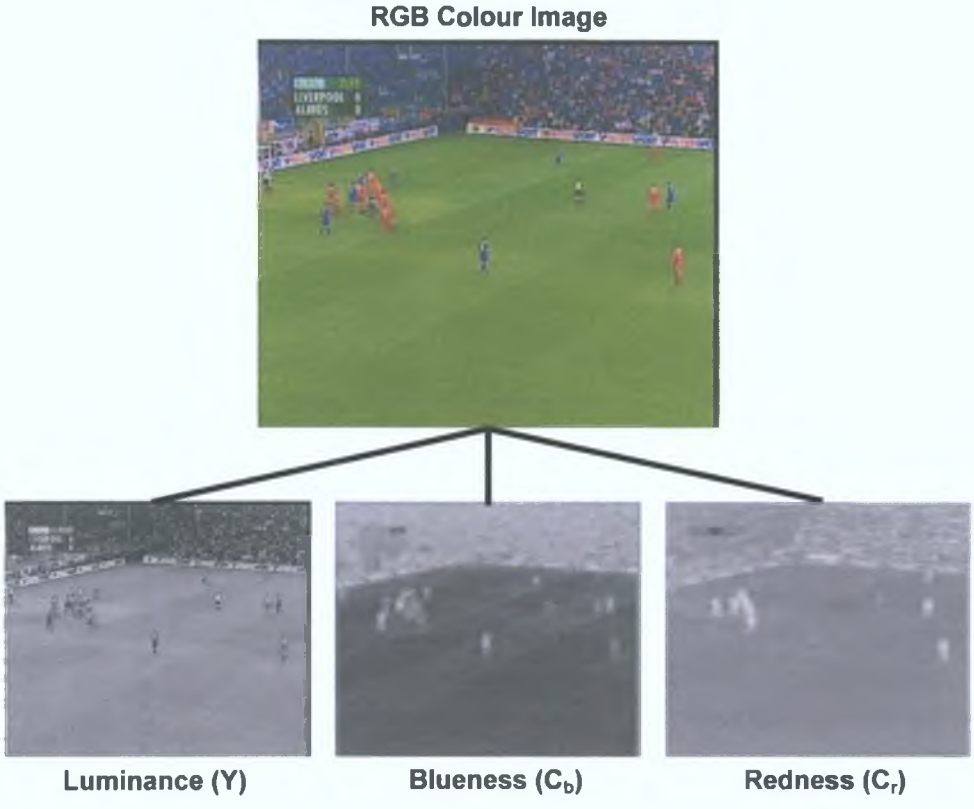


Fig. 3.1. RGB image and corresponding YC_bC_r colour-space components.

3.2.3. HSV Colour-Space Format

While not strictly related to the subject of digital video encoding, the HSV colour space format has been shown to be very useful in terms of colour-based analysis of images. In short, the HSV colour space was purposely designed to be more closely related to the way that the human visual system perceives colour, compared to the other traditional schemes (e.g. RGB, YUV, etc.). Given this, it is in the hue space where colours that are perceptively similar tend to cluster best, hence its usefulness from an analysis point of view. It has three fundamental bands, which according to Munsell [53] may be described as follows. Hue (H) is that quality by which we distinguish one colour family from another (as red from yellow, or green from blue or purple). Saturation (S) is that quality by which we distinguish a strong colour from a weak one, i.e. the degree of departure of a colour sensation from that of a white or gray (i.e. the intensity of a distinctive hue). Value (V) is that quality by which we distinguish a light colour from a dark one. HSV signals may be derived from gamma-corrected RGB space as shown in **(3.3)**. From this it may be shown that the hue component is measured as an angle within the range [0°-

360°] For instance, within this range the primary colours typically reside as shown in **Table 3 1** Meanwhile, S is generally deemed to range between the values [0-1], where 0 represents pure grey and 1 is the pure primary colour, and V typically lies within the range [0-255], with higher values representing brighter colours **Fig 3 2** illustrates a colour FSV video image and its equivalent HSV space components From this figure the similarity between the value (V) component and the luminance (Y) component in YC_bC_r space is evident Note that the large playing field object, while having large fluctuations in both saturation and value, tends to maintain a uniform hue level throughout

$$
\begin{aligned}
&\bullet \textit{If } R = \max(RGB) \\
&\quad H = 60 * [(G - B)/(\max(RGB) - \mathrm{mm}(RGB))] \\
&\bullet \textit{Else if } G = \max(RGB) \\
&\quad H = 60 * [2 + ((B - R)/(\max(RGB) - \mathrm{min}(RGB)))] \\
&\bullet \textit{Else if } B = \max(RGB) \\
&\quad H = 60 * [4 + ((R - G)/(\max(RGB) - \mathrm{min}(RGB)))] \\
&\bullet S = (\max(RGB) - \mathrm{min}(RGB)) / \max(RGB) \\
&\bullet V = \max(RGB)
\end{aligned}
\qquad (3\ 3)
$$

Table 3 1 Hue positions for primary colours

Colour	Hue
Red	0° (360°)
Yellow	60°
Green	120°
Cyan	180°
Blue	240°
Magenta	300°

3.3. Video Structure Modelling

To provide for any level of content-based analysis of video, it is first required that some objective standard of video structure be inferred, towards breaking up the material into its constituent elements To this end, a bottom-up description of the conventional video structural hierarchy is presented in **Fig 3 3**, and to varying degrees, the work described in this thesis performs video analysis operations at each layer of this model

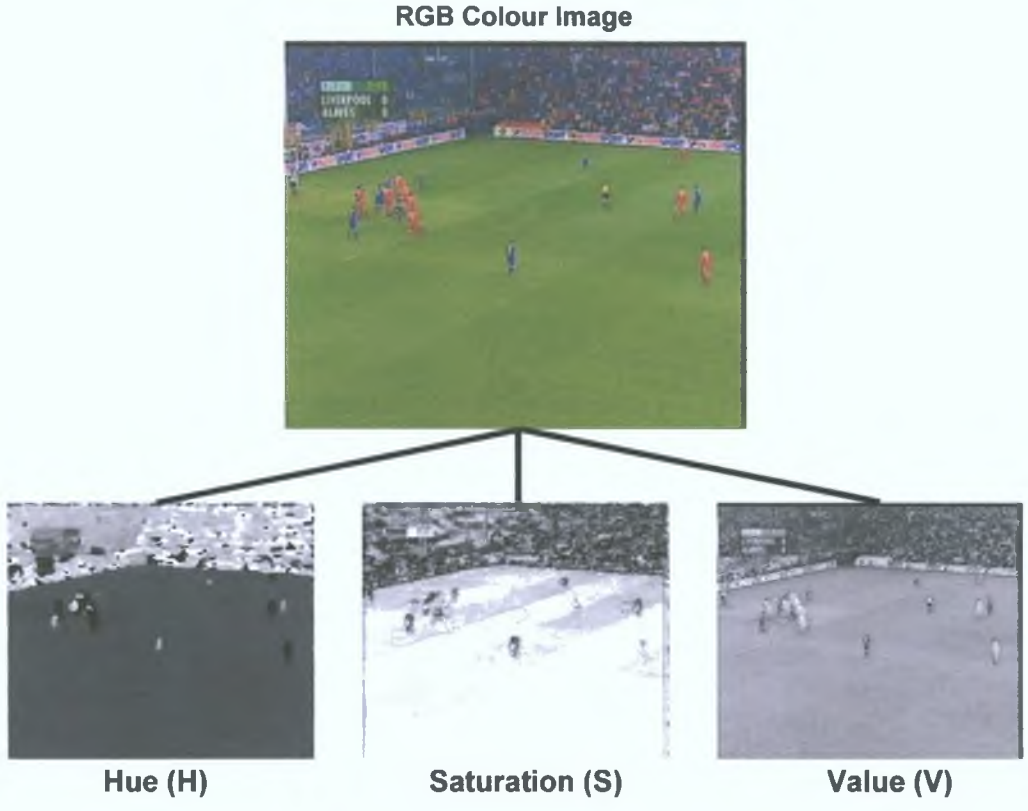


Fig 3.2. Decomposition of colour image into HSV colour-space components.

3.3.1. Pixels

Image pixels are the most fundamental elements of digital video. Short for ***picture element***, a pixel corresponds to a single point in a video image. Specifically, pixels represent the luminance and chrominance information for particular points in image space. Video images are comprised of a dense concentration of pixels, typically arranged in a row and column format, as shown in Fig 3.3. The resolution of an image corresponds to the density of pixels within a given area space.

3.3.2. Image Objects

Pixels unite to form objects, which correspond to the discrete semantic entities that comprise the image environment. However, an object may also simply relate to a logically linked spatio-temporal region, such as the image background (e.g. in the absence of foreground objects in landscape images). As shown in Fig 3.3, it is the blend of foreground/background image-objects that comprises a completed picture.

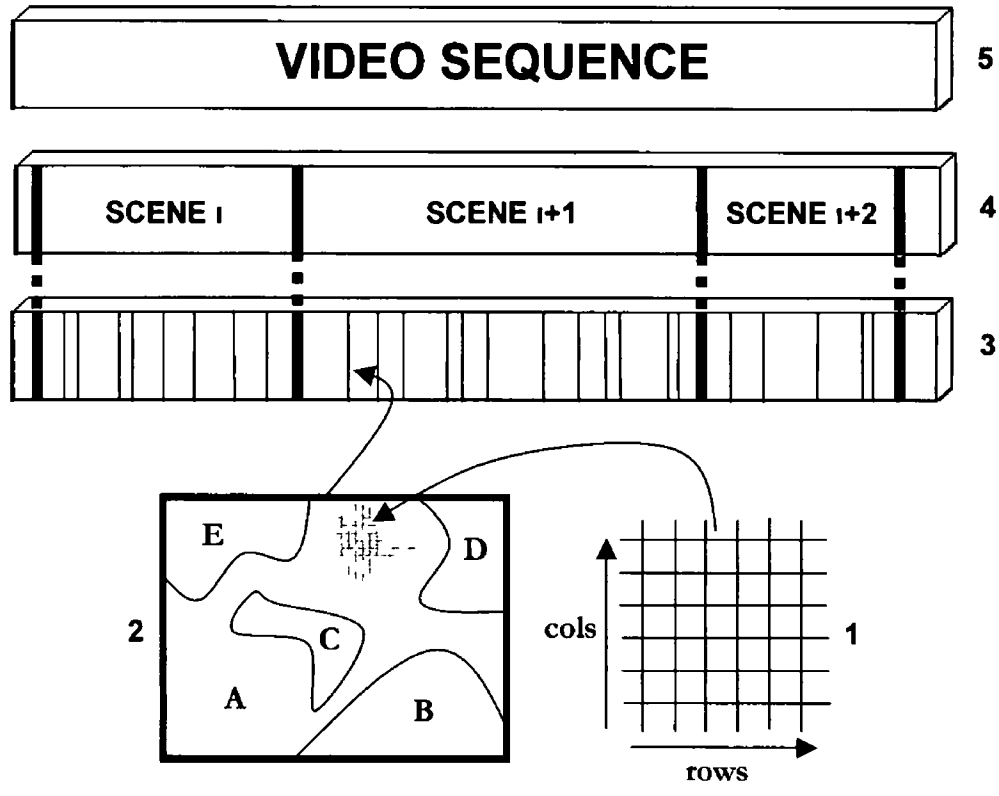


Fig 3 3 Video Structure Hierarchy 1 Pixel level, 2 Image-objects and Frame level, 3 Shot level, 4 Scene level, 5 Video sequence level

3.3.3. Video-Frames

The term ***video-frame*** historically comes from movie films, i e a video-frame is one complete picture or image within a reel of film A complete film may be described as a sequence of frames, which are typically synchronised to an accompanying audio track The frames are presented in a rapid manner, such that to the human eye, visual motion is represented with sufficient fluidity The frequency of these discrete images is called the video ***framerate***, which is typically measured in frames-per-second (fps) Common framerate conventions are 25fps (corresponding to the PAL and SECAM video broadcasting standards), and 30fps (corresponding to the NTSC video broadcasting standard)

3.3.4 Camera Shots

Moving further up the value chain of the video structure hierarchy leads to the shot level A ***camera-shot*** (or simply ***shot***) may be defined as the video resulting from a

continuous, unbroken recording by a single video camera [54] Hence, shots exhibit a flow of video images, which from frame-to-frame are successively very similar to each other Many algorithms exist that aim to delimit the locations of shots within video [55]-[58] Most of these schemes exploit this characteristic of successive frame similarity

3 3 5. Video Scenes

The largest semantic unit within the video structure hierarchy corresponds to the ***video-scene*** A video-scene may be defined as a succession of semantically related camera-shots, which together constitute a single unit of action Video-scenes typically exhibit a consistency in both context and environment, and are generally situated in unique locations However, video-scenes are high-level semantic concepts, the nature of which can be ambiguous Thus scene recognition in video is not always a totally objective task The difficulty concerning this issue hampers the development of reliable automatic video-scene delimiting tools [59]

3.4. Data Coding & Compression

Consider a multimedia article, e g a video sequence, for digital representation As well as digitizing the content, a further desirable objective of an encoding scheme, is to reduce the amount of data that is required to realize an accurate representation of it That is, it is desirable that it be ***compressed***, such that its associated bit-rate demands are reduced The compression should provide for increased efficiency in article archiving, and thus also combat the problems concerning transmission of large articles across limited bandwidth channels

The standard approaches to data compression are typically two-fold The most basic involve techniques for statistical coding, towards the generation of optimized compact representations of the digitized data However more sophisticated approaches concern methodologies for the front-end reduction of source content redundancy that is intrinsic to the characteristics of the article itself

3.4.1. Data Redundancy

Data redundancy is a concept that is common amongst many types of multimedia articles For example, consider a standard black-on-white text manuscript Such

documents naturally exhibit large areas of white space that correspond to the background Simply encoding the black and white regions using a binary digital scheme would clearly yield large redundancy, since the incidents of white space regions would be encoded independently of each other In this situation, a more efficient encoding approach would be to exploit the spatial concentrations of white space within the manuscript Such an approach would thus be expected to yield a much more compact (compressed) representation

Still images and audiovisual sequences exhibit substantial aspects of data redundancy In the case of digital images, most tend to contain redundancy in the spatial domain due to the typically high correlation between neighbouring pixels Furthermore, taking into account the perceptual limits of the human visual system, it may be argued that for a given image, data representing its most intricate detail may not be important to the human eye, and therefore may be rendered expendable In the case of high-framerate video, which is characterised by a rapid sequencing of images, the subsequent frames differ very little from each other Hence significant redundancy may be eliminated by encoding each frame, not in isolation, but with reference to previous and/or subsequent frames Audio sequences, like visual media, also exhibit perceptual redundancy, due to the limitation of the human aural system Similarly, this limitation may be exploited, such that the associated redundancy may be eliminated in the encoding of audio sequences The standard multimedia data compression algorithms typically integrate such data redundancy techniques in realizing digital domain content representations

Any method of redundancy reduction may be categorised as either (i) lossless, or (ii) lossy The decision whether to target either lossless or lossy compression is generally based on the requirements of the target application and/or the nature of the redundancy involved However, it should be noted that the overall performance of any compression technique is usually directly proportional to the amount of redundancy originally contained in the material

3.4.2. Statistical Coding For Lossless Compression

In some situations, while it is desirable for the content to be compressed, it is also required that it be possible for the material to be perfectly returned to its original state, without detriment, by the decoding (decompression) process Example scenarios include document encoding, medical/satellite based imaging etc, i e any situation whereby it is

required that the original article remains wholly intact, from the encoding phase through to the decoding phase Hence, such circumstances require ***lossless*** data compression techniques

The most basic approaches to lossless compression, involve methods that simply exploit repetitive sequences within the content, e g ***Run-Length Encoding*** (**RLE**) Central to these schemes is the substitution of successive series' of similar data value entries within data sequences The repetitive entries are replaced by single data values, coupled with an associated occurrence ("run-length") count, thus representing the data in a more compact fashion While these algorithms are ostensibly conducive to the compression of pixel images, the overall compression performance of these algorithms significantly depends upon the nature of the material involved Whilst uncomplicated, in general these compression methods do not provide high compression ratio performances

A more sophisticated approach concerns ***pattern substitution***, which is effectively a basic mode of statistical encoding In this instance, regularly occurring data patterns are substituted with a short code or flag To achieve compression, the code is selected such that it is small relative to the original data pattern At the most basic level the codes may be statically defined in advance However, a more advanced approach involves the dynamic assignment of the flags ***Entropy encoding*** schemes are techniques that attempt to optimise the assignment of codes, such that the best compression ratios are achieved for content concerned Examples of such schemes include ***Huffman Coding*** and ***Arithmetic Coding***, descriptions of which may be found in [51] These entropy-encoding techniques are inherently based in classical information theoretic methodologies

3.4.3. Source Coding For Lossy Compression

Source-coding algorithms interpret the actual contents (signals) of the raw material While it is feasible to employ these methods in losslessly encoding data, the compression performances of source-coding techniques truly excel when generating ***lossy*** content representations, albeit at a cost of a (tolerable) degradation of the original material That is, with lossy compression the reconstructed article is never an exact replica of the original However, in general, the aim is to obtain the best possible representation of the source material, for a given target bit-rate

Video and audio sequences are conducive to lossy compression since, as outlined, they typically exhibit high degrees of perceptual redundancy In practice, lossy compression concerns an optimisation of the trade-off between data compression ratio, decompressed quality, and scheme simplicity There are three broad techniques involved in lossy source coding, i e ***transform coding***, ***predictive encoding***, and ***vector quantisation***, the first two of which are of primary concern to this discourse

3 4 3 1 Transform Encoding

In terms of multimedia encoding, domain transforms have become central to the most popular lossy compression methods Transforming from one domain (e g time/spatial) to the frequency domain, typically yields a decorrelation of the data as represented in the original domain Consequently, when digitally representing the content, this allows for transform components to be encoded instead of the original data values Perceptual redundancy may then be reduced by appropriately suppressing the least significant components, which are typically more discernible in the transformed domain than in the original

In terms of the development of multimedia coding standards, amongst a pool of many alternatives, the ***Discrete Cosine Transform*** (**DCT**) [60] has become the most popular transform algorithm Its popularity is primarily due to its excellent combined performance in both data decorrelation and in speed of computation

Fourier theory [61] illustrates how a complex function may be represented reasonably accurately by a small set of values (coefficients), which control the weighted superposition of a set of (relatively simplistic) basis functions It may be shown that by projecting a signal onto an orthonormal basis, an efficient signal representation is produced that is optimal [62] Furthermore, it has been shown that the cosine basis, as an instance of an orthonormal basis, is most appropriate for the projection of 2-D spatial image data [63] The 2-D DCT implements cosine basis projection in transforming blocks of spatial image data In the transformed domain, the block data is represented as a superposition of weighted basis functions At the decoder, given a known input array size, the corresponding set of basis functions may be precomputed and stored **Fig 3 4** illustrates an image representation of the basis functions of the 2-D DCT for an (8x8) block, which is the array size typically utilized in most image processing scenarios The upper-left corner component is the zero-frequency (or DC) basis function of the transform For a given block of transformed data, the

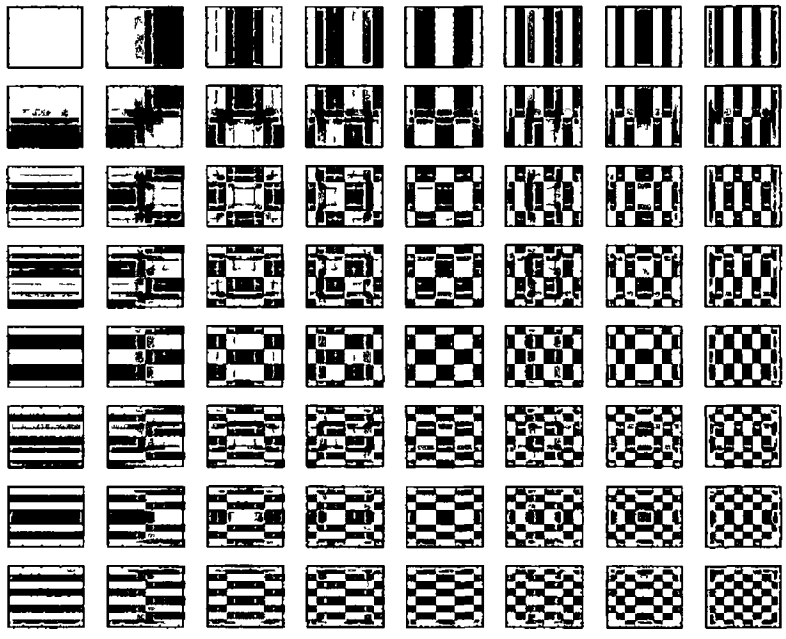


Fig 3 4 Plot of the basis images for an 8x8 2-D DCT

corresponding DC-DCT coefficient represents the weight or energy of this basis component, which corresponds to the mean overall intensity level of the original spatial block As illustrated, the remaining basis functions represent non- zero-frequency (AC) components, i e the rows and columns represent vertical and horizontal edges, respectively

While the DCT coefficients provide a mechanism for the reconstruction of images from the known set of basis functions, at the most basic level this is hardly significant, since the number of DCT coefficients produced equals the number of input pixels from the original array However in general, it may be shown that for natural images, most of the energy converges in the upper-left corner of DCT space (i e the low frequency DCT coefficients) Furthermore, the human visual system is more sensitive to reconstruction errors related to low spatial frequencies than to high frequencies, therefore the significance of the coefficients to the human eye decays with increased distance [64] Hence, this characteristic may be exploited towards compression To this end, since the higher DCT coefficients tend to be relatively less important, it is usually feasible to disregard them, and to rely purely on the subset of remaining significant components for block (image) reconstruction Hence this slight information loss should be either indiscernible or at least tolerable to the user

3 4 3 2 Predictive Encoding

Predictive encoding exploits redundancy in data that is temporal in nature This approach is most suitable in situations in which, during the evolution of data, the successive signal samples do not differ significantly within short time periods Clearly predictive encoding is very useful for video compression, where due to high video framerates, subsequent video frames differ very little from each other Hence individual pixel values differ little from frame-to-frame On the basis of this strong correlation between successive pixels, it is generally more economic to encode the difference between the pixel values rather than the values themselves, since these amounts will be smaller and hence require fewer bits There are varied approaches to this technique [65], which primarily differ in prediction generation The most common format is ***Differential Pulse Code Modulation*** **(DPCM)** In DPCM the prediction for a future value is based on that of the currently held value, and it is simply their variance that is encoded Therefore, if successive samples are sufficiently close to each other we only need to encode the first sample with a large number of bits, and the prediction with a relatively smaller number of bits Other variant schemes of predictive encoding include ***Delta Modulation*** **(DM)** and ***Adaptive Differential Pulse Code Modulation*** **(ADPCM)** [63]

3.5. MPEG-1 Compression

The ***Motion Picture Experts Group*** **(MPEG)** generate international standards for digital video and audio compression, and convene under the auspices of the International Standards Organisation (ISO) Due to their generic applicability, the MPEG standards have become the most popular in real world scenarios MPEG-1 is a finalized standard, which is presently being utilized in a large number of real world applications In essence, it is a technology for digitally coding audiovisual content for the purposes of storage The standard, also known as ISO/IEC 11172, builds, improves and generalises upon the earlier H 261 video telecommunications standard Specifically, the objective of MPEG-1 is to deliver digitised and compressed video signals at the maximum sustained data-transfer rate that could be handled by CD-ROM drives at the time of development, i e up to approximately 1 5Mbps

MPEG-1 is a standard in five parts [66] Part-1 (Systems ISO/IEC 11172-1 1993) addresses the problem of combining one or more data streams from the video

and audio parts of the MPEG-1 standard with timing information to form a single stream, i e multiplexing and synchronization of audio/video data Once combined into a single stream, the data is well suited to digital storage and transmission Part-2 (Video ISO/IEC 11172-2 1993) specifies a coded representation that can be used for compressing video sequences to the principal target bit-rate of 1 5Mbps However, since the approaches undertaken are generic in nature, the standard may be used more widely than the specified bitrate Part-3 (Audio ISO/IEC 11172-3 1993) specifies a coded representation that can be used for compressing audio sequences – both mono and stereo A psycho-acoustic model creates a set of data to control the quantifier and coding Part-4 (Conformance Testing ISO/IEC 11172-4 1995) specifies how tests can be designed to verify whether bit-streams and decoders meet the requirements as specified in parts 1, 2 and 3 of the standard Part-5 (Software Simulation ISO/IEC TR 11172-5) is technically not a standard, but rather a technical report It provides a full software implementation of the first three parts of the MPEG-1 standard The source code is not publicly available

The subsequent sections provide an overview of how Parts-2 and –3 (i e the video and audio compression processes) are realised However, as described above (i e under the banner of Part-1), once the audiovisual signals have been compressed, in practice the processed signals are time-stamped and interleaved, thus constituting a combined audiovisual stream, known as a ***system-layer*** MPEG-1 bitstream

It should be also noted that the MPEG group have successively developed many other related video standards, geared not only towards compression, but also content interaction and description The ***MPEG-2*** standard is a compression standard similar to MPEG-1 in that it is also based on motion compensated block-based transform coding techniques It was finalized in 1994, and addresses issues directly related to digital television broadcasting, e g the efficient coding of field-interlaced video and scalability In addition, the target bit-rate was raised to between 4 and 9 Mb/sec, resulting in potentially very high quality video ***MPEG-4***, which was finalized in 1998, targets very low bitrate applications It deviates from the more traditional approaches in its ability to independently encode individual objects present in the scene Further work has been focused on standardising a multimedia content description interface, i e ***MPEG-7***, and in developing a new standard called "A Multimedia Framework," also known as ***MPEG-21*** The abovementioned standards are not further

described since the work presented in this thesis utilizes MPEG-1 exclusively as the audiovisual representation

3.6. MPEG-1 Video Compression

3 6 1. Overview

As outlined earlier, video sequences exhibit substantial levels of redundancy, i e spatial, perceptual and temporal Spatial redundancy exists in images due to the typically high correlation between neighbouring pixels Perceptual redundancy is manifested in the limitations of the human visual system, in that data representing the fine detail of a given image may not be perceptible to the naked human eye Temporal-based redundancy is consequential of the high video framerates typically used That is, due to the rapid sequencing of images, there tends to be little difference between adjacent frames This is evident even for dynamic scenes involving substantial motion/activity Therefore, at the pixel level there is typically a high correlation between the successive value entries for the fixed pixel locations of the frames All of the standard video compression algorithms established to date exploit this tri-fold redundancy

3.6 2. Implementation

3 6 2 1 MPEG Colour-Space

Prior research into the perceptual quality of the human visual system [50] suggests that The human eye is inherently more sensitive to variations in luminance than to chrominance Hence, to increase the compression performance, the MPEG video algorithms (and the H 26x standards alike) exploit this characteristic in utilizing a colour space representation, i e YC_bC_r, which takes advantage of this perceptual trait Armed with this specialised colour space, the perceptual redundancy of the chrominance domain may be eliminated, independently of the luminance information To this end, the chrominance domain space is subsampled, while the luminance space remains unaltered A typical luminance/chrominance sampling ratio, which has been shown to be adequate for most practical scenarios, comprises four luminance pixels to a single twin colour-difference pixel - a scheme commonly known as ***mode-4 2 0*** Since mode-4 2 0 comprises one quarter of the chrominance information contained in a

corresponding full bandwidth RGB representation, the scheme yields lossy data compression

3 6 2 2 MPEG Video Structure

Each video frame in MPEG video is of one of three particular types The most basic of these are ***intra-coded (I-) frames***, which are video images that are coded independently, in a manner similar to that of the still image compression standard JPEG [67] However, encoding video with frame prediction yields much higher compression efficiency than that yielded by merely intra-coding all frames To this end, ***Predicted (P-) frames*** are encoded as pseudo-differences from the data comprising a prior frame ('forward referencing') Yet, forward-based prediction is limited in the sense that in many cases, the predicted frames could benefit significantly from reference information that is not evident in prior frames, but is in subsequent frames MPEG video addresses this issue by defining a third frame type ***Bi-directionally predicted (B-) frames*** are those predicted from data comprising both prior and subsequent frames (combined forward and backward referencing)

3 6 2 3 I-frame Coding

In encoding I-frames the spatial and perceptual redundancy contained in images is exploited The first step involved in implementing the image compression is the DCT transformation of the spatial data of the image Specifically, images are subdivided into regions of size [16x16] pixels, which are called **macroblocks** Thus, in mode-4 2 0 video, each macroblock is comprised of one [8x8] block for each of the colour difference signals (C_b, C_r), and four [8x8] blocks for the luminance component (Y) In the encoding process, each macroblocks constituent [8x8] blocks are transformed via the DCT Following this, a quantization process is performed, which aims to retain only the most significant bits of the DCT coefficients While quantization error is the main source of the data loss, it is proposed that the degradation to the content following this process should be reasonably indiscernible to the viewer Subsequent to quantisation, a zig-zag scan of DCT space is performed such that in mapping two-dimensional (8x8) space to a one-dimensional (1x64) vector, the low-frequency coefficients, which are of most significance to the human eye, are collectively grouped towards the top That is, they occupy the most significant vector positions - see **Fig 3 5** [51] While the DC (zero-frequency) DCT coefficients are large and varied for most images, neighbouring

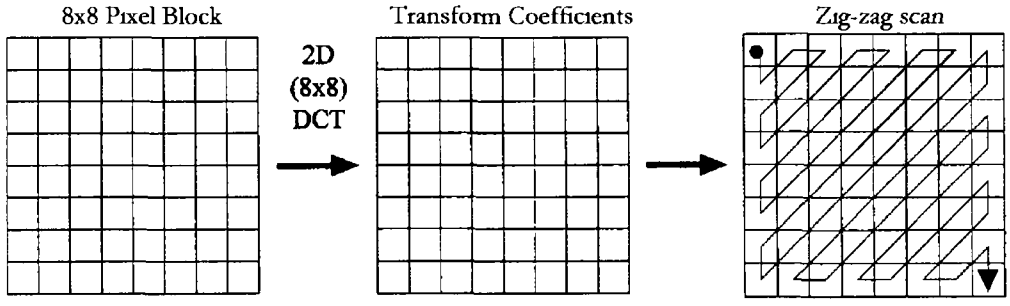


Fig 3 5 Zig-zag scanning of 2D (8x8) DCT coefficients

values are often close in value Thus DPCM is applied to the DC-DCT coefficients, such that only the difference from previous 8x8 blocks is encoded Following this, for each block, RLE is applied to the AC-DCT coefficients, since it is not uncommon for a 1x64 vector to contain many zeros Finally the data is then entropy encoded such that the DCT coefficients are represented by an even smaller number of bits

I-frames are independently coded from any other frame in the MPEG video sequence Therefore, while spatial and perceptual redundancy is eliminated, the fact that each I-frame is encoded in isolation, implies that there is not great efficiency achieved in exclusively intra-encoding frames Nevertheless, I-frames are very important elements of the MPEG video stream, since they are used as reference frames for the prediction techniques employed by other frame types Furthermore, their occurrence in the video sequence facilitates random access points within the encoded video stream Overall, the frequency of occurrence of I-frames within the video sequence represents a trade-off between compression intensity and error propagation

3 6 2 4 P-Frame Coding

As well as exploiting spatial and perceptual redundancy, P-frame encoding involves the elimination of temporal redundancy in video, via a process called inter-frame coding Given an encoded I-frame, the encoder estimates or predicts a future frame, i e a P-frame, which in turn, may then also be used as a reference in predicting further P-frames in a forward manner - see **Fig 3 6** [65] In implementing this technique, the target image is subtracted from the reference image, yielding a prediction residual Given the reference frame, this residual frame is further compressed as in the case of I-frames, i e via the quantisation of its equivalent DCT coefficients This data is then coupled with

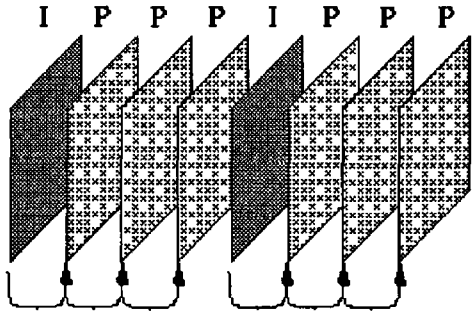


Fig 3 6 Inter-frame coding in video sequences

the information required to reconstruct the prediction, and the ensemble encoded accordingly The addition of temporal-based compression yields a much higher video encoding efficiency than that of I-frames, since the prediction residual requires fewer bits for representation than an independently encoded image

The inter-frame coding process may be further enhanced by a technique called ***Motion Estimation*** (ME), based on the argument that successive video frames are generally very similar, except for small variations produced by the movement of objects within frames, plus movement of the camera itself ME is usually implemented as a pixel-block matching technique, the objective of which is to gauge the motion between reference and target frames, prior to the generation of a frame residual This estimated motion is then subsequently 'undone' (compensated for) in generating a more efficient prediction[#] Typically, a comprehensive two-dimensional spatial search is performed for each luminance domain macroblock Once an adequate match has been located, the encoder assigns ***motion vectors*** (**MVs**) to the macroblock, which describe the direction and distance of the displacement in 2-D Note that the search algorithm is not employed in the chrominance domain, since it is assumed that the colour motion may be sufficiently represented from the motion estimated in the luminance space Clearly, not every search results in an acceptable macroblock match If the encoder decides that no acceptable match exists then the option of coding that particular macroblock as a standalone intra-coded macroblock may be instigated In doing so, high image quality may be sustained at a minor cost in compression efficiency In practice, the MV data is then tagged with the DCT information of the residual frame, and encoded using

[#] There exists a wide range of motion analysis techniques, i e optic-flow, polynomial motion modelling, etc However, the description of those other than that characteristic of a typical MPEG encoder are outside the scope of this thesis

variable length codes During frame reconstruction, MVs designate the extent to which the macroblock in the reference frame must be shifted in the horizontal and vertical planes, at the decoder's ***motion compensation*** stage

3 6 2 5 B-Frame Coding

B-frames are encoded based on a forward prediction from a preceding I/P-frame, as well as a backward prediction from a succeeding I/P-frame When backward prediction is used, this requires that the future reference frames are encoded and transmitted first, out of display order **Fig 3 7** [65] illustrates the frame referencing scheme between I/P/B-frames in MPEG video B-frames employ motion estimation/compensation either from a preceding or succeeding frame, or both Therefore the macroblock matching search takes place in both past and future frames As a result, each forward and backward predicted macroblock may contain two MVs, so true bi-directionally predicted macroblocks will utilize four MVs B-frames achieve the highest compression efficiency However, a further advantage is that backward prediction allows the encoder to make more intelligent decisions regarding video content, e g when moving objects reveal hidden areas during the sequence

3 6 2 6 Group of Pictures

A ***group of pictures*** (**GOP**) is a set of pictures that are in continuous display order in a video sequence It begins with an I/B-frame and ends with an I/P-frame A GOP example is illustrated in Fig 3 7 A GOP may be deemed open or closed, depending on whether inter-GOP ME frame referencing is permitted The smallest GOP size is a single I-frame While there is no upper limit in GOP length, typical MPEG-1 GOP lengths are between 10-18 frames The encoder makes the decision on how often the

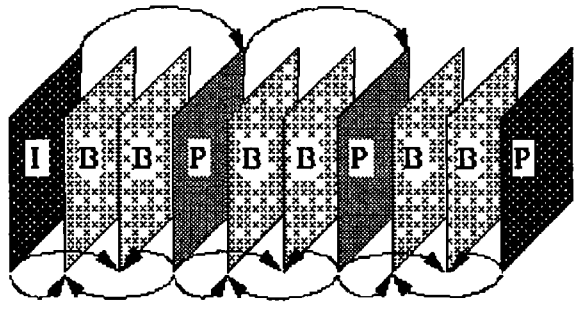


Fig 3 7 Referencing between I-, P- and B-frames in MPEG video

different picture types occur and if a high compression ratio is required, then many B-frames will be used However, most broadcast quality applications tend to use two consecutive B-frames as the ideal trade-off between compression efficiency and video quality – as illustrated in Fig 3 7

3.7. MPEG Audio Compression

3.7.1 Overview

As in the case of video, audio sequences benefit most from lossy compression, as the lossless-based techniques tend not to yield much gain in terms of compactness ratio For example, ADPCM may be used in exploiting the temporal redundancy between successive audio samples Specifically, the encoding scheme targets the difference between consecutive audio samples, and adapts the quantization such that fewer bits are used when the value is smaller, thus yielding compression

However, a more obscure type of audio redundancy exists, which corresponds to the psycho-acoustic perceptual properties of the human audio sensory system [68] In essence, the human ear exhibits a frequency masking property, whereby the presence of one frequency component can mask the perception of another nearby component, in both time and frequency It is accepted that this characteristic is a form of audio redundancy That is, if it is possible to accurately discern which components have a high probability of being masked, then compression may be achieved by discarding these, without a noticeable detriment in the perception of the signal This psycho-acoustic redundancy forms the primary source of compression in MPEG audio encoding

3 7 2 Implementation

MPEG audio compression is defined in three layers For each layer the basic model is the same, however the scheme complexity increases accordingly In Layer-I, a filter bank is employed to decompose the frequency spectrum of the audio signal into thirty-two equally spaced subbands, which approximate the ear's critical bands The subbands are subsequently assigned individually weighted bit-allocations according to the audibility of quantisation noise within each subband A psychoacoustic model of the ear analyses the audio signal and provides this information to the quantiser However, the audio data is firstly segmented into frames of length 384 samples, i e 12 samples from each of the 32

subbands Each group of 12 samples gets a bit allocation and, if this is non-zero, a scalefactor ***Scalefactors*** are weights that normalize groups of audio samples such that they use the full range of the quantiser The scalefactor for such a group is determined by the next largest value (given in a look-up table) to the maximum of the absolute values of the sample group

In Layer-II, the psychoacoustic analysis attempts to model temporal frequency masking as well as static masking To this end, Layer-II analyses three Layer-I-sized frames at a time in the filtering process, which correspond to previous, current, and subsequent frames Therefore, Layer-II frames consist of 1152 samples, 3 groups of 12 samples from each of 32 subbands, corresponding to 36 (3x12) samples per subband (or 12 ***granules*** per subband as shown in **Fig 3 8** [51]) In this scenario, the encoder uses a different scalefactor for each of the three groups of 12 samples within each subband only if necessary The complete Layer-II data bitstream structure is illustrated in **Fig 3 9** [51]

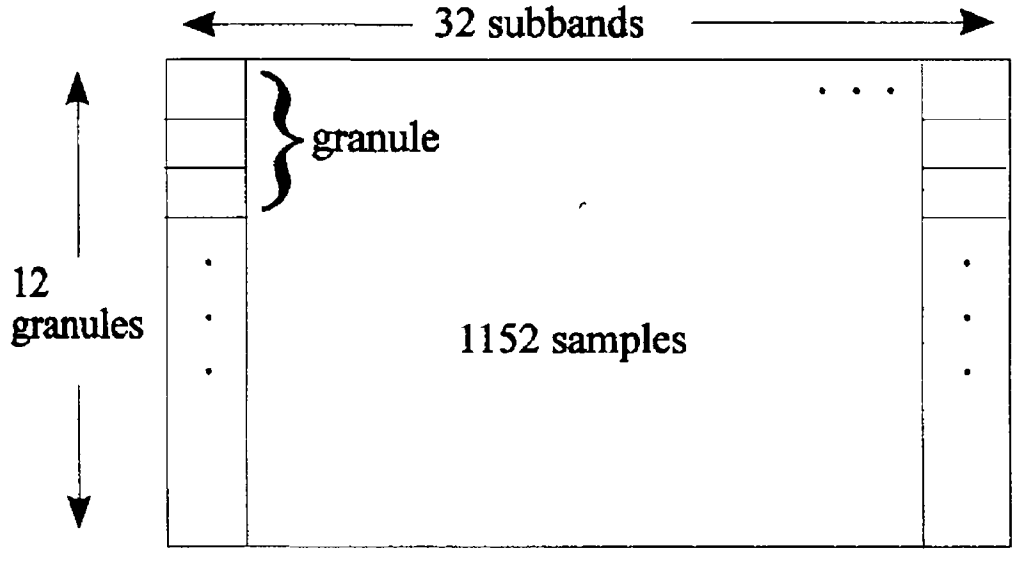


Fig 3 8 Structure of Layer-II subband samples

Data (Layer II)	Bit Allocation (2 ~ 4 bits)	Scale factor Select Information (2 bits)	Scale Factor (6 bits)	Samples (2 ~ 16 bits)	Ancillary

Fig 3 9 The data bitstream structure of Layer-II

Layer-III exhibits non-uniform frequency band division, which improves the approximation to the ear's critical bands It also exploits stereo redundancy, and an entropy encoding mechanism is utilized

3.8. Chapter Summary

In this chapter an introduction to the basic principles of digital video was provided towards facilitating a more complete understanding of the concepts exploited in subsequent chapters Initially, a description of the appropriate colour-space formats was provided, explaining why in the field of video encoding, a luminance independent format is more favourable to the basic RGB representation Next, an introduction to the standard video structure hierarchy was presented, including definitions of video concepts such as pixels, frames, shots, scenes, etc This was then followed by a discourse on the various methodologies for data compression, with particular emphasis placed on those pertinent to video coding Given this background, the audiovisual compression technologies specifically underpinning the MPEG-1 video-encoding standard (the representation used in this work) were then described

Chapter 4

A Hypothesis for the Generic Summarization of Field-Sports-Video

As introduced in *Chapter 1*, the requirement of a genre-independent solution to the problem of event detection-based sports-video summarization represents the primary motivation for the work undertaken in this thesis However, it has been explained that in approaching the development of this, at least some restriction in genre scope is necessary, such that the underlying event concept definitions (as well as the general aspects of the games) may remain robust throughout the domain of operability Given this, a compromised scenario was proposed whereby characteristically similar sports genres are convened under the ambit of a 'supergenre' It is anticipated that this approach should provide for event concept definitions that are robust across constituent genres, such that the supergenre may be treated as a single entity in relation to the summarisation task

As explained in *Section 1 5 1*, as a target case study, the specific research objective of this thesis is to develop a generic solution for event detection-based summarization in the field-sports-video (FSV) supergenre A framework overview of the approach to be taken towards realising this objective was presented in *Section 1 5 2*, however, this chapter presents a complete account of the overall hypothesis via which it is proposed this objective may be accomplished It begins by describing an investigation into the characteristics that describe a field-sport Given this, and then given the assumption that the narrative of field-sport games may be sufficiently represented by the score-update episodes (SUEs), an investigation into determining what features

consistently characterise these events across all field-sport broadcasts is then described, which is based on observations inferred across the five FSV genres constituting the training corpus Given a set of features deemed critical to the indication of SUEs, hypotheses for the frame-level detection/quantification of such in FSV content are then described, based on exploiting relevant signal-level attributes of the content Reasons for employing a pre-processing filter stage are then proposed This is then followed by a description of how frame-level critical feature evidence might be aggregated at the shot-level, towards generating a critical shot-level description of the content, upon which it is envisaged that SUE-shots might be discernable

4.1. Field-Sports-Video Summarisation

4.1 1. The Boundaries Of The Field-Sports-Video Supergenre

Given the requirement of a summarisation solution that is generically operable throughout the FSV supergenre, as explained in *Section 1 5 2 2*, it is clearly necessary to explicitly specify the bounds of what is meant by the 'field-sports' description Recall that the data corpus obtained is comprised of the following genres, soccer, rugby, Gaelic football, field hockey, and hurling (see *Section 1 5 2 1*) It was required to determine what are the qualities that link these sports Following an observation of the abovementioned training-corpus genres (coupled with a limited exposure to the other field-sport genres of Table 1 1 not represented, i e Australian rules football and American football), it was recognised that field-sports in general are linked by the fact that they each exhibit an intrinsic set of common characteristics These are as follows,

(i) Two opposing teams + referee(s)
(ii) Enclosed playing area
(iii) Grass pitch
(iv) Field lines
(v) Commentator voice-over
(vi) Spectators
(vii) On screen video-text graphics (scoreboard)
(viii) Three well-defined styles of camera shot global (main), zoom-in and extreme close-up

(ix) Game objectives concerned with territorial advancement, and directing an object (e g ball) towards a specific target

(x) Score tally

Clearly some of these features may be found in sports genres that are not listed above However, what is significant is that all ten features are exhibited in the abovementioned genres On this basis, it is thus proposed that these criteria are both necessary and sufficient in characterising a FSV, i e they define the boundaries of the FSV supergenre In terms of developing the hypothesis for the FSV summarization task, the corresponding challenge is that any derived solution should thus operate with consistent performance for any sports genre exhibiting all ten of these features

4 1.2 The Summarisation Methodology (Narrative-Critical Events)

As explained in *Section 1 5 2 3*, in terms of the adopted FSV summarisation methodology, it is SUEs alone that are targeted for detection That is, although it is not uncommon for interesting events to contribute to the narrative of a field-sport game that are non-score related, it is recognised that, in general, SUEs represent the most objectively critical elements of the narrative, and therefore their detection alone should provide for a favourable summarisation solution

Examples of SUEs for several field-sports genres are listed in **Table 4 1** From this inventory it is evident that SUEs exhibit many guises across the spectrum of FSV genres Hence, it was recognized that obtaining a generic solution for SUE detection would require the development of a hypothesis that exploits what is common to all scenarios, as opposed to what individually defines them

Table 4 1 Field-sports genres and corresponding score-update episodes

Field-Sport Genre	Score Update Episodes
Soccer	Goal
Hockey	Goal
Rugby	Try, Conversion, Drop-goal, Penalty kick
Hurling	Goal, Point
American Football	Touchdown, Conversion, Field Goal,
Gaelic Football	Goal, Point
Australian Rules	Goal, Behind

4.2. Score-Update Episode Characteristics

Given the different guises of SUEs across the different genres, it was required to investigate what features are most apparent in generically characterizing SUEs in FSV content, with a view that their combined detection/quantification might form the crux of a general SUE identification hypothesis To this end, SUEs were surveyed from the training-corpus in equal proportions from each individual training-corpus genre Given that there exists many circumstances in which SUEs may be manifested, the SUE characteristics probed did not relate to individual scenarios, but rather related to modeling what was common to all situations, irrespective of circumstance

4.2 1. Action Replays

At the outset, the most immediately obvious SUE-related characteristic concerned the high probability that they are followed by an action replay segment Towards gauging the extent of this, evidence pertaining to this characteristic was acquired from across all genres of the training-corpus and is listed in **Table 4 2** It was observed from this data that the cross-genre variance was small and that, on average, 97% of all training-corpus SUEs were followed by an action replay This phenomenon suggested that by simply locating replay segments within the content, SUEs could be retrieved with high statistical recall accuracy However, it was also observed that replay segments are highly prevalent throughout FSV content whether SUEs occur or not Therefore, it was concluded that the precision accuracy offered by employing this retrieval methodology alone would be unsatisfactory Moreover, it was recognized that the detection of action replays remains a challenging aspect of sports-video processing, especially given a genre-independent domain requirement Recall from the literature review in *Chapter 2* that the

Table 4 2 Percentage of training corpus SUEs followed by action replays

Field-Sport Genre	% SUEs Followed By Action Replay
Soccer	100%
Gaelic Football	95%
Rugby	97%
Hurling	94%
Hockey	100%
	Average = 97%

classical approach to action replay detection is based on two assumptions [37] The first of these is that the replay segments exhibit slow motion playback, and then secondly, that the mechanisms implementing slow motion, are based on video frame repetition and/or drops However, increasingly, high-speed camera technology is prominent in live sports broadcasting Consequently, the conventional techniques, such as frame repetition, have been replaced by more sophisticated variable-speed playback solutions Thus, the classical assumptions breakdown, and the schemes are therefore vulnerable to failure Other methodologies, e g [69], attempt to detect replay segments based on spatial domain algorithms, which detect digital video graphical effects While such analyses may be genre-independent, they tend not to be generic, in the sense that they are dependent on the characteristics of a particular broadcaster Thus, while the topic of action replay detection has already attracted some research attention, there are evidently some aspects that remain unsolved in terms of a broadcast/genre-independent framework, the challenges of which serve to discourage further pursuit of this approach in developing this work

4 2 2 Reaction-Phase

Given the high coincidence between SUEs and action replay segments as described, a further consistent feature observed from the training-corpus, was the play-break lag time that immediately follows a SUE before the cut to replay It was found that, in the main, the programme director utilizes this '***reaction-phase***' (**RP**) segment to capture the responses of players and/or crowds to the significance of the event that just occurred Furthermore, it was noted that in direct response to this significance, the RP segments tend to exhibit several prominent characteristics (the details of which are expounded below) Given these, it was proposed that the prevalence of the observed RP features may be exploited towards the development of a SUE model hypothesis To facilitate an investigation of this, an analysis into the attributes of training-corpus RP segments was performed **Fig 4 1** illustrates the distribution of RP durations across an equal number of SUEs extracted from the field-sports genres of the training-corpus From this distribution it is clear that the mode RP duration is in the range 15s-16s, corresponding to approximately 15% of all examined cases However, more significantly, it is evident that a negligible amount of RP durations are in excess of 24s It is thus proposed that this 24s upper limit constitutes a post-SUE ***reaction-phase seek-window*** (**RPSW**), i e specifying an appropriate temporal domain for the probing

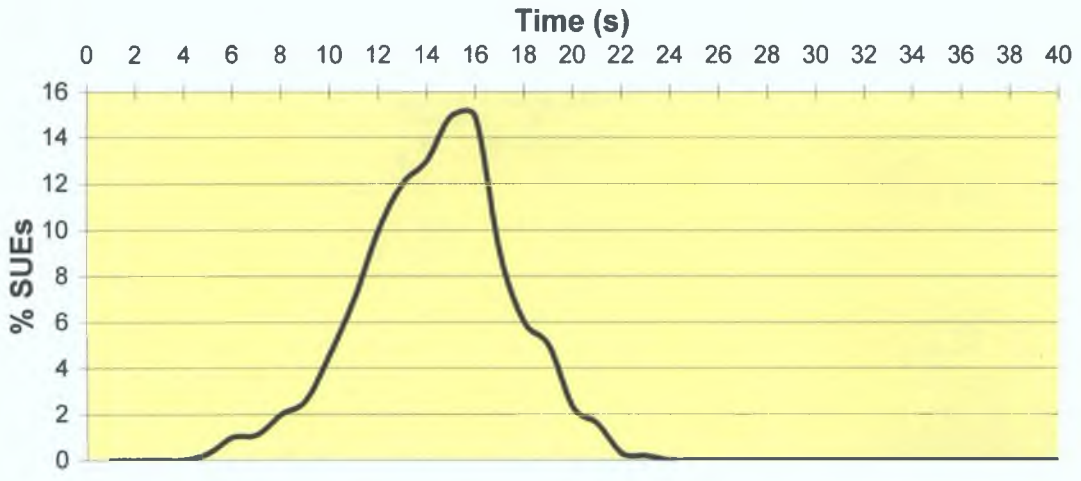


Fig. 4.1. Distribution of SUE reaction-phase durations across all field-sport genres within the training-corpus.

of RP-orientated characteristics. Given this, a detailed description of the observed RP features follows, which is coupled with a manual training-corpus quantification of the potency of each during the RPSW.

4.2.2.1. Close-Up & Crowd Views

As mentioned above, following an SUE occurrence in FSV, the director typically endeavours to convey the immediate reactions of the relevant parties to the viewer. Consequently, frequent interspersions between player close-up-view-shots and crowd-view-shots were found to be prevalent during SUE RPs. To explicitly quantify their incidences within these segments, it was manually ascertained (across all field-sports genres of the training-corpus) exactly what ratio of the SUEs exhibited (i) close-up image sequences, and (ii) crowd image sequences within their respective RPSWs. **Tables 4.3** and **4.4** list the results of this manual investigation. From this data it is evident that, on average, approximately 98% of all training-corpus SUEs exhibit a close-up image sequence within the specified timeframe, and likewise approximately 71% exhibit a crowd image sequence.

4.2.2.2. Visual Activity

It was observed that a consequence of the prevalence of close-up views during SUE RPs was that these segments were characterized by increased visual activity, a phenomenon that tended to be further accentuated by the typically celebratory

Table 4 3 Percentage of SUE-RPSWs exhibiting close-up image sequences

Field-Sports Genre	% SUE-RPSWs containing Close-Up Sequences
Soccer	100%
Gaelic Football	97 5%
Rugby	96 4%
Hurling	96 2%
Hockey	98 4%
	Average = 97 7%

Table 4 4 Percentage of SUE-RPSWs exhibiting crowd image sequences

Field-Sports Genre	% SUE-RPSWs containing Crowd Sequences
Soccer	54 2%
Gaelic Football	79 1%
Rugby	73 1%
Hurling	79 5%
Hockey	71 2%
	Average = 71 4%

behaviour of the scoring player Further sources of increased post-SUE visual activity were found to correspond to the use of zoomed-in/close-up views during the action replay segments, and the use of video effects in delimiting their multiple viewing angles Again, it was considered desirable to explicitly quantify this feature within the training-corpus data To this end, it was determined exactly what ratio of all training-corpus SUE-RPSWs exhibited peak near-field motion activity measures in excess of their respective broadcast mean levels[1] **Table 4 5** lists the results of this investigation From

Table 4 5 Percentage of SUE-RPSWs exhibiting near-field motion activity surges

Field-Sports Genre	% SUE-RPSWs exhibiting motion activity surges
Soccer	96 2%
Gaelic Football	85 6%
Rugby	91 0%
Hurling	83 1%
Hockey	93 5%
	Average = 89 8%

[1] An automatic visual activity quantification tool was used to facilitate this measurement procedure, the details of which will be formally introduced at a later stage

this data it was observed that, on average, approximately 90% of all observed cases exhibited this trait

4 2 2 3 Audio Activity

Another consistent characteristic observed was the perceptible increase in audio activity that tends to characterize RP segments in FSV content That is, it was found that, in direct response to the significance of SUEs, there tends to be a prominent surge in audio level, which is generally attributable to the energy dynamics of the commentator voice over and that of the cheering spectators To explicitly quantify this it was determined exactly what ratio of all training-corpus SUE-RPSWs exhibited audio track levels in excess of their respective broadcast mean levels[2] **Table 4 6** lists the results of this investigation From this data it was observed that, on average, 85% of all cases exhibited this trait

Table 4 6 Percentage of post-SUE RPSWs exhibiting audio energy peaks

Field-Sports Genre	% SUE-RPSWs exhibiting audio energy peaks
Soccer	94 7%
Gaelic Football	80 0%
Rugby	85 3%
Hurling	76 3%
Hockey	90 8%
	Average = 85 4%

4 2 2 4 Scoreboard Graphic

Finally, for many of the training-corpus field-sport broadcasts, it was found that it was not uncommon for the on-screen scoreboard graphic to be temporarily suppressed during its update procedure Moreover, it was found that in such circumstances, the scoreboard suppression was most frequently apparent during the RP segments Again was considered desirable to explicitly quantify this correlation Hence, it was manually determined exactly what ratio of training-corpus SUE-RPSWs exhibited a temporary suppression of the on-screen scoreboard graphic **Table 4 7** lists the results of this

[2] An automatic audio energy quantification tool was used to facilitate this measurement procedure, the details of which will be formally introduced at a later stage

Table 4 7 Percentage of SUE-RPSWs exhibiting scoreboard suppression

Field-Sports Genre	% SUE-RPSWs exhibiting scoreboard suppression
Soccer	79 5%
Gaelic Football	34 2%
Rugby	96 8%
Hurling	30 7%
Hockey	63 9%
	Average = 61 0%

investigation, and from this data it was concluded that, on average, 61% of cases adhered to this paradigm

4.2.3. Field End-Zone Activity

As described above, the SUE RP segments exhibit several characteristics, which suggested a basis for the development of a SUE model hypothesis However, a further potential SUE indicative feature was also observed that differs from those already mentioned in that it does not relate to characteristics of the post-SUE RP segments Specifically, it corresponds to the typical in-field location of SUE activity Recall that feature (ix) in *Section 4 1 1* alludes to fact that the objective of FSV games is concerned with territorial advancement, and with directing an object (e g ball) towards a specific target It is clear that all field-sport SUEs obey this paradigm For example, the SUEs referenced in Table 4 1, i e *goals*, *tries*, *points*, *conversions* etc, are achieved either by (i) directing the ball towards a target in the field end-zone, or (ii) player, with ball, advancing towards the end-zone On this basis, it was observed that as such events unfold, it is typical for the camera following the action to use a global view perspective and be focused on the end-zone region of the playing field SUE scenarios contradicting this paradigm included placed kicks, where the camera assumes an abnormal view (e g behind the target) Once again, it was considered desirable to explicitly quantify the prevalence of this phenomenon for the training-corpus content, and to this end **Table 4 8** presents the results of a manual investigation From this data it is evident that, on average, 74% of all training corpus SUEs adhered to the circumstances described

Table 4 8 Percentage of SUEs occurring with camera in global view and focused on field end-zone region

Field-Sports Genre	% SUEs with camera focused on field end-zone
Soccer	85 6%
Gaelic Football	71 7%
Rugby	61 5%
Hurling	69 4%
Hockey	82 1%
	Average = 74 1%

4.3. Score-Update Episode Shot Model

Inferred from a manual training-corpus investigation, the previous section describes several features whose occurrences exhibit a high correlation with that of SUEs, which are consistent across multiple FSV genres Specifically, five features corresponding to the (pre-action replay) 'reaction-phase' segments were documented, as well as the association with field end-zone activity Although the statistics were recorded in terms of SUE coincidence as opposed to (the more powerful aspect of) SUE discriminance, given the high values recorded, it is proposed that the combined detection/quantification of these characteristics should provide a reliable basis for the automatic identification of SUEs in FSV That is, while it was noted that it was not uncommon for any of the aforementioned features to occur sporadically throughout any genre of FSV content (the mark-up task of which would be hugely time-consuming), the assumption is that given the correlation statistics recorded, it is when some of these features are found occurring within close proximity of each other, it may be concluded that the SUE occurrence probability has increased

The proposed SUE model hypothesis exploits the above assumption by applying appropriately restricted temporal probing domains for the detection of the aforementioned features - hereafter known as the ***critical features*** **(CFs)** That is, it is suggested that the locations of SUEs in FSV content may be discerned based on the quantification of the sustained prevalence and/or intensity of the CFs within relevant temporal seek-windows Specifically, while it is not uncommon for the build up of SUEs to occur over more than one shot, the shots immediately preceding the RP segments are generally the most vital to the conveyance of the event narrative (hereafter known as

SUE-shots). Therefore the detection of SUE-shots should provide for SUE retrieval at a sufficient level. Towards the detection of such, it is proposed that, for a given shot, the prevalence/intensity of the RP-orientated features be quantified within the RPSW immediately following its end-boundary. In addition, to quantify the probability that a given shot culminates in a SUE, it is proposed that the prevalence of the field end-zone feature be quantified towards the shot-end-bound (i.e. within some appropriate shot-end seek-window to be specified at a later stage). Given the statistics observed from the training-corpus investigation, shots directly prior to SUE RP segments should be discernable from others on the basis that they will tend to exhibit a significantly higher prevalence/intensity of CFs within their respective seek-windows. This represents the proposed hypothesis for the generic detection of SUE-shots in FSV content as illustrated in **Fig. 4.2**.

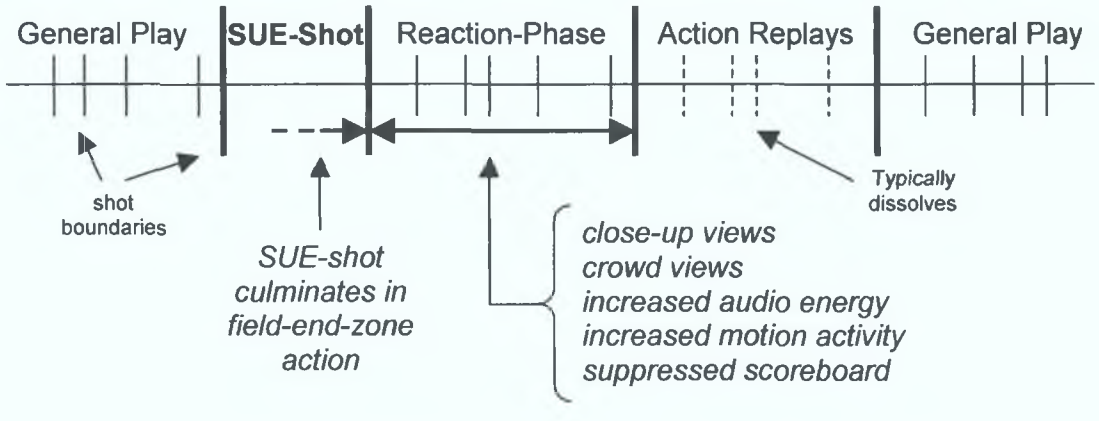


Fig. 4.2. Model hypothesis for the detection of SUEs in FSV.

4.4. Frame-Level Critical Feature Extraction

In the previous section a hypothesis for the detection of SUEs in FSV was proposed based on the extraction of evidence pertaining to a set of six high-level critical features (CFs), the justification of which was inferred from observed training-corpus statistics. In short, the CFs correspond to close-up views, crowd views, scoreboard suppression, field end-zone activity, and the quantification of motion activity and audio energy. In this section frame-level extraction methodologies are proposed for these CFs based on the

exploitation of key signal-level audiovisual data However, it should be noted that the discovery of the correlation between such features and scoring events in sports-video content is not a novel observation In fact, these feature types have been commonly exploited in the development of most prior art schemes (see *Chapter 2*) For example, the correlation between sports-video highlights and increased audio energy is argued in [30], [32], [42], [43], [47], and [48], that of close-up views in [15], [20], [27], [34], [40], [45], crowd views in [40] and [45], and that of on-screen graphics (scoreboard) activity in [16], [2], and [33] Furthermore, in [17], [19], [22], [23], and [27], various ways are described of exploiting knowledge inferred by the tracking of field-lines towards extracting semantic concepts from sports-videos Likewise, in [18], [23], [24], [25], [26], [28], [2], [32], [39], [43], [45], [47], and [48], where motion dynamics and/or the quantification of visual activity in general is shown to be exploitable towards realizing a variety of event detection tasks in sports-video content While most of these existing critical feature extraction methodologies alluded to above have been shown to be useful in fulfilling their purpose within the overall scheme objectives specified in each case, in terms of design and implementation, many originate from a genre-specific disposition (i e of or relating to the genre-specific schemes described in *Section 2 2*) Hence, in exploiting such features in terms of the development of the generic field-sports scheme herein, it was decided to design original extraction methodologies for such, in order to ensure reliable and consistent responses across all sports-genres within the remit of the field-sport domain

4.4.1 CF1 Close-Up Image Detection

The first critical feature (CF1) corresponds to the detection of close-up images To this end, a colour-based approach is proposed Although chrominance-based classification may not be practical in many video scenarios, it is suitable for FSV, where colours are purposely used to differentiate players, and clearly defined rules constrain the action [70] As a result, the colours of the objects concerned, such as the playing surface, players/referee shirts, etc, usually consist of one or two (striped) dominant colours, as illustrated in **Fig 4 3** On this basis, it is proposed that, given a video frame, a close-up view confidence value may be derived via an analysis of pixel hue evidence (Note as explained in *Section 3 2 3*, the analysis of the HSV colour-space is preferred over others since it is in the hue space where colours that are perceptively similar tend to cluster best)

4.4.1.1. Close-Up Image Characteristics

Within the specific domain of FSV content, a close-up image is defined as a zoomed-in view, which principally displays a player's head and shoulders. Two such images (A and B) are displayed in **Fig. 4.4**. From these examples it is evident that the salient characteristics of close-up images are (i) the presence of a face in the top-middle-centre region (i.e. the focus) of the frame, together with (ii) a jersey in the bottom-middle region of the frame (occluding an arbitrary background). It is the combined potency of these two critical characteristics, which forms the basis of the detection hypothesis for FSV close-up views.

Fig. 4.3. A field-sports-video image. Within this image the acute dominant-colour differention between players, referee and playing field is apparent.

Fig. 4.4. Two close-up image samples.

4 4 1 2 Close-up Image Modeling

The proposed close-up modeling approach is based on bounding the two abovementioned characteristics, via the segmentation of video frames into ***regions of expectancy*** (**ROE**) The approximate positioning of these inferred regions are illustrated in **Fig 4 5**, where the specific positioning/dimensions are left to be determined at the implementation stage In this segmentation, ***Region-1*** (**R1**) corresponds to the estimated region of expectancy for the location of the player's face in a close-up image ***Region-2*** (R2) is the estimated ROE for the location of the player's jersey in a close-up view Finally, ***Region-3*** (R3) corresponds to the ROE for the image background

As described, it is desirable that a confidence measure be computed for a given video frame, the value of which infers the probability of the image representing a close-up view On the basis of the salient characteristics discussed, and the corresponding ROE outlined, it is proposed that in modeling close-up views, the ***close-up confidence*** (**CuC**) value should represent the degree to which the image exhibits both of the following attributes

(i) a skin-toned entity in R1 (i e indicating a face)

(ii) a dominant colour in R2, not so dominant in R3 (i e indicating a jersey occluding an arbitrary background)

In the field of computer vision, it is a commonly held argument that skin-colour clusters well in the hue space, i e in [71] it is explicitly illustrated that irrespective of race or

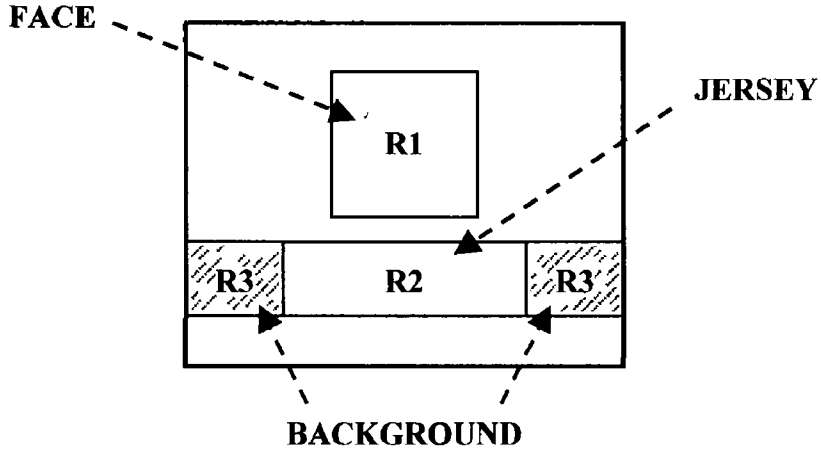


Fig 4 5 Approximate regions of expectancy for face, jersey, and occluded background for generic close-up image

nationality, the majority of skin pixels reside in the hue interval [10°-55°] Therefore, it is proposed that via the analysis of low-level pixel hue data, skin-toned pixels may be discriminated from other pixel colours in video images Hence, to quantify attribute (i) above, it is proposed that values for the skin-hue pixel ratios (SHPRs) be calculated within R1 These values ($\mathbf{SHPR_{R1}}$) may be computed using **(4 1)** below, where 'SkinPixels' correspond to those that exhibit hue within the critical interval [10°-55°]

$$SHPR_{R1} = \frac{\#\,SkinPixels_{R1}}{\#\,Pixels_{R1}} \qquad \textbf{(4 1)}$$

In exploiting the strong colours present in players' jerseys, the first part of attribute (ii) above concerns the degree to which R2 is dominated by one colour To quantify this, it is proposed that low-level pixel hue data should again prove useful in ascertaining the dominant hue that R2 exhibits ($\mathbf{DH^{R2}}$), and then in determining its corresponding overall level of mono-chromaticity Once $\mathbf{DH^{R2}}$ has been determined, a value for the dominant-hue pixel ratio ($\mathbf{DH^{R2}PR_{R2}}$) may be generated using **(4 2)** below This value represents the overall level of mono-chromaticity for the region In this formula a 'DomHuePixel' corresponds to one that exhibits hue within the interval [$\mathbf{DH^{R2}} \pm \xi$], where ξ is a specified tolerance variable (to be specified at the implementation stage)

$$DH^{R2}PR_{R2} = \frac{\#\,DomHuePixels_{R2}}{\#\,Pixels_{R2}} \qquad \textbf{(4 2)}$$

The second part of attribute (ii) concerns the extent to which this mono-chromaticity is bound to region R2, i e not found within R3 To quantify this, it is proposed that the degree to which the dominant hue is not prevalent in regions R3 is measured That is, values for the $\mathbf{DH^{R2}}$ pixel ratios are calculated for R3 These values, $\mathbf{DH^{R2}PR_{R3}}$, are computed using **(4 3)** below, where again, a 'DomHuePixel' corresponds to one that exhibits hue in the interval [$\mathbf{DH^{R2}} \pm \xi$]

$$DH^{R2}PR_{R3} = \frac{\#\,DomHue^{R2}Pixels_{R3}}{\#\,Pixels_{R3}} \qquad \textbf{(4 3)}$$

Clearly, an ideal close-up image would exhibit the player's face and jersey perfectly encapsulated in the appropriate ROE In this ideal case, both $\mathbf{SHPR_{R2}}$ & $\mathbf{DH^{R2}PR_{R2}}$ would be expected to have relatively large values, while the descriptor $\mathbf{DH^{R2}PR_{R3}}$ should be relatively small These characteristics were exploited in developing the

arithmetic for the formulation of a close-up confidence (**CuC**) measure, which is defined in **(4 4)** It is expected that within the limited image domain context of FSV, this scheme should work well in generating confidence values that facilitate the reliable detection of close-up images

$$CuC = SHPR_{R1} * (DH^{R2}PR_{R2} - DH^{R2}PR_{R3}) \qquad \textbf{(4 4)}$$

4.4.2. CF2: Crowd Image Detection

The second critical feature (CF2) corresponds to the detection of crowd view images To this end, a texture-based approach is proposed, and as with CF1 it is required that for a given image, a confidence value be generated

4 4 2 1 Crowd Image Characteristics

It is recognized that the classification of crowd images based on texture characteristics alone may not be feasible for generic video However, as noted in *Section 4 1 1*, one of the defining characteristics of FSV content is that, in general, it is constrained to three well-defined camera views Consequently, within this limited context, the majority of video images tend to capture relatively sizeable, monochromatic, homogeneous regions (e g grassy pitch, player's shirts, etc) On the contrary, crowd images tend to be relatively more complex in terms of scene detail, i e exhibiting a large collection of small heterogeneous objects (spectators) These differing traits are illustrated in **Fig 4 6**, in which a series of generic FSV images from the three standard camera perspectives are presented with sample crowd image views On this basis, it is proposed that the required confidence values may be derived purely via an analysis of image texture

A crowd image is hereafter defined as a camera view that principally displays approximately 20 or more spectators simultaneously within a single frame As illustrated above, compared to other images in FSV content, crowd images exhibit a relatively higher degree of visual detail In image processing terms, this characteristic manifests itself as high texture density However, image texture may be more canonically described as an edge proliferation attribute, since it is the abundance (or paucity) of such that predominantly determines this quality [72] To illustrate this concept, consider **Fig 4 7**, which presents both a mildly textured image sample, and an intensely textured sample For each image, the relationship between their texture densities and the edge pixel densities of their corresponding edge-detected equivalents is clearly evident Given

Fig. 4.6. Images from the three standard field-sports-video camera perspectives (1: close-up, 2: zoom-in, 3: global view), and sample crowd image views (4, 5, 6).

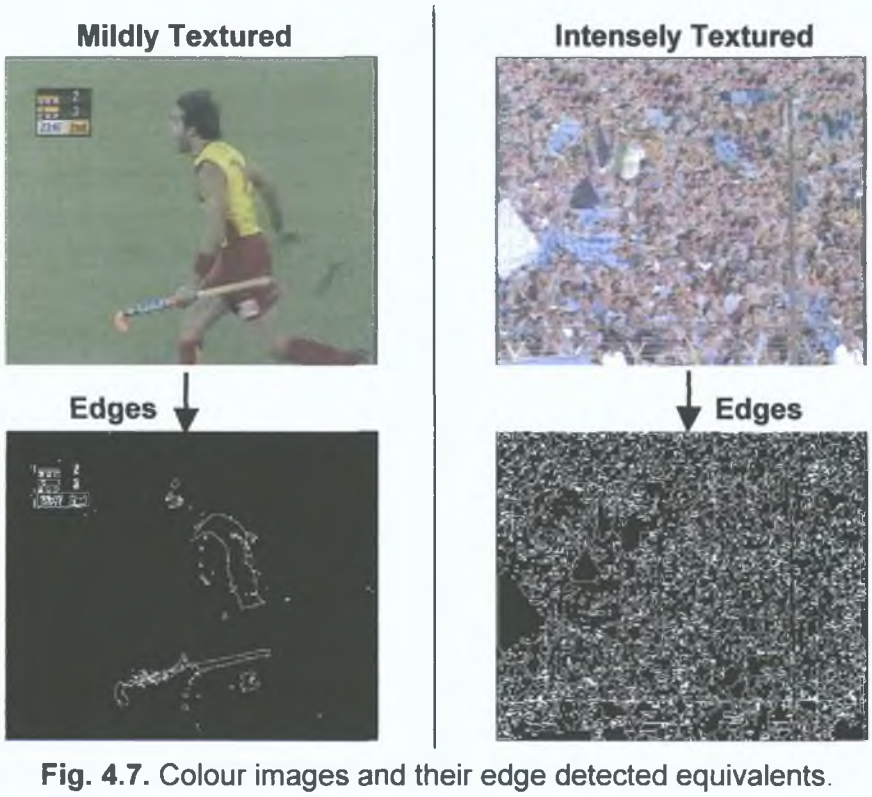


Fig. 4.7. Colour images and their edge detected equivalents.

this, it is proposed that the discrimination of crowd-view images in FSV may be achieved on the basis of edge density quantification

It is also evident from the above crowd view samples that in addition to being conspicuously high, the texture density tends to be uniform throughout the images Hence, as an addendum to the quantification of edge proliferation, it is further proposed that the spatial consistency of the image texture is also exploited, in bolstering the discrimination process It is thus the combined influence of both of these salient crowd image characteristics that forms the basis of their detection

4 4 2 2 Crowd Image Modeling

The proposed crowd image modeling approach is rooted in the generalization of the texture-based characteristics alluded to above Specifically, for a given video frame, it is proposed that an associated ***crowd image confidence*** (**CIC**) measure be generated, according to the degree to which the image exhibits both of the following attributes

(i) an abundance of edges

(ii) spatial consistency in edge intensity

To facilitate the quantification of these attributes for a video frame, it is proposed that it be divided into five ***regions of interest*** (**ROI**), representing both the centre and the four extreme corner regions of the image The approximate positioning of the ROI is described in **Fig 4 8** (it is left to precisely specify the parameters **x** and **y** at the implementation stage) To quantify the abovementioned attributes, it is proposed that ***edge-pixel ratio*** (**EPR**) values be calculated for each region of interest (R_n) using **(4 5)** below, where for a given image, 'EdgePixels' correspond to those that exhibit the value 1 its corresponding binarised edge-detected equivalent

$$EPR_{R_n} = \frac{\#EdgePixels_{R_n}}{\#Pixels_{R_n}} \qquad \textbf{(4 5)}$$

Given the EPR values for each region, a mean value (**μEPR**) is computed via **(4 6)**, which averages their sum (**ΣEPR**) across each of the five ROI It is proposed that **μEPR** quantifies attribute (i) above

$$\mu EPR = \frac{\sum EPR}{5} \qquad \textbf{(4.6)}$$

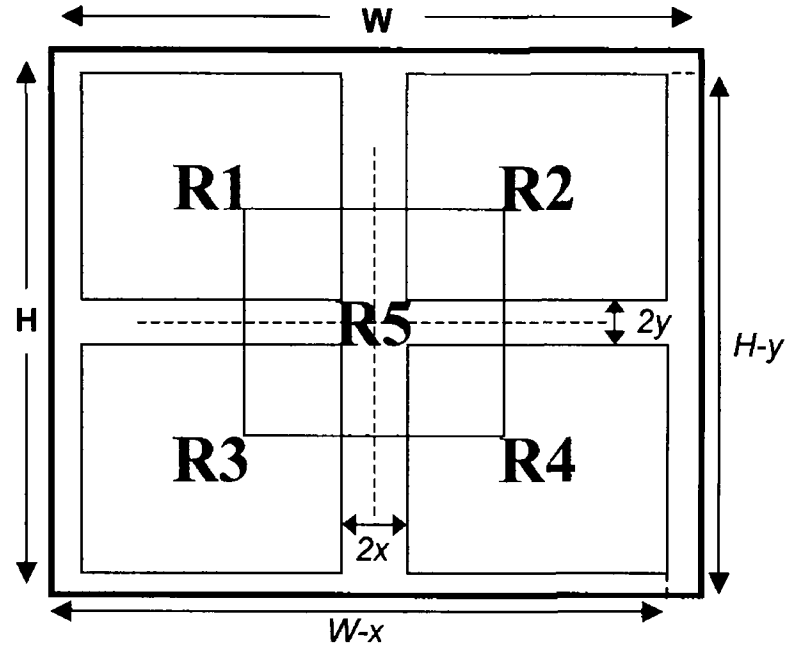


Fig 4 8 Dividing a video frame into five regions of interest (R1-R5)

Furthermore, a value representing the maximum absolute difference between the EPR values for any two ROI may be computed via **(4 7)** It is proposed that this value (**ΔEPR**) should satisfactorily characterize attribute (ii) above

$$\Delta EPR = \left|\max(EPR_{R_n}) - \min(EPR_{R_n})\right|, \quad \forall\, n \tag{4 7}$$

In devising the CIC measure, considering attributes (i) and (ii) above, it was noted that, **μEPR** represents a positive aspect, while **ΔEPR** represents a negative aspect On this basis, a formulation for CIC was proposed and is presented in **(4 8)** Note, in this formulation, to reinforce recall, the term representing spatial inconsistency is weighted by the inverse of the overall ROI edge density, **ΣEPR**

$$CIC = \mu EPR - \frac{\Delta EPR}{\sum EPR} \tag{4 8}$$

4 4.3. CF3: Speech-Band Audio Level Measure

The third critical feature to be developed (CF3) corresponds to the quantification of audio activity However, it is proposed that by making the quantification process

frequency selective, it may be possible to extract ***speech band audio levels*** **(SBALs)** which, based on the following reasoning, is of primary interest in this scenario

4 4 3 1 Speech-Band Focus

FSV audio tracks predominantly exhibit commentator vocalizations, which overlay a background noise ensemble generated from multiple sound sources Thus, by strictly focusing the analysis on the content that resides within the speech-band (approximately 0 5kHz-4kHz [73]), the influence of the commentator vocal source on the energy envelope should be increased This is clearly desirable since it is assumed that the patterns of commentator speech represent the most reliable (i e impartial) noise-level indicators of event significance An additional benefit of limiting the spectral focus of the analysis in this way, is that the processing efficiency should be significantly increased, since it is only a small proportion of the overall audio spectrum that is taken into account

4 4 3 2 Audio Level Extraction

Given an audio signal, the process of quantifying its energy levels may be quite simply performed by adding up the values corresponding to the power spectrum of the audio samples However, given a particular encoded audio representation, it is proposed that there normally exists components of such that lend themselves to exploitation towards providing a more efficient means of extracting the energy levels of an encoded audio signal, than that offered by the process of first decoding it and then analysing at the sample level For example, as outlined in *Section 3 7 2*, a fundamental component of MPEG audio bitstreams is the scalefactor, which are variables that normalize small groups (typically 12) of audio samples, such that they use the full width of the quantiser Recall that the scalefactor for such a group is determined by the next largest value to the maximum of the absolute values of the samples Hence, they provide an indication of the maximum power (volume) of any sample within the group Furthermore, the scalefactors may be individually extracted from any one of 32 equally spaced frequency subbands, which uniformly divide up the input audio bandwidth Hence, the extraction of compressed domain scalefactor data from the bitstreams should prove useful in providing for an efficient frequency-selective means of obtaining knowledge pertaining to the energy envelope of an MPEG encoded audio signal It is proposed that there

exists equivalently exploitable bitstream components for most encoded audio representations

4 4.4. CF4: Scoreboard Suppression Detection

The fourth critical feature (CF4) concerns the process of flagging the suppression of the scoreboard graphic Again, given a video frame, it is required that a confidence value be generated, the value of which indicates the probability that the given image exhibits a suppressed scoreboard To this end, an approach is proposed based on the analysis of pixel-luminance data

4 4 4 1 Scoreboard Graphic Characteristics

The on-screen scoreboard is a synthesized graphical component placed over the images of a video sequence As such, the video footage and over-laid graphics are not broadcast as two separate components, i e the graphic is a constituent of the video signal Furthermore, the format of FSV scoreboards is particular to each broadcaster, and may even occasionally change appearance on an intra-broadcaster basis Hence the prospect of scoreboard analysis based upon the assumption of a known template is unfeasible

However, a salient characteristic of scoreboard graphics is that they exhibit textual data Clearly, for text to be visible, it is required that there exists a strong luminance contrast between the foreground and background This is illustrated in the sample scoreboard graphic presented in **Fig 4 9** Furthermore, for a given FSV broadcast, while the scoreboard graphic may occasionally be suppressed, its location within the frame tends to be static for the entirety of the video On the basis of these

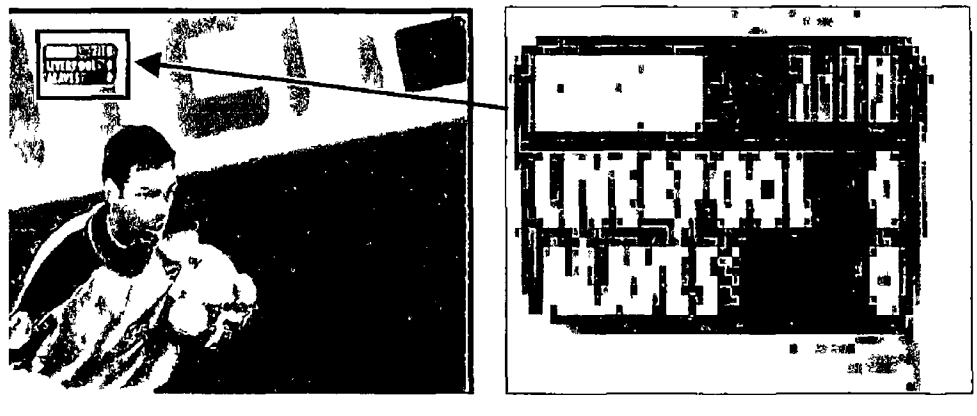


Fig 4 9 Scoreboard graphic of a FSV image showing acute luminance contrast variation in realizing text

two observations, a hypothesis for the automatic positional detection of scoreboards within FSV is proposed With positional knowledge to hand, it is then proposed that a scoreboard suppression detection procedure may be realized by a differencing metric

4 4 4 2 Scoreboard Recognition

The training-corpus was investigated to ascertain the average length of time the scoreboard graphics are displayed on-screen, across various FSV genres **Table 4 9** presents the observations as an average percentage of total duration for each genre From this data it is concluded that for FSV in general, the scoreboard graphics tend to remain on-screen for the majority of the broadcasts It is proposed that this characteristic is exploitable in approaching the task of scoreboard recognition

As mentioned, the scoreboard graphic has a fixed frame position for each particular FSV broadcast Furthermore, it has been explained why scoreboard related pixel blocks must exhibit a high-density variance in luminance intensity, such that textual information may be conveyed Therefore, since the scoreboard graphics tend to be present on-screen for the majority of the time, for a particular video, its scoreboard related pixel blocks should thus exhibit high luminance intensity variances consistently throughout In contrast, non-scoreboard related pixel blocks, will naturally over the course of a broadcast, constitute many different aspects of the images captured Hence, they will generally not exhibit such a consistently high luminance intensity variance

FSV scoreboard graphics must be (i) large enough to convey the textual information, and (ii) small enough such that the occlusion disturbance to the viewer is limited Therefore, if a reliable value representing the average number of pixels used to represent the scoreboard graphics could be determined, then for a particular broadcast, this should provide a reliable means of determining its potential scoreboard pixels (PSPs), by simply finding this number of pixels that exhibit the highest cumulative

Table 4 9 Percentage durations of FSV genres with scoreboards on-screen

Training Corpus Genre	Proportion of content with scoreboard on-screen
Soccer	92%
Hockey	86%
Hurling	98%
Rugby	79%
Gaelic F	97%

luminance intensity variance throughout the course of the broadcast It is then proposed that the recognized scoreboard pixels (RSPs) correspond to the largest spatially connected group of the detected PSPs

The process of quantifying the pixel intensity variance of an image may be quite simply performed by analyzing the individual pixel values However, given a particular encoded video representation, it is proposed that there normally exists components of such that lend themselves to exploitation towards providing a more efficient means of extracting measures of pixel intensity variance of an encoded image, than that offered by the process of first decoding it and then analyzing at the pixel level For example, as outlined in as outlined in *Section 3 6 2*, MPEG-1 video encodes (8x8) image pixel blocks using the Discrete Cosine Transform (DCT) That is, pixel block contents are represented by DCT coefficients in the bitstream, which at the decoder impart knowledge pertaining to the intensity contribution of a set of 64 frequency adapted basis-function components – see *Section 3 4 3 1* The set of basis-functions includes a zero-frequency (DC) component, of which the corresponding ***DC-coefficient*** level indicates the mean overall intensity of the transformed block The remaining 63 basis-functions correspond to non-zero-frequency (AC) components, which are weighted by corresponding ***AC-coefficients*** Given a DCT transformed pixel block, it is the combination of the AC basis-functions that indicates the overall nature of its intensity variance On this basis, it may be extrapolated that the number of non-zero AC-coefficients used to represent the block, is somewhat proportional to its level of intensity variance Given this, it is proposed that the intensity variance level of a pixel block may be characterized reliably without requiring specific knowledge of AC-coefficient values, but simply with knowledge of the amount of AC-coefficients used to represent it It is proposed that there exists equivalently exploitable bitstream components for most encoded video (image) representations

4 4 4 3 Scoreboard Suppression Detection

Assuming the RSPs are reliably detected by some means, in this section a scheme for the suppression detection of scoreboards is proposed, which is based on the luminance domain processing of the detected RSPs

Since FSV scoreboard graphics are on-screen for the major part of the broadcasts, the RSPs convey the scoreboard graphic more often than not Therefore, the mode values of the RSPs, computed across the images of the entire sequence,

should be highly representative of the scoreboards characteristics This is illustrated in **Fig 4 10**, which presents the luminance component of an extracted scoreboard, and the equivalent mode luminance values of the same pixels computed across the images of the corresponding sequence Extraction of the RSP mode luminance values thus provides for the generation of a reliable scoreboard template, which, as further explained below, forms the basis of the proposed scoreboard suppression detection technique

By extracting pixel luminance data (Y) from the images of a FSV sequence, the mode luminance values for the RSPs ($\mathbf{Y^M}$) may be easily computed It is proposed that the spectrum of $\mathbf{Y^M}$ ([0-255]) be quantised into five relatively equal length bins corresponding to very-dark, dark, grey, bright, and very-bright Using the values of the quantised mode $\hat{\mathbf{Y}}^{\mathbf{M}}$ as a scoreboard template, for a given frame (x), a ***mode-variance measure*** (**MVM**) may be computed using **(4 9)**, which effectively quantifies the inconsistency between the similarly quantised RSP luminance values of the given frame ($\hat{Y}^x$), and those of the mode ($\hat{Y}^M$)

$$MVM^x = \frac{\#\,Discrepancies\,(\hat{Y}^x_{RSPs}, \hat{Y}^M_{RSPs})}{\#\,RSPs} \qquad \textbf{(4 9)}$$

Given that a high value for MVM^x suggests a high inconsistency between its RSP luminance values and those of the mode, it is proposed that suppressed scoreboard graphics may be characterized by high values of MVM, and vice-versa

4.4.5 CF5: Visual Activity Measure

The fifth critical feature (CF5) to be developed corresponds to the quantification of motion activity However, for the reasons explained below, sports-video sequences,

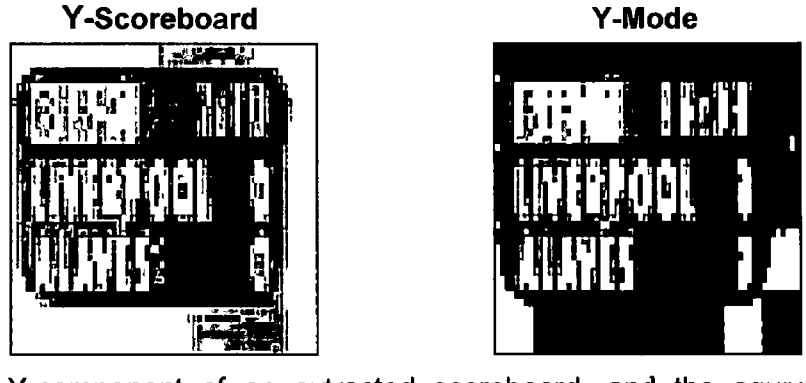


Fig 4 10 Y-component of an extracted scoreboard, and the equivalent mode luminance values computed across all images of the corresponding sequence

especially FSVs, require special attention when it comes to extracting ***visual activity measures*** **(VAMs)** since they are generally characterized by an abundance of motion, corresponding to the movement of the camera.

4.4.5.1. Motion Type Focus

As discussed in *Section 4.1.1*, FSV is characterized by three main camera views, i.e. close-up, zoom-in, and global-view. Following a training-corpus investigation it was observed that whilst zoomed-in and close-up views tend to be employed during relatively stagnant game action and during play-breaks, global views tend to be used to capture the dynamic live-action moments of the games. Global-views are produced by cameras held in a fixed overhead position. Examples of such images taken from training-corpus soccer, rugby, and hockey-video sequences are presented in **Fig. 4.11**.

As outlined in *Section 4.2.2.2*, the post-SUE segments are typically characterized by intense visual activity, particularly during their corresponding reaction-phases. Clearly, it is this type of activity, as opposed to the smooth camera motion of global views, which is of primary interest for detection. Hence, if possible, it is desirable to limit the quantification of VAMs to that concerning motion activity of this class.

4.4.5.2. Visual Activity Extraction

Given a particular encoded video representation, it is proposed that there normally exists components of such that lend themselves to exploitation towards quantifying visual activity. For example, as outlined in *Section 3.6.2.4*, the MPEG video standard employs an inter-frame dependency scheme for the predictive coding of video frames. As explained, in order to increase the compression ratios achievable in frame prediction, a motion estimation (ME) process is employed, in which a luminance domain pixel-block matching technique is used to gauge the motion between the target and a

Fig. 4.11. Video images illustrating global view perspective in FSV.

reference frame In representing this motion, the ME process yields a set of motion vectors (MVs), which indicate the estimated displacement of small image regions (macroblocks) between the frames Following the ME process, the difference between the reference and predicted frame is calculated (residual frame) This is coupled with the MVs, and the ensemble is encoded together At the decoder, the predicted frames are reconstructed via a compensation process, which uses the information contained in the MVs to 'undo' the motion between the frames Therefore, since MVs represent an estimation of the temporal displacement of macroblocks from their original reference frame positioning, they provide a valuable indication of the dynamic activity between the frames Furthermore, since intra-coded (i-) macroblocks represent fresh data, i e data not matched within the ME search space, their presence also represents significant activity Hence, it is proposed that both MV magnitude and macroblock type provide a useful basis, upon which knowledge pertaining to the low-level temporal video attribute of visual activity intensity may be extrapolated It is proposed that there exists equivalently exploitable bitstream components for most encoded video representations

4.4.6. CF6: Field-Line Orientation Detection

The sixth and final critical feature (CF6) relates to the detection of field-end zone action in FSV As alluded to in *Section 4 1 1*, field-lines are standard objects comprising the images of all genres of FSV It is proposed that knowledge relating to the location of the action within the playing field may be inferred from data pertaining to field-line orientation To this end, CF6 specifically corresponds to the detection and extraction of field-lines in FSV images The proposed approach is based upon the analysis of both pixel hue and luminance data, as well as extracted edge and Hough line space data Once the field-lines have been detected for the images of a sequence, it is proposed that the corresponding angles of the most prominent detected lines may be used as input to a higher-level process, concerning the recognition of field end-zone action This inference process will be described at a later stage

4 4 6 1 Field End-Zone Characterisation

As explained in *Section 4 4 5 1*, global views tend to be used to capture the dynamic live-action moments of field-sport games Due to the fixed position of the camera for global-views, the resulting perspective is such that for action situated in the field end-zone, the visible field lines tend to assume certain angles To illustrate this, field end-

zone video images, extracted from training-corpus soccer, rugby and hockey sequences, are presented in **Fig 4.12**. Within these images, the orientations of some of the most prominent visible field lines are highlighted. From these examples it is inferred that, for global view perspectives, the visible field-lines for field-end zone regions tend to lie within a particular narrow interval relative to the point of observation. It is proposed that this suggests a basis for the recognition of field-end zone views.

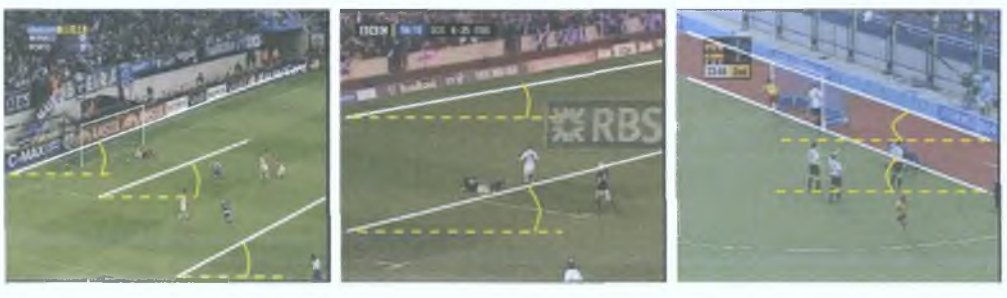


Fig. 4.12. Video images displaying field end-zone action from soccer, rugby, and hockey video sequences.

4.4.6.2. Playing Field Segmentation

The first step in the detection of field-lines in FSV content concerns the segmentation of the playing field from the other objects comprising the images. To this end, an approach is proposed based on the analysis of image hue space data.

Given a FSV sequence, since the most frequent camera perspectives correspond to global-views, it is assumed that the mode pixel hue value occurring, ψ, corresponds to the prevailing hue of the playing field grass. This is a reasonable assumption since the playing field is clearly the largest reoccurring object in FSV content. Hence, given ψ for a sequence, it is proposed that grass pixels may be segmented from non-grass pixels by comparing each individual pixel hue to ψ. Specifically, allowing for small fluctuations, a pixel is deemed a ***field pixel candidate*** (**FPC**) if its hue is within the range $[\psi \pm \eta]$, where η is a tolerance to be specified at the implementation stage. **Fig. 4.13** presents a video image from a training-corpus soccer-video. The value of ψ was determined for the sequence, and based on the abovementioned analysis (taking $\eta = 20°$) the FPCs were detected for this image as shown.

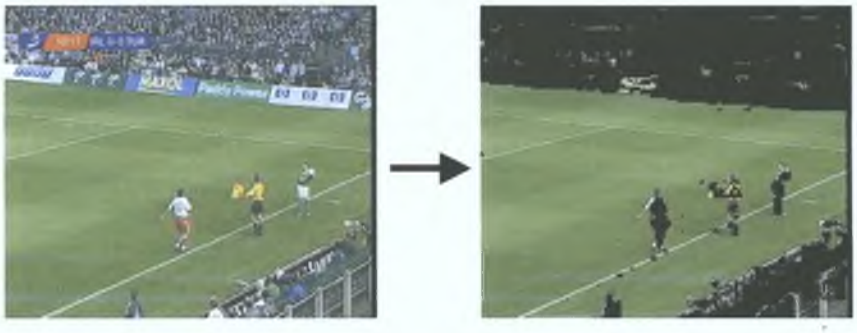


Fig. 4.13. Soccer-video image illustrating the segmentation of FPCs.

From the above example it is evident that the detected FPCs primarily correspond to the majority of the grass related pixels of the image. However, also segmented are various non-grass related pixels, whose hue values happen to lie within the critical interval. Therefore, to avoid the possibility of such elements affecting the subsequent analysis, it is proposed that some type of morphological filtering/erosion will be required to reduce noise in the segmentation map, i.e. towards yielding a set of ***refined field pixel candidates*** **(RFPCs).**

4.4.6.3. RFPC Luminance Binarisation

Assuming reliable extraction of the RFPCs, the next step will involve the segmentation of the field-lines from the set of RFPCs. Given the RFPCs of an image, it is proposed that the field-line pixels may be segmented from the grass pixels in the luminance domain. That is, since the field-line pixels are brighter than those of the grass their segmentation should be feasible via a binarisation of the luminance space of the image. However, a fixed threshold may be unreliable for varying image brightness/contrast, which is typically a consequence of varying weather conditions. Therefore, a methodology for dynamically assigning a threshold is proposed on the following basis.

Since grass pixels constitute the majority of the RFPCs, it is assumed the mode luminance of this set corresponds to the prevailing luminance of the grass, i.e. not the field-lines. Using this mode luminance value as a threshold, the RFPC luminance values are binarised into bright and dark pixels. On this basis, the bright field-lines should be discernable from the darker grass. **Fig. 4.14** illustrates this process applied to an image where the RFPCs were manually segmented (for illustrative purposes non-RFPCs are coloured white). In this example it is evident that via the dynamic thresholding, the

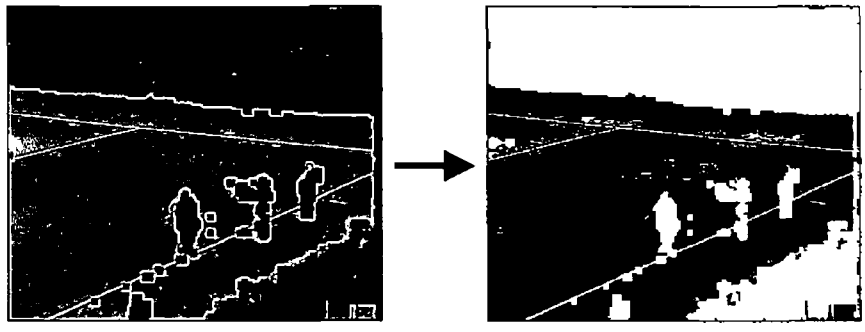


Fig 4 14 Binarisation of RFPC luminance data using dynamic threshold

majority of the luminances of both the RFPC field-line pixels and grass pixels have been binarised appropriately

4 4 6 4 Edge Detection

It is then proposed that the binarised luminance values of the RFPCs be edge detected towards outlining the edges of the field-lines There exist many well-established solutions to the problem of edge detection in digital images Examples include the Sobel [74] and Canny [75] algorithms However, the approach utilized herein is the Roberts Cross [76] method since (i) it exhibits a more simplistic methodology (i e comparatively more computationally efficient than the aforementioned methods), (ii) field-lines exhibit a sharp change in intensity and the Roberts method has been shown to be reliable for the detection of sharp edges in digital images [76], and finally (iii) instead of responding maximally to vertical/horizontal edges like the Sobel algorithm (and by proxy the Canny solution) the Roberts method responds maximally to lines running at 45° to the pixel grid [76], a characteristic that correlates well with the end-zone lines that are required to be detected

4 4 6 5 Hough Line Transform

In terms of detecting the most prominent line from the edge map of the binarised luminance RFPCs, it is proposed that an appropriate scheme is the ***Hough Line Transform*** (**HLT**), which corresponds to a particular instance of the ***Hough Transform*** (**HT**) [77] The HT is a generic image processing methodology for the recognition of specific types of visual features within digital images, such as lines, circles, etc The algorithm was developed by Paul Hough in 1962 and subsequently patented by

IBM Corp [78] In its operation, a description of the feature concerned (e g line, circle) is appropriately parameterized, with respect to its characterisation in Cartesian image space In doing so, this process spawns a Hough space lattice, defined by all potential values of the description parameters For a given image, a tally is maintained at each lattice point - the value of which suggests how well the feature described by the parameters, defined at that particular lattice point, matches the data in the original image On this basis, knowledge of the occurrences and characteristics of the feature concerned may be extracted from amongst larger amounts of other data within the image The most basic mode of the HT is the HLT In this scenario the normal-form line representation is exploited in generating the Hough space lattice, which is defined by the corresponding polar parameters that characterise this line description From this, knowledge concerning the position and orientation of lines within the original image space may be inferred

It is proposed that given the edge detected binarised luminance RFPCs, the data be transferred to Hough line space as described, within which it is anticipated the most prominent lines (and their associated orientations) may be discerned as those corresponding to the highest lattice intersection tallies

4.5. Shot-Boundary Detection

In order to process CF data at the shot level, it is required that knowledge of the frame-level boundaries of such be determined This section introduces the methodology that is proposed to realise this in terms of the nature of the content to be dealt with

As introduced in *Section 3 3 4*, the camera shot, which corresponds to the video resulting from a continuous, unbroken recording by a single video camera [54], is the basic syntactical unit of a video sequence Shots may be delimited by a variety of boundary transition types, e g hard cuts, fades, dissolves, and wipes However, it was recognised that due to the generally high tempo nature of FSV games, during the live action segments the broadcast director has little chance to utilize shot transition types other than abrupt shot cuts In fact, it was manually quantified that at least 95% of all shot transitions within the multi-genre FSV training-corpus were of this nature In contrast, it was found that video effects transitions such as dissolves, wipes etc tend only to occur when the director has time to be more creative, i e during a break in the

play or during a break in the live action (e g during action replays) Given this, the shot-boundary detection analysis in this thesis is primarily concerned with shot-cut detection

To this end, it is proposed that an externally developed shot-cut detection tool [79] be employed for this task A comprehensive description of this tool is presented in *Appendix A*, along with a general introduction to the topic Also in *Appendix A*, via an appraisal of the performance of the tool on the training-corpus, it is shown that it provides for a very reliable means of detecting hard shot cuts in field-sports-video content

4.6. Pre-Processor Filter

It is proposed that the summarization performance that is expected to be yielded by the CF pattern analysis based detection of SUE-shots may be improved upon, or at least bolstered, by the incorporation of a pre-processing content filter By and large, the main task of the pre-processor would be to reject outright any periods of FSV content that are clearly irrelevant to the SUE detection task, i e periods that, without resorting to a detailed pattern analysis of the CF combinations, may be robustly classed as most likely not exhibiting a SUE Given this knowledge, the scope (i e the '***probing domain***') of the subsequent CF pattern analysis may be then restricted accordingly The positive consequence of this is that the quantity of content considered for further CF pattern analysis phase would be reduced, such that increased efficiency and hence performance speed might be attained Moreover, improved performance may be yielded for the retrieval task, since any potential false-positives contained within these segments would be eradicated beforehand

4.6 1. Advertisement Detection

It is proposed that the first stage of the preprocessing filter should concern the removal of advertisement breaks A scheme providing for the removal of advertisement breaks from broadcast television programmes was developed by this author in another work [80] The solution has been shown to operate successfully across a wide-varying corpus of generic video, including news, sports, chat show, game show, and cartoon [80] Specifically, the methodology is rooted in a pattern recognition concept, which models the frequency of detected audiovisual signal depressions, which tend to delimit the individual 'ad' segments that comprise completed advertisement breaks The scheme is

inherently biased towards precision, in that in testing, the results state average precision of 100% and a corresponding average recall value of 94 8% [80] It was decided to incorporate this ad-break detection scheme into the FSV pre-processing stage in this work, such that any advertisement breaks within the content that are detected and flagged, are subsequently de-listed from the probing domain for the CF pattern analysis stage

4.6.2. Close-Up Based Content Filter

As outlined in *Section 4 2 2 1*, it was estimated that on average, almost 98% of all training corpus SUE-shots exhibited a close-up image sequence during their reaction-phase segments In view of this high correlation, it was decided to exploit this critical feature (i e CF1) at the preprocessor stage, in defining a retention condition for potential SUE-shots Specifically, the proposed stipulation requires that for a given shot to be retained for further CF pattern analysis, it must be followed by an instance of a close-up sequence within its post shot-end boundary (SEB) reaction-phase seek-window (RPSW) – as defined in *Section 4 2 2* Clearly, while it is not uncommon for many non-SUE-shots to be followed by close-up views, on the basis of the high correlation percentage observed, it was envisaged that this condition should provide for a favorable trade-off in the retention of potential SUE-shots, and the rejection of others within FSV content It was proposed that in terms of implementing the reaction-phase close-up detection task, a CuC threshold ($\mathbf{T_{CuC}}$) be defined Then, for a given shot **i**, the maximum CuC exhibited by any of its respective RPSW images ($\mathbf{[CuC_{MAX}]_{RPSW_i}}$), be compared to $\mathbf{T_{CuC}}$ towards determining whether the shot should be retained or rejected - see **(4 10)**

$$If \quad [CuC_{MAX}]_{RPSW_i} \geq T_{CuC} \quad \Rightarrow \quad Shot-i \ is \ retained \qquad \textbf{(4 10)}$$

4.7. Shot-Level Critical Feature Aggregation

It was described in *Section 4 6 2*, how it is proposed to exploit CF1 evidence towards content rejection at the pre-processing stage It is thus proposed that the actual SUE-shot detection process relies on the indicative combinations of the remaining five CFs (i e CFs2-6) in a more sophisticated pattern analysis phase of the scheme Recall from *Section 4 3*, that this corresponds to the process of quantifying the shot-level prevalence/intensity of these features within appropriate temporal windows (i e CFs2-5

within the reaction-phase seek window (RPSW), and CF6 within a (so far unspecified) shot-end seek window In terms of realising this, it is proposed that, for a given video, frame-level evidence for CFs2-6 be extracted at some appropriate level Then, assuming reliable shot-boundary detection, this evidence be appropriately processed as described towards generating ***shot feature vectors*** (**SFVs**), in which the individual ***vector component coefficients*** (**VCCs**) represent a critical quantification of the prevalence/intensity of CFs2-6 (e g maximum confidence) within the key intervals – see **(4 11)**

$$SFV = [VCC_1, VCC_2, VCC_3, VCC_4, VCC_5] \qquad \textbf{(4 11)}$$

Given the training-corpus observations presented in *Section 4 2*, it is envisaged that the SUE-shots should exhibit certain indicative SFV patterns, and therefore on the basis of some higher-level SFV pattern analysis method, they should be discernable from other shots

4.8. Chapter Summary

In this chapter a hypothesis for event detection-based summarization in the field-sports-video supergenre was proposed and justified Initially, the features deemed both necessary and sufficient in characterizing field-sport-video were described Next, given the target of detecting the score-update episodes (which were recognized as constituting the major narrative-critical events of field-sport-video), the features that were deemed to generically characterize all SUE manifestations were inferred via a training-corpus investigation Specifically, these related to close-up views, crowd views, suppressed scoreboards, increased visual activity, increased audio activity, and field end-zone action A hypothesis for the detection of SUE-shots was then proposed on the basis of the quantifying the intensity/prevalence of these critical features within appropriate seek windows To this end, methodologies were proposed for the frame-level extraction of these six critical features from field-sports-video content Next, on the basis of the extremely high correlation observed between close-up views and SUEs, it was proposed that extracted confidence values pertaining to this critical feature in particular be exploited in constituting a filter component of a proposed preprocessor stage It was then described exactly what format the proposed shot-level aggregation process will take for the extracted frame-level evidence of the remaining CFs

Chapter 5

Hypothesis Implementation

The hypothesis for the automatic summarization of field-sports-video (FSV) was outlined in *Chapter 4*, and as described, it is rooted in the detection of score-update episode shots (SUE-shots) based on the quantification of the prevalence/intensity of six frame-level critical features (CFs) within specific temporal seek windows Specifically, the proposal is that evidence corresponding to one CF be exploited in constituting a shot filter component of a pre-processing stage (in conjunction with an ad-break detection algorithm), the aim of which is to bolster both precision accuracy and the overall computation efficiency of the scheme Then, it is proposed that evidence relating to the remaining CFs be aggregated at the shot-level, towards providing a critical shot-level description of the (pre-processed) content, upon which it is anticipated that SUE-shots may be discerned In this chapter, it is described how each element of this overall proposed hypothesis is implemented with respect to the field-sports-video data corpus obtained and the nature of the content representation used, i e MPEG-1 (see *Section 1 5 2 1*) Although the representation used is specific, it is maintained that no feature is exploited in particular that is not characteristic of many other representations, e g MPEG-2/4 and H 26x, and hence it is anticipated that the implementation remains transferable on this basis

5.1. Implementation of CF Extractors

Proposed methodologies for the extraction of the frame-level critical features (CFs) were presented in *Section 4 4* In this section it is fully described how these proposals are implemented in terms of their extraction from the MPEG-1 encoded FSV data corpus

In each case, an illustration of the effectiveness of the CF extraction process is provided It is important to note that descriptions regarding the development of tools for the extraction of the relevant audiovisual signal-level data, upon which the CFs are derived, are provided in *Appendix B* The reader should be familiar with the concepts of these signal-level features and their extraction process In short, they relate to methods for the extraction of DCT coefficients, motion vectors, and audio subband scalefactors (directly from the MPEG encoded bitstream), and from the uncompressed domain, pixel luminance/chrominance data, edge pixel data, and Hough line space data

5 1.1. CF1 Close-Up Confidence (CuC) Measure

In *Section 4 4 1*, a colour-based approach to generating close-up confidence (CuC) measures was proposed, which was based on segmenting the images into regions of expectancy (ROE) for face and jersey entities, and quantifying the degree to which both have a strong presence within these regions

5 1 1 1 Implementation & Parameter Settings

Based upon evidence from numerous close-up images (carefully chosen in proportion from all five FSV genres constituting the training-corpus), the best-fit ROE for these characteristics were estimated The dimensions and positioning of these inferred regions are illustrated in **Fig 5 1**, where **W** and **H** represent the frame width and height, respectively Specifically, the best-fit frame position for R1 was delineated empirically as

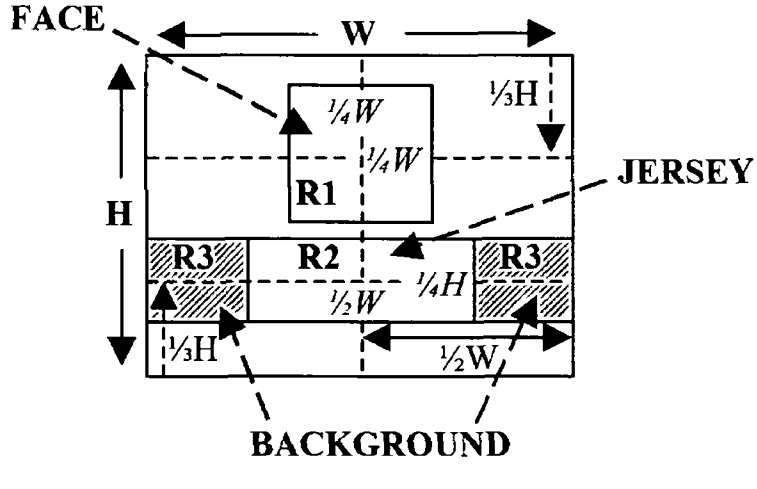


Fig 5 1 Estimations for the best-fit regions of expectancy for face, jersey, and occluded background for generic close-up image

a square of dimension ¼**W** centred on the vertical median, at a horizontal position corresponding to ⅓**H** from the top of the frame. The best-fit frame position for R2 was a rectangle of dimensions ¼**H** x ½**W** centred on the vertical median, at a horizontal position corresponding to ⅓**H** from the bottom of the frame. R3 was simply defined as the outstanding regions that are generated by a bi-directional extension of the dimensions of R2 to the image border. In determining the dominant hue pixel ratios as described in *Section 4.4.1.2*, it was empirically determined that optimum results were obtained by setting the tolerance value $\xi = 10°$.

To specifically implement the procedure of generating CuC values as described in *Section 4.4.1.2*, a software tool called ***CloseUpConfExtract*** was designed and built in the C programming language. Given a FSV sequence to be analysed, ***CloseUpConfExtract*** takes low-level pixel hue data for each frame as input (see *Section B.5* of *Appendix B* for information on how the pixel hue data was extracted) and then executes the procedures as outlined, thus yielding resultant CuC values for each input frame. To verify its effectiveness in this task, an evaluation is provided in the following section.

5.1.1.2. Effectiveness

To evaluate the effectiveness of ***CloseUpConfExtract***, consider again the sample close-up images-A and -B presented in **Fig. 5.2**. The critical ROE for close-up images as defined above were applied to these images - see **Fig. 5.3**. For each respective region the critical pixel ratio analyses were performed as described above and are presented in **Fig. 5.4**. Using the respective values for images-A and -B in **(4.4)** yields resultant close-up confidence values of $\mathbf{CuC_A = 0.3474}$, for image-A, and $\mathbf{CuC_B = 0.3078}$, for image-B.

Fig. 5.2. Two close-up image samples.

Fig. 5.3. Close-up image regions of expectancy applied to sample images A and B.

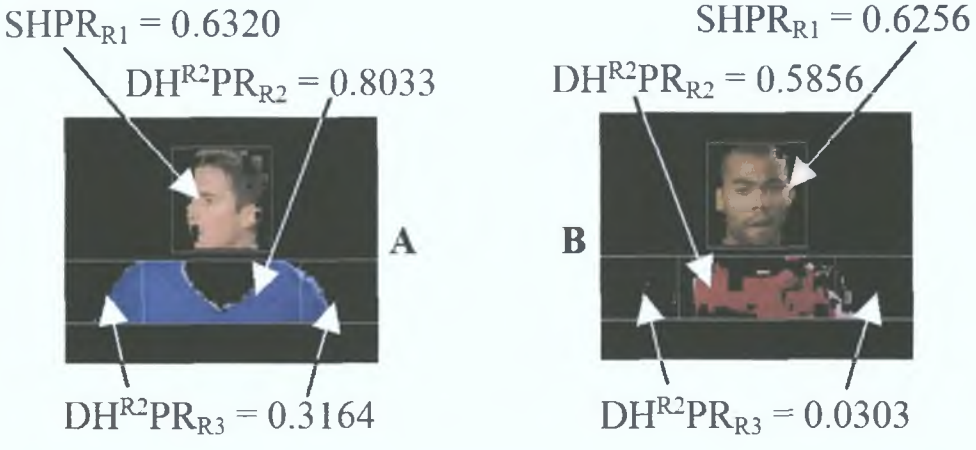


Fig. 5.4. Pixel ratios for close-up model ROE applied to images A and B.

However, to gauge the significance of the magnitude of these values in discerning close-up views, it is necessary to demonstrate the execution of the scheme across both close-up and non-close-up images alike. To this end, four distinctly non-close-up images (representing different levels of camera view), as well as four more close-up images, were extracted for comparison from the various genres that comprise the FSV training corpus. **Fig. 5.5** presents the close-up images, M, N, O, P. Also presented in this figure are the critical pixel ratios for each of the ROE.

Given these, the corresponding CuC values are tabulated in **Table 5.1**. It is evident that these values are of a similar magnitude to those of the earlier close-up images-A and –B, which for the purposes of comparison are also tabulated. The four arbitrarily chosen non-close-up images, W, X, Y, & Z, are presented in **Fig. 5.6**. Upon

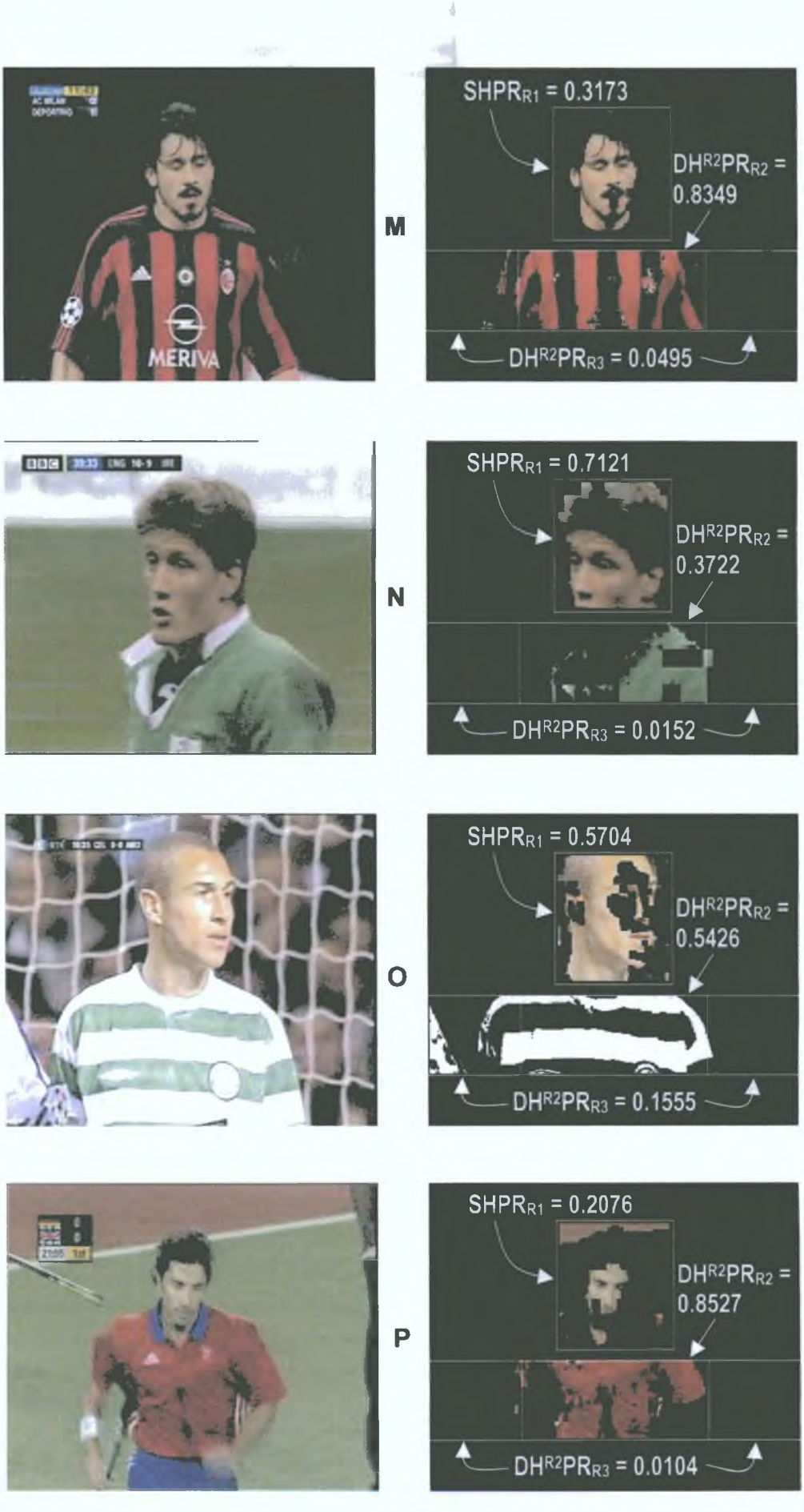


Fig. 5.5. Pixel ratio analysis of four close-up images.

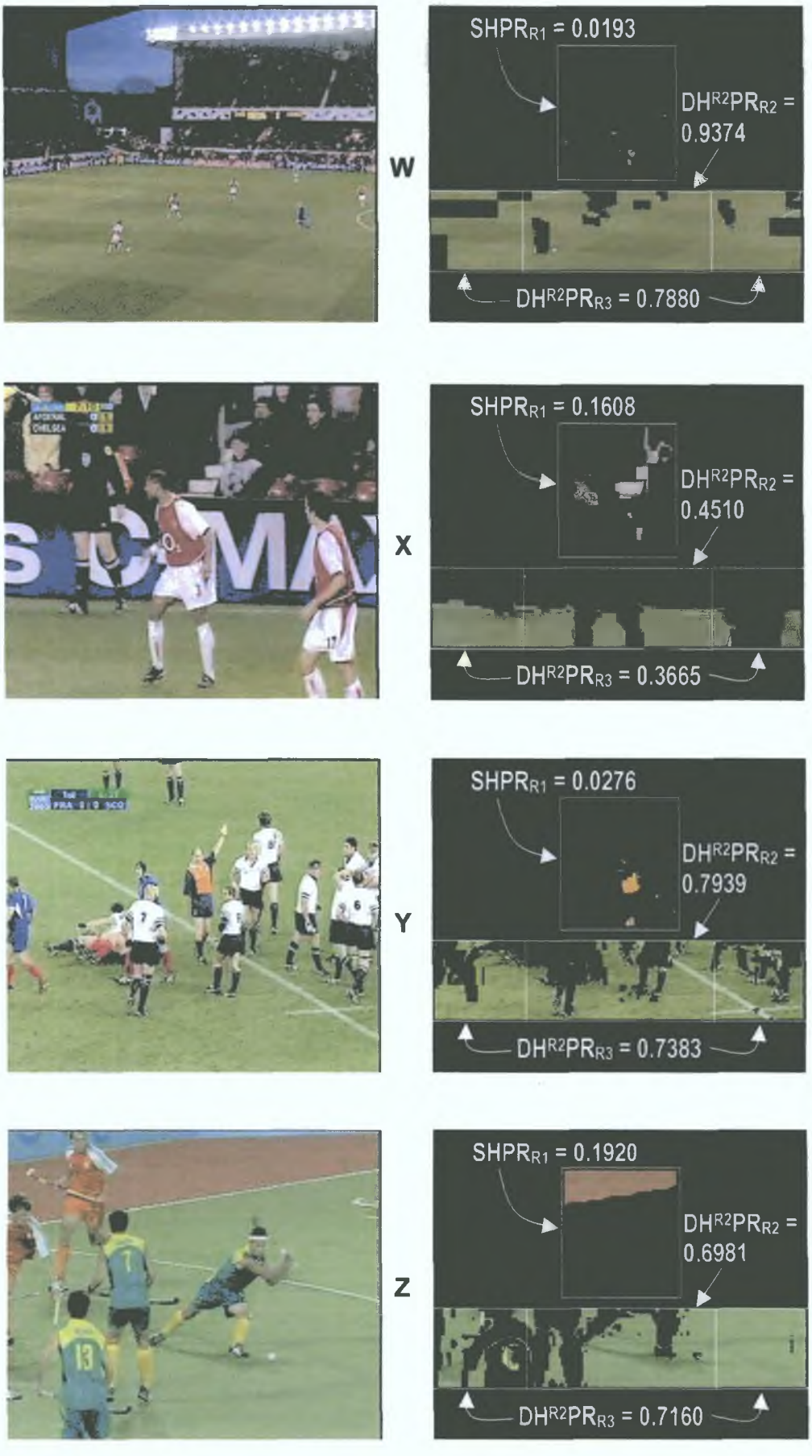


Fig. 5.6. Pixel ratio analysis of arbitrarily chosen non-close-up images.

analysis of their respective CuC values (also presented in Table 5 1), it is evident that there is at least a factor of 10 in difference between the values for close-up and the non-close-up images Based on this, it is concluded that the model is effective in the discrimination between such views in terms of the examples shown However, it must be recognized that the nature of the close-up images upon which this model is based on (exemplified by those used in the above illustration) are those of a very well defined type, i e an ideally centred player with little or no occlusion Therefore, such sharp discrimination involving close-up views of a non-ideal nature cannot be expected, and it is accepted that this observation must be taken into account when exploiting such evidence

Table 5 1 Close-up confidence values for assessed images

Image	Type	CuC
M	Close-Up	0 2493
N	Close-Up	0 2543
O	Close-Up	0 2208
P	Close-Up	0 1748
A	Close-Up	0 3474
B	Close-Up	0 3078
W	Non Close-Up	0 0029
X	Non Close-Up	0 0136
Y	Non Close-Up	0 0015
Z	Non Close-Up	0 0000

5.1 2. CF2: Crowd Image Confidence (CIC)

In *Section 4 4 2* a texture-based (edge-based) approach to generating crowd image confidence (CIC) measures was proposed, which involved segmenting the images into five regions of interest (ROI), and quantifying the degree to which the images have a high texture density that is spatially uniform

5 1 2 1 Implementation & Parameter Settings

In terms of the specific positioning of the ROI segmentation illustrated in Fig 4 8, the parameters x and y were chosen as follows, **x = 0 025W** and **y = 0 025H** (this provides for the deliberate exclusion of pixels residing close to the image edges, which is desirable since these occasionally contain high-frequency noise that tends to interfere with the

texture analysis). To specifically implement the extraction of CIC measures as described in *Section 4.4.2.2*, a software tool called ***CrowdConfExtract*** was designed and built in the C programming language. Given an MPEG-1 video sequence to be analyzed, ***CrowdConfExtract*** takes pixel edge data for each frame as input (see *Section B.6* of *Appendix B* for information on how the pixel edge data was extracted) and then yields resultant CIC values for each image. The effectiveness of this tool is evaluated in the following section.

5.1.2.2. Effectiveness

To illustrate the effectiveness of ***CrowdConfExtract*** in the discrimination of FSV crowd image views, consider the crowd images-P and -Q presented in **Fig. 5.7**, which display differing levels of camera zoom. **Fig. 5.8** illustrates the demarcations of the ROI applied to the edge-detected equivalent of image-P. The EPRs were determined for each

Fig. 5.7. Two crowd images, P & Q.

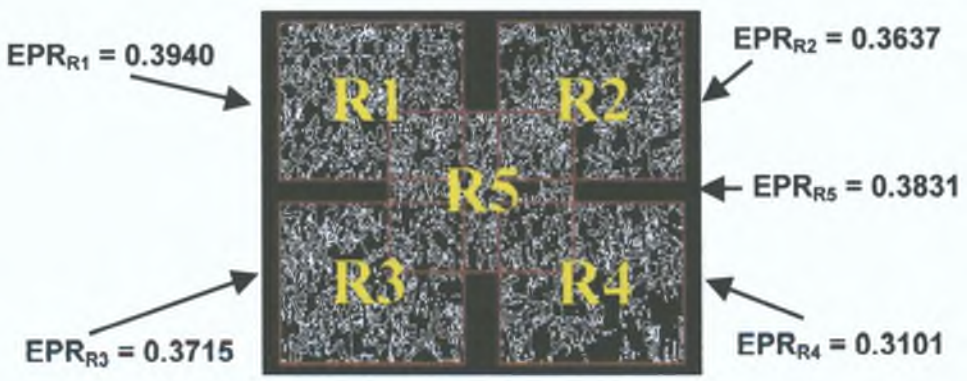


Fig. 5.8. Edge-pixel analysis of image-P.

ROI, and these are also displayed in the figure. From these values it was determined that $\Sigma EPR_P = 1.8225$, $\mu EPR_P = 0.3645$, and that the maximum EPR difference between any two zones is that between R1 and R4, i.e. $\Delta EPR_P = 0.0839$. Using these values in **(4.8)** gives a resultant crowd image confidence value of $CIC_P = 0.3184$.

Similarly, **Fig. 5.9** illustrates the ROI division of the edge-detected equivalent of image-Q. Again the EPRs were determined for each ROI, and these are also displayed in the figure. From these values it was determined that $\Sigma EPR_Q = 1.2380$, $\mu EPR_Q = 0.2476$, and that the maximum EPR difference between any two zones is that between R3 and R5, is $\Delta EPR_Q = 0.0238$. Using these values in **(4.8)** gives a resultant crowd image confidence value of $CIC_Q = 0.2284$.

Fig. 5.10 presents two more crowd images (R, S), and four distinctly non-crowd views (H, I, J, K). Also illustrated in this figure are the ROI-divided edge-detected equivalents of the images. The resultant CIC values are tabulated in **Table 5.2**,

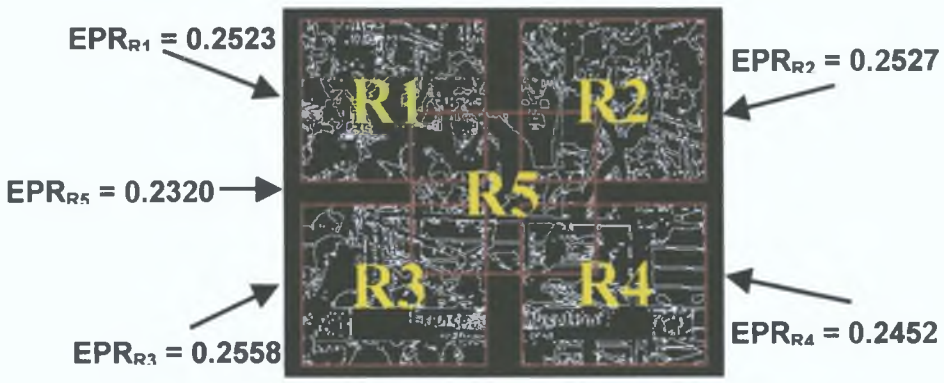


Fig. 5.9. Edge-pixel analysis of image-Q.

Table 5.2. Crowd image confidence values for assessed images.

Image	Type	CuC
P	Crowd	0.3184
Q	Crowd	0.2284
R	Crowd	0.1578
S	Crowd	0.2467
H	Non-Crowd	-0.2975
I	Non-Crowd	-0.1239
J	Non-Crowd	0.0144
K	Non-Crowd	-0.1203

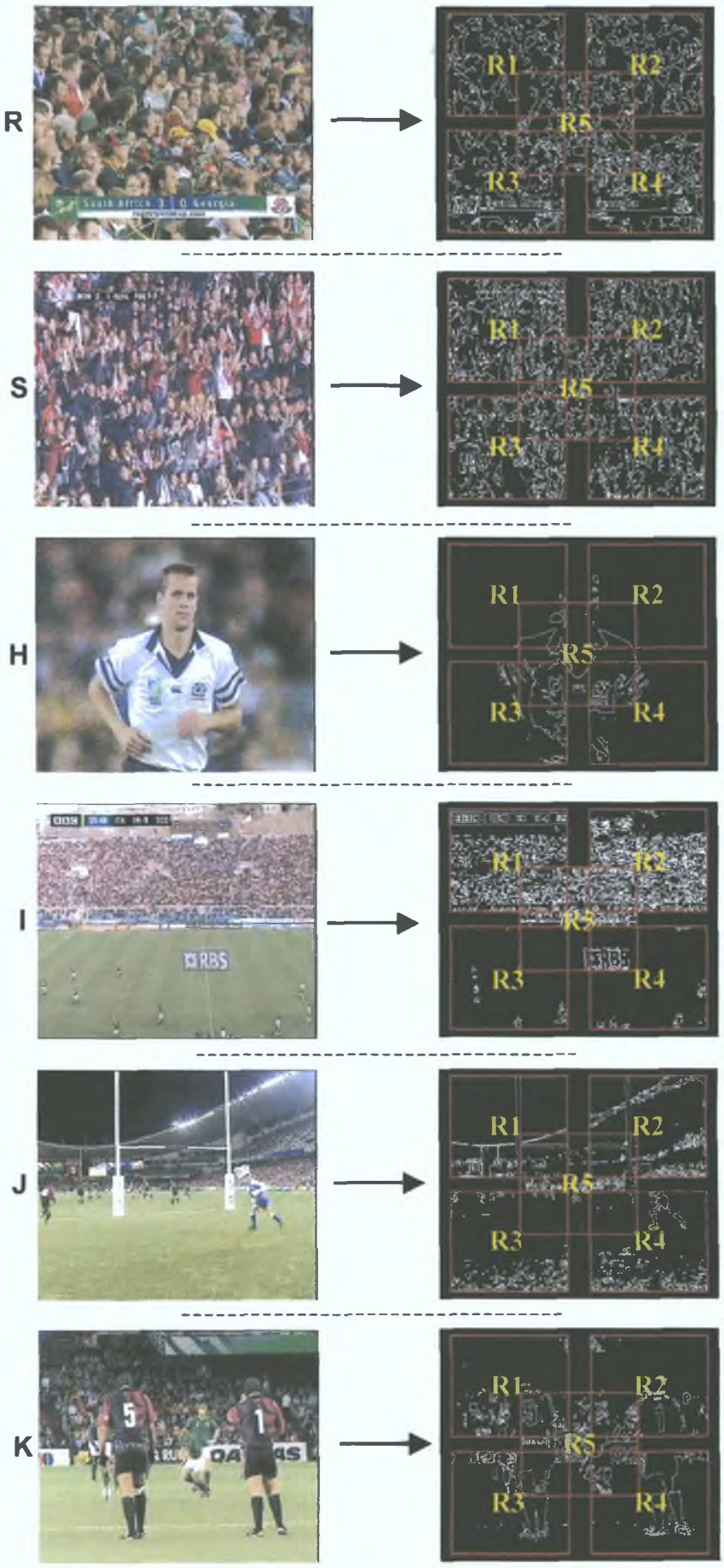


Fig. 5.10. Edge-pixel analysis applied to crowd images (R, S) and non-crowd images (H, I, J, K).

together with those of images-P and –Q for comparison It is evident that a difference of at least a factor of 10 exists between those corresponding to crowd image samples and those of a non-crowd nature On this basis, it is concluded that within the limited domain context of FSV, the proposed approach provides for an excellent discrimination between the two classes

5.1.3. CF3 Speech-Band Audio Level (SBAL)

In *Section 4 4 3 2*, it was explained how, given a particular encoded audio representation, there normally exists bitstream components of such that lend themselves to exploitation towards providing an efficient frequency-selective means of extracting the energy levels of an encoded audio signal As mentioned previously, the representation used in this work is MPEG-1, and therefore in terms of extracting speech-band audio levels (SBALs) as required, the procedure implemented involves the manipulation of subband scalefactor data, as described in *Section 4 4 3 2*

5 1 3 1 Implementation & Parameter Settings

As noted in *Section 3 7 2*, in the MPEG encoding of audio sequences, the input frequency spectrum is divided into 32 equally spaced subbands Since, the input spectrum is band-limited to [0-20kHz], it is thus concluded that subbands 2 through 7 represent the frequency range from [0 625kHz – 4 375Khz] The span of these six subbands approximates the spectrum of human speech [73] An additional benefit of limiting the spectral focus to these selected subbands, is that the processing efficiency of the analysis should be significantly increased, since it is only scalefactors from 6 of a possible 32 subbands that are taken into account

Following the extraction of the scalefactors from subbands 2-7, a value for the ratio of extracted scalefactors to the number of video frames (Ψ), may be determined using **(5 1)**

$$\Psi = \frac{\#ScalefactorsExtractedFromAudioTrack}{\#FramesInVideoSequence} \qquad \textbf{(5 1)}$$

Given Ψ, ***speech-band audio levels*** (**SBALs**) may be then determined for any video frame of the sequence as the root-mean-square (RMS) equivalents of their corresponding Ψ scalefactors (S_i), as shown in **(5 2)**

$$SBAL = \sqrt{\frac{\sum_{i=1}^{\Psi} S_i^2}{\Psi}} \qquad \textbf{(5 2)}$$

However, given that FSV audio tracks are comprised of sounds from multiple sources, there is the potential for their corresponding energy envelopes to exhibit irregular noise spikes To combat this, the video frame-level SBAL values are subjected to a smoothing procedure As described, the feature is primarily concerned with reflecting the energy dynamic of commentator vocalisations within FSV audio tracks, however, due to the limits in the capacity of human responsiveness, the variation rate of vocal dynamics exhibits an upper bound In fact in [81], it is argued that the average human responds to a stimulus within 0 75s - 1 0s On this basis, it was assumed reasonable to suggest a 0 5s sliding window for the smoothing of SBAL values Such an interval should provide for a reasonable trade-off in being short enough to capture the dynamics of human responsiveness, and long enough to facilitate the suppression of fleeting noise spikes Given a 1-D data set, **(5 3)** defines the arithmetic for a mean-filtering (smoothing) operation, where $\mathbf{x}_i$ is the data entry currently being filtered, and **N** is the number of elements within the prescribed interval

$$\textit{If } N \textit{ even} \quad \overline{x_i} = \frac{\sum_{j=i-\frac{N}{2}}^{i+\frac{N}{2}} x_j}{N+1}$$

$$\textit{If } N \textit{ odd} \quad \overline{x_i} = \frac{\sum_{j=i-\frac{N-1}{2}}^{i+\frac{N-1}{2}} x_j}{N} \qquad \textbf{(5 3)}$$

Given the framerate of the data corpus (i e 25fps), a 0 5s interval corresponds to 12 5 MPEG-1 video frames Hence, in accordance, it is proposed that extracted frame-level SBAL values are mean-filtered via the formula in **(5 3)**, with N set to 13 – see **(5 4)**

$$\overline{SBAL_i} = \frac{\sum_{j=i-6}^{i+6} SBAL_j}{13} \qquad \textbf{(5 4)}$$

To implement the abovementioned procedures, a software tool called ***SpeechBandEnergyExtract*** was designed and built in the C programming language Given the scalefactor input (see *Section B 8* of *Appendix B* for information on how the scalefactor data was extracted), this tool yields mean-filtered SBAL values for the frames of a video sequence Since it operates purely on compressed bitstream data, and only a partial segment of the audio spectrum is considered, this novel approach to audio envelope energy tracking exhibits excellent computational efficiency compared to the more conventional sample-based approaches

5 1 3 2 Effectiveness

The correlated relationship between the envelope of an audio signal waveform and its corresponding scalefactor data is illustrated in *Section B 8 2* of *Appendix B* The abovementioned procedures involved in the execution of ***SpeechBandEnergyExtract*** merely concern the manipulation of such data into a cogent mean-filtered frame-level feature Therefore, it is assumed that the effectiveness of this tool in the objectives outlined may be implied from this illustration

5.1.4 CF4: Scoreboard Suppression Confidence (MVM)

In *Section 4 4 4* a luminance-based approach to generating scoreboard suppression confidence measures was proposed Specifically, it was first proposed that the potential scoreboard pixel blocks (PSBs) be determined as those exhibiting the highest cumulative luminance variance intensities throughout the broadcasts The recognised pixel blocks (RSBs) are then the largest spatially connected group of PSBs It was then proposed that the mode luminance values of the RSB pixels be calculated thus providing a scoreboard template, and on this basis, mode variance measures (MVMs), which represent the average discrepancies between the luminance values of the RSB pixels of a given image and the mode values, be calculated towards indicating whether or not the scoreboard is present/absent

In *Section 4 4 4 2*, it was explained how, given a particular encoded video representation, there normally exists bitstream components of such that lend themselves to exploitation towards providing an efficient means of indicating the level of pixel variance intensity As mentioned previously, the representation used in this work is MPEG-1, and therefore in terms of quantifying luminance variance intensity as required,

the procedure implemented involves the analysis of DCT coefficient data, as described in *Section 4.4.4.2.*

5.1.4.1. Implementation & Parameter Settings

As explained in *Section 4.4.4.2*, it is anticipated that due to their high level of luminance intensity variance, the luminance components of DCT encoded scoreboard pixel blocks should necessitate a high number of AC coefficients in their DCT representation. In contrast, given that non-scoreboard related pixel blocks, over the course of a broadcast, constitute many different aspects of the images captured, they will generally not exhibit such a consistently high profusion of AC-DCT coefficients. Hence, this trait forms the basis of this particular implementation of the scoreboard recognition process. That is, the quantification of luminance intensity variance is performed at the pixel block level, rather than the pixel level as initially introduced in *Section 4.4.4.2.*

The luminance domain DCT coefficients were extracted from several diverse scoreboards, which were manually selected from all FSV genres constituting the training-corpus. In all, fourteen different broadcaster scoreboard formats were observed from this corpus. For each case, the number of AC-DCT coefficients used to represent each of its constituent pixel blocks was recorded. **Fig. 5.11** illustrates the distribution of these counts across all blocks analysed. From this distribution it is evident that, for the given bitrate, over the 14 formats analysed, a negligible percentage of the scoreboard pixel blocks exhibited an AC-DCT coefficient count of less than 10 (dashed line). This

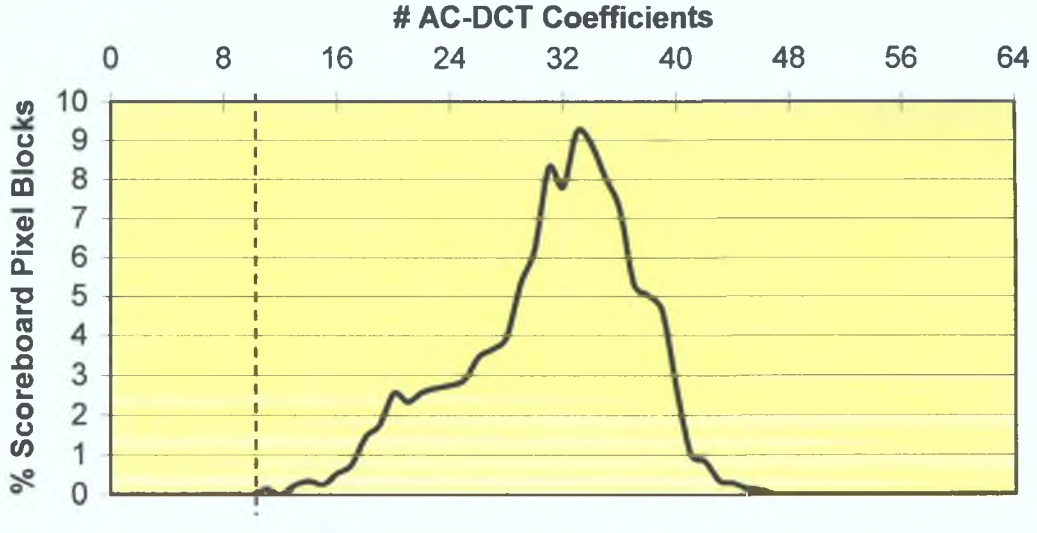


Fig. 5.11. Distribution of AC-DCT coefficients for scoreboard related pixel blocks, corresponding to 14 different scoreboard formats extracted from the training corpus.

limit is used as a discriminatory threshold in the development of a scheme for the detection of scoreboard related pixel blocks Specifically, it is proposed that the Y-DCT coefficients are extracted for each I-frame of the sequence Then, for each pixel block address (b), a tally (τ_b) is accumulated, which represents the number of times throughout the sequence the blocks' AC-DCT coefficient counts exceed the threshold 10 Since the scoreboard graphic is present on-screen for the majority duration, as the sequence progresses, scoreboard related pixel blocks should become obvious, as those exhibiting higher values of τ However, to extend this analysis towards a complete scheme for the recognition of scoreboards, it is required to have knowledge of typical scoreboard size

For each of the fourteen different formats observed in the (CIF resolution) training-corpus, the number of pixel blocks used was determined, and this data is tabulated in **Table 5 3** From this data, it was noted that the mean number of pixel blocks required to represent training corpus scoreboard graphics was 48 Based on this average, for a given FSV sequence, the 48 blocks that exhibit the highest values of τ have a high probability of constituting the scoreboard, and are deemed the ***potential scoreboard blocks*** (**PSBs**) Finally, it is further proposed that of the 48 detected PSBs, the ***recognized scoreboard blocks*** (**RSBs**) of a sequence correspond to those that constitute the largest spatially connected group

For a given broadcast, the mode luminance values of the RSB pixels (RSBPs) are computed across all images of the entire sequence On the basis of the resulting scoreboard template, frame-level mode variance measures are computed as described in *Section 4 4 4 3*, i e by quantifying the inconsistency between the quantised luminance values of the mode RSBPs and current image RSBPs, where the quantisation levels are as shown in **Table 5 4**

Table 5 3 Pixel block counts for 14 observed scoreboard formats

Scoreboard	# Blocks	Scoreboard	# Blocks
A	57	H	55
B	35	I	47
C	42	J	53
D	48	K	49
E	56	L	41
F	48	M	36
G	59	N	53
Average Size = 48 blocks			

Table 5.4. Five bin quantisation of [0-255] luminance spectrum

Band	Interval
Very Dark	0-50
Dark	51-100
Grey	101-154
Bright	155-204
Very Bright	205-255

It was anticipated that this process, as implemented above, should provide for the reliable detection of scoreboard suppression. However, following a closer investigation of the training-corpus scoreboards, it was noted that it is not uncommon for many of the graphics to exhibit some degree of transparency. This is usually performed to limit the occlusion disturbance to the viewer. A consequence of this is that RSBP luminance values are subject to transparency-noise, which can destructively interfere with the mode-discrepancy count in **(4.9)**. Hence, to combat the effects of potential transparency-noise on the analysis, the contrast of the luminance spectrum [0-255] of the RSBPs is warped (enhanced) prior to quantisation, such that the effects of fleeting luminance variations are suppressed. Specifically, a 256-bin scaling operator characteristic based on a 180° cycle period of the sine function is used to perform this task – see **(5.5)**. This characteristic, is illustrated in **Fig. 5.12**.

$$1 + Sin(\omega) \quad ; \quad \frac{3\pi}{2} \leq \omega \leq \frac{5\pi}{2} \tag{5.5}$$

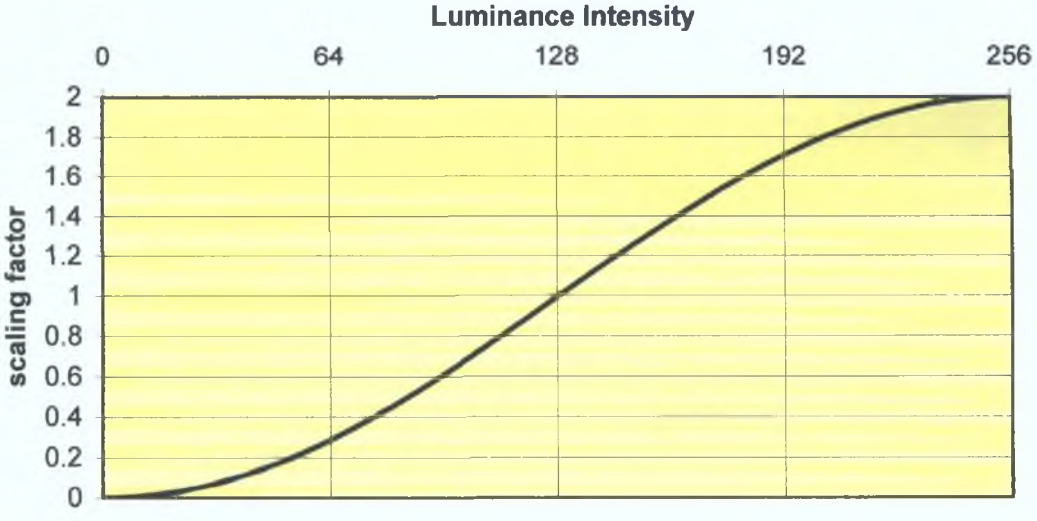


Fig. 5.12. Contrast scaling characteristic, based on 180° cycle of sine function.

The effect of this scaling operation in the luminance domain is to push reasonably dark RSBPs to very dark, reasonably bright RSBPs to very bright, while leaving mid-luminance values relatively unaffected Note, resultant pixel values that reside outside the permitted range [0-255] are clipped accordingly **Fig 5 13** illustrates the luminance component of an extracted scoreboard and its contrast-enhanced equivalent Hence, prior to the quantised discrepancy count, the pixel luminance values are contrast-enhanced in this way That is, the mode-variance analysis is actually performed in the quantised contrast-enhanced luminance domain of **(4 9)**

To specifically implement the extraction of MVM measures as described, a software tool called ***ScrbrdMVMextract*** was designed and built in the C programming language Given an MPEG video sequence to be analysed, ***ScrbrdMVMextract*** exploits both low-level AC-DCT data, and pixel luminance data as input (see *Sections B 2* and *B 4* of *Appendix B* for information on how both the DCT coefficients and pixel luminance data was extracted), in yielding resultant MVM values for each frame analysed The effectiveness of this tool is evaluated in the following section

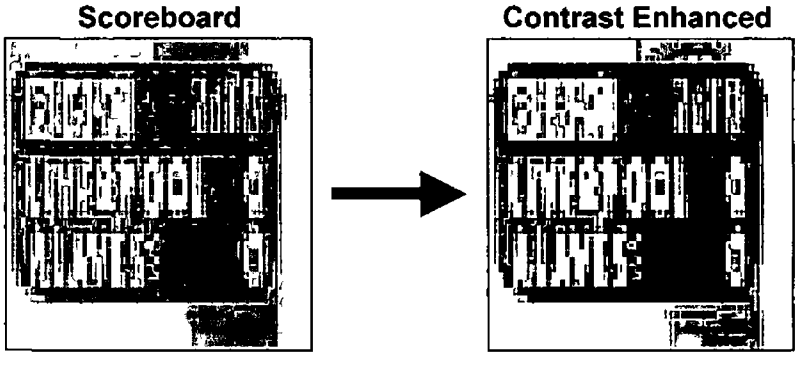


Fig 5 13 The luminance component of an extracted scoreboard and its contrast-enhanced equivalent

5 1 4 2 Effectiveness

Image-1 in **Fig 5 14 A** was selected from a training-corpus hockey-video By analysing the AC-DCT luminance coefficients of the I-frames of this sequence, the 48 PSBs were discerned based on their respective values of τ, as described above The 48 detected PSBs are illustrated in image-2 of this figure From the PSBs, the RSBs were determined as those constituting the largest spatially connected group In this case there are 46

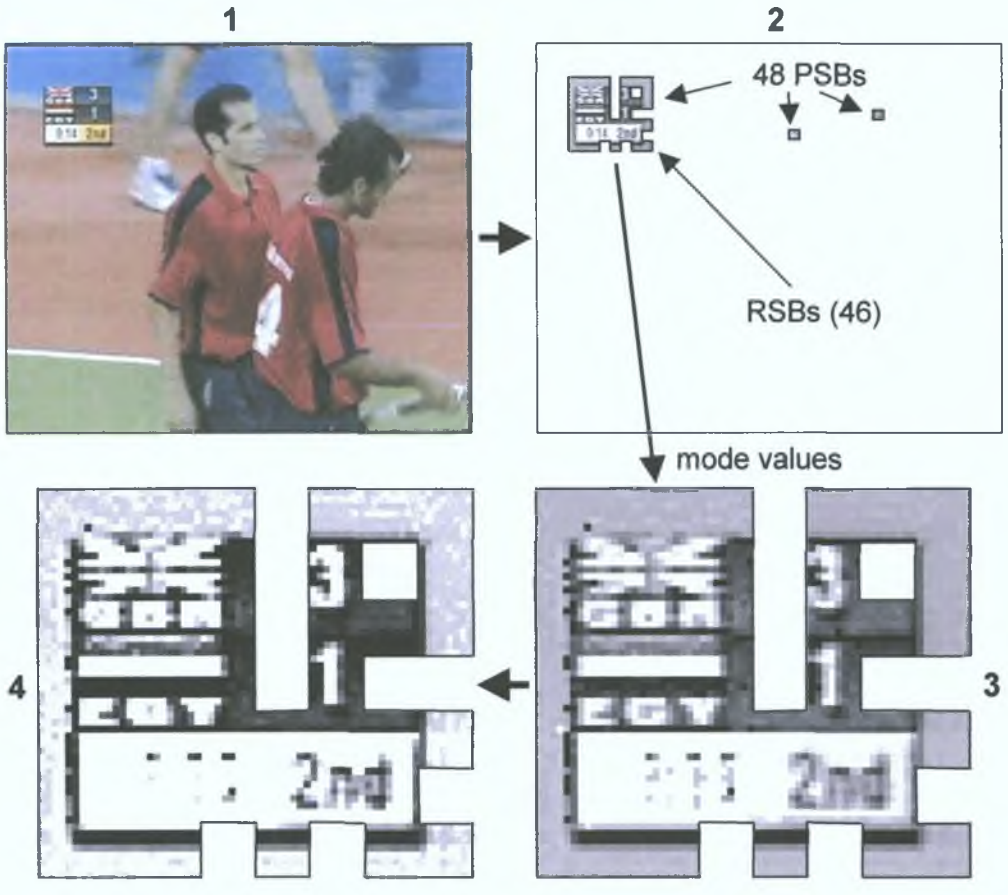


Fig. 5.14.A. Image-1: video image from a training corpus hockey video sequence. Image-2: PSBs and RSBs. Image-3: RSBP luminance mode values for the sequence. Image-4: contrast-enhanced RSBP mode values.

RSBs, and as can be seen from the illustration, these correlate well with the actual scoreboard position. The mode luminance values of the RSBPs, computed across the I-frames of the sequence are illustrated in image-3. These values were subsequently scaled using the contrast-enhancement operator of **(5.5)**, and the resultant contrast-enhanced mode RSBP luminance values are illustrated in image-4. **Fig. 5.14.B** presents two successive I-frames, which were extracted from the same hockey sequence. In the first (image-A) the scoreboard is on-screen, however in the second (image-B) it has been suppressed for update. In each case, the luminance values for the detected RSBPs were extracted. Images-A1/B1 illustrate these for the cases of images-A and –B, respectively. Similarly, these values were scaled using **(5.5)** and the resultant contrast-enhanced RSBP luminance values are illustrated in images-A2/B2, respectively. For each case (A2/B2), such were compared with the contrast-enhanced RSBP luminance values of the RSBP

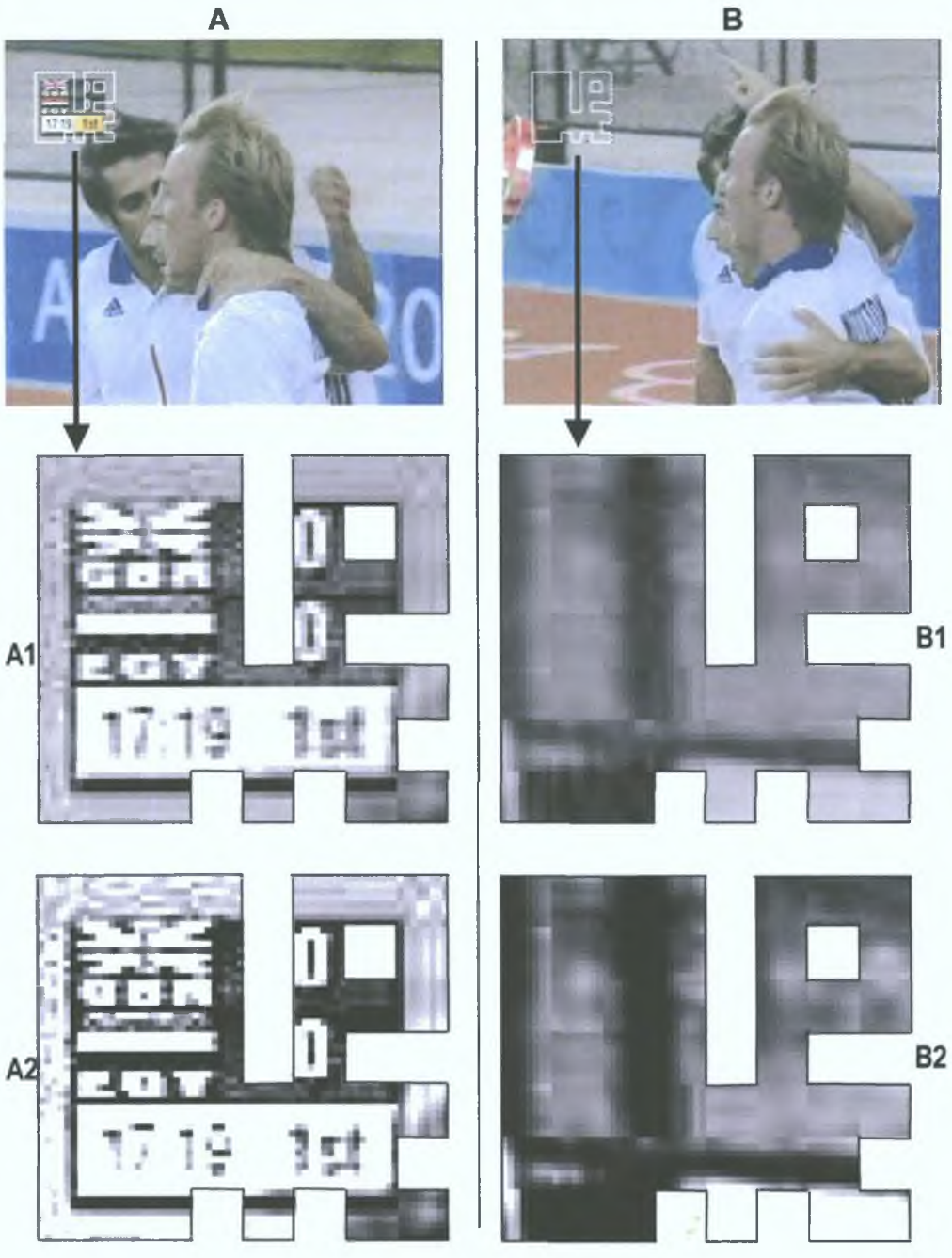


Fig. 5.14.B. Two successive I-frame images (A & B), the luminance pixel values of their RSBPs (A1 & B1), and their contrast-enhanced equivalents (A2 & B2).

sequence mode values (image-4 of Fig. 5.14.A). The number of discrepancies between these were determined, and using **(4.9)**, it was established that $\mathbf{MVM^A = 0.4406}$ and $\mathbf{MVM^B = 0.9891}$. From this data it is evident that there is at least a factor of 2 difference between the respective MVM values for the scoreboard present and scoreboard suppressed cases.

Image-1 in **Fig. 5.15.A** was extracted from a training-corpus rugby sequence. Image-2 illustrates the detected PSBs, RSBs, and the contrast-enhanced RSB mode values. **Fig. 5.15.B** presents two successive I-frames from this same sequence. Again, in the first (image-C) the scoreboard is on-screen, however in the second (image-D) it has been suppressed for update. Also illustrated for each case are the contrast-enhanced luminance values of the detected RSBPs of the sequence. For each of the images (C & D), the contrast-enhanced RSBP luminance values were compared with those of the RSBP mode values of the sequence. Using **(4.9)**, it was determined that $\mathbf{MVM^C = 0.4215}$ and $\mathbf{MVM^D = 0.9162}$, which again exhibit a factor of at least two difference.

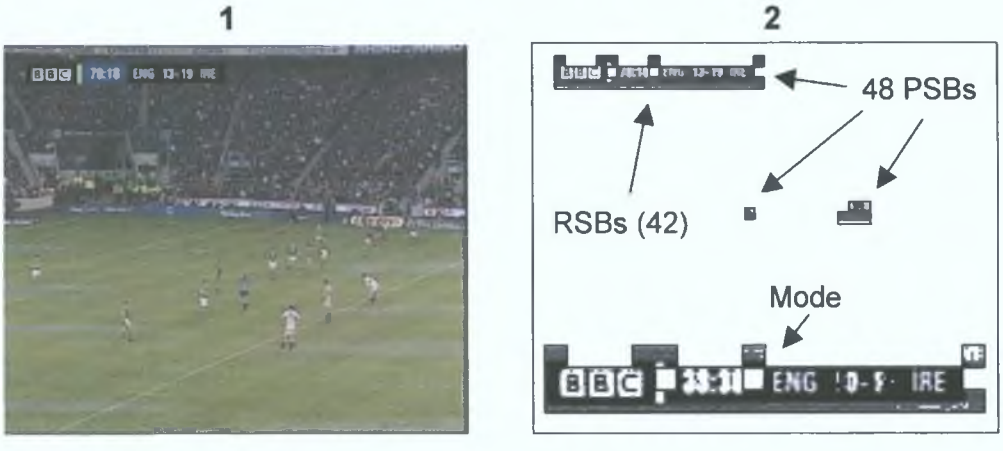


Fig. 5.15.A. Image-1: video image from a training-corpus rugby video. Image-2: detected PSBs, RSBs, and contrast-enhanced RSBP mode values.

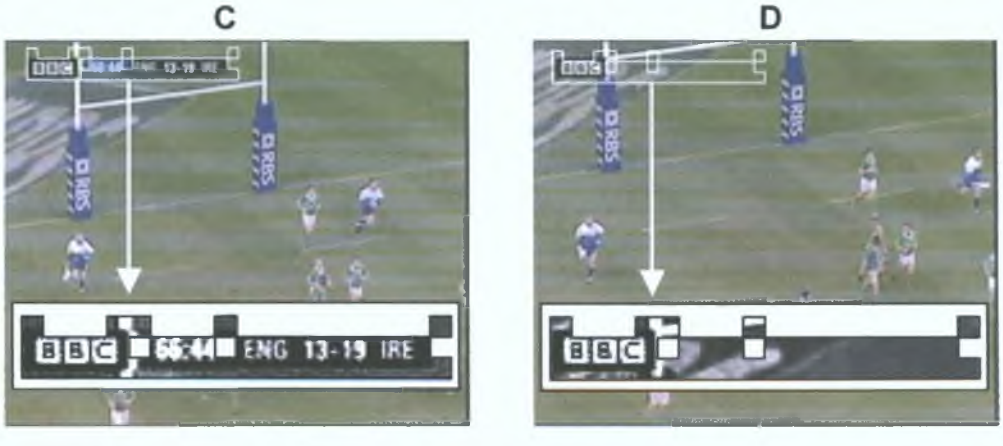


Fig. 5.15.B. Images C & D: two successive I-frame images. Inserts: the contrast-enhanced luminance pixel values of their RSBPs.

Image-1 of **Fig. 5.16.A** was extracted from a training-corpus soccer-video. Image-2 illustrates the detected PSBs, RSBs, and the contrast-enhanced RSBP mode values. Images-E & -F of **Fig. 5.16.B** are two successive I-frames from this same sequence, during the interval between which the scoreboard is suppressed. Also illustrated are their contrast-enhanced luminance values for the RSBPs of the sequence. Likewise it was determined that $\mathbf{MVM^{E} = 0.5781}$ and $\mathbf{MVM^{F} = 0.9824}$.

Finally, image-1 of **Fig. 5.17.A** was extracted from a training-corpus Gaelic football-video. Image-2 illustrates the detected PSBs, RSBs, and the contrast-enhanced RSBP mode values. Images-G & -H in **Fig. 5.17.B** are two successive I-frames from this same sequence, between which the scoreboard is suppressed. Also illustrated are their contrast-enhanced luminance values for the RSBPs of the sequence. In a similar fashion it was determined that $\mathbf{MVM^{G} = 0.5064}$ and $\mathbf{MVM^{H} = 0.8177}$.

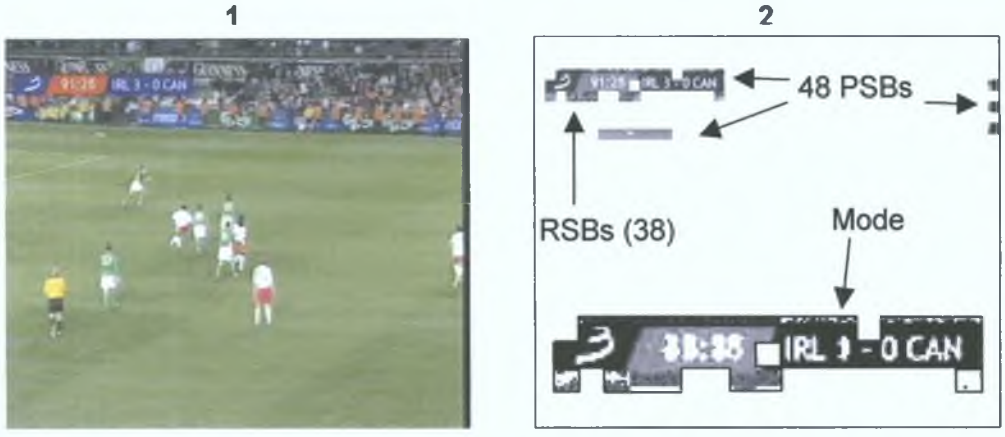


Fig. 5.16.A. Image-1: video image from a training-corpus soccer video. Image-2: detected PSBs, RSBs, and contrast-enhanced RSBP mode values.

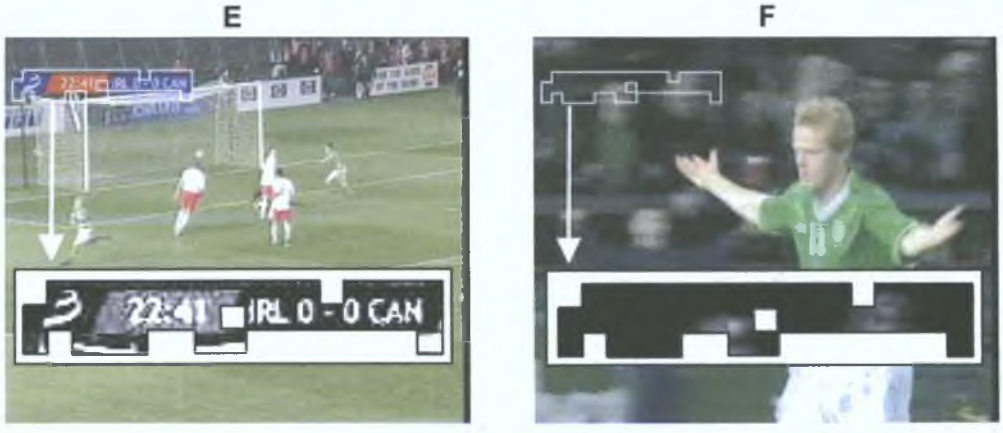


Fig. 5.16.B. Images E & F: two successive I-frame images. Inserts: the contrast-enhanced luminance pixel values of their RSBPs.

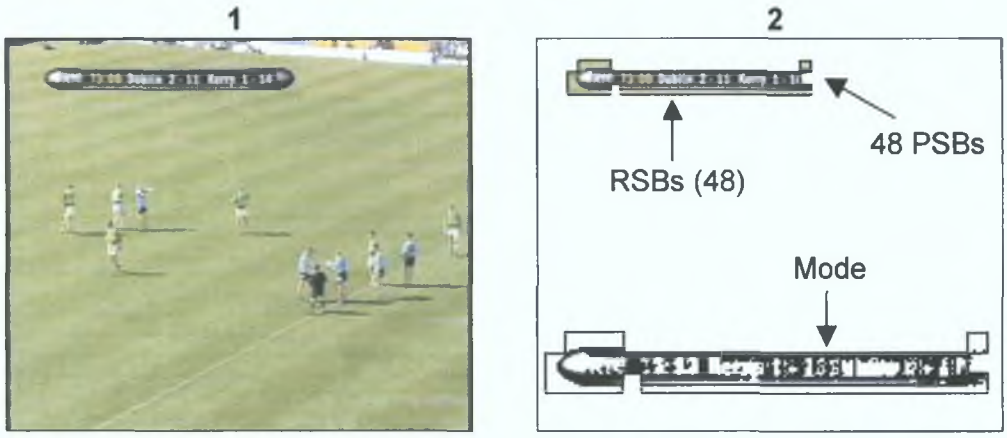


Fig. 5.17.A. Image-1: video image from a training-corpus Gaelic football video. Image-2: detected PSBs, RSBs, and contrast-enhanced RSBP mode values.

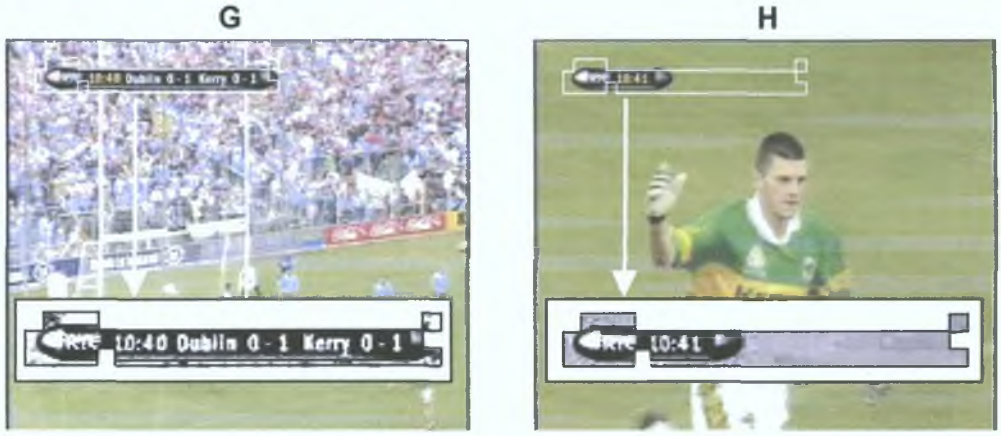


Fig. 5.17.B. Images G & H: two successive I-frame images. Inserts: the contrast-enhanced luminance pixel values of their RSBPs.

Table 5.5 presents a summary of the MVM values for the illustrated examples. From this data it is evident that for the scenarios illustrated, there is a consistent magnitude variance between the MVM values for the suppressed and present cases. On this basis, it is concluded that the scheme provides for the generation of MVM values that reliably infers the confidence of scoreboard suppression in FSV content.

5.1.5. CF5: Visual Activity Measure (VAM)

As described in *Section 4.4.5*, in terms of estimating visual activity, it was proposed for the reasons outlined, that the quantification be focused towards that of intense visual activity, while ignoring smooth camera motion.

Table 5 5 A summary of the MVM values for the illustrated examples

Image	Scoreboard	MVM
A	Present	0 4406
B	Suppressed	0 9891
C	Present	0 4215
D	Suppressed	0 9162
E	Present	0 5781
F	Suppressed	0 9824
G	Present	0 5064
H	Suppressed	0 8177

In *Section 4 4 5 2*, it was explained how, given a particular encoded video representation, there normally exists bitstream components of such that lend themselves to exploitation towards providing means for the quantification of visual activity As mentioned previously, the representation used in this work is MPEG-1, and therefore in terms of quantifying visual activity measures (VAMs) as required, the procedure implemented involves the analysis of motion vector (MV) data, as described in *Section 4 4 5 2*

5 1 5 1 Implementation & Parameter Settings

Although MVs are provided for both P- and B-frames in MPEG-1 video, given that typical GOP structure is used (see *Section 3 6 2 6*), and the framerate of the data corpus is 25fps, it was proposed that in terms of sampling the dynamics of visual activity from the video content, it should be sufficient to rely on MVs extracted from P-frames alone In terms of implementing the process of visual activity quantification from P-frame MVs, it is described below how a non-zero MV count is calculated, the resultant of which is representative of the frame's overall visual activity level This statistic is similar to that developed by Sun et al [82] However, a novel addition is that by employing a relatively large 'zero' threshold, it is proposed that this metric should be capable of discriminating between smooth camera motion and intense visual activity, as required

Recall that associated with each vector pair is the attribute of magnitude, which may computed as in **(5 6)**

$$v = a\vec{i} + b\vec{j} \quad , \quad |v| = \sqrt{a^2 + b^2} \qquad \textbf{(5 6)}$$

Hence, to numerically quantify visual activity, a critical statistic, i e the ***non-zero motion vector count*** **(NZMVC)**, is proposed based on this attribute Specifically, for a predicted frame, the NZMVC is determined by counting the number of macroblocks within the frame whose MV magnitude exceeds that of a pre-selected ***'zero'-threshold*** **(Z)** However, recall that by default, intracoded (i-) macroblocks are assigned zero-length MVs by the encoder, but as outlined in *Section 4 4 5 2*, i-macroblocks do not represent zero motion Therefore in quantifying frame activity, the abovementioned statistic must be augmented such that the incidences of i-macroblocks are accounted for in the calculations That is, for a predicted frame, its NZMVC is thus defined as the number of non-zero predicted macroblocks (i e whose MV magnitude is greater than **Z**), plus the number of i-macroblocks used to encode the image – see **(5 7)**

$$NZMVC = \#NonZeroMacroblocks^{P} + \#Macroblocks^{I} \qquad \textbf{(5 7)}$$

Given the NZMVC for a predicted frame, this statistic is then normalized by the total number of macroblocks used to encode the image, yielding its ***visual activity measure*** **(VAM)**, as shown in **(5 8)**

$$VAM = \frac{NZMVC}{TotNumMacroblocks} \qquad \textbf{(5 8)}$$

In [82] the authors propose that p-macroblocks may be reliably categorized into zero and non-zero types by defining a 'zero'-threshold that corresponds to the average of the observed MV magnitudes The authors maintain that in using this scheme, the activity dynamics of a generic video signal should be reliably characterized However, towards targeting intense visual activity as described, it is proposed that if **Z** is chosen large enough, it may be feasible for slow, smooth, far-field motion to be ignored, whilst jerky, uneven, near-field motion is detected To facilitate the selection of a suitably large threshold, the following training-corpus evaluation was undertaken

A number of global-view segments were extracted in equal proportions from the multiple FSV genres of the training-corpus For these segments alone, P-frame VAMs were calculated as **Z** traversed the range [$0 \leq Z \leq 100$] A similar analysis was performed for the reaction-phase content of an equal number of training-corpus SUEs **Fig 5 18** presents the variances of the average peak P-frame VAM observed with Z, for both cases From this figure, it is evident that, although the disparity is large throughout a range of values, the maximum disparity observed between the average values for the

Fig. 5.18. Variance of average P-frame VAM with Z, for global-view content and SUE reaction-phase content respectively.

specific content of the two scenarios, corresponds to a 'zero' threshold of 45 (indicated), and hence it is this maximally discriminating threshold (i.e. **Z=45**), is that used in this implementation.

In implementing the above procedures for the extraction of VAMs, a software tool called ***VAM_extract*** was designed and built in the C programming language. For a particular input video, given the P-frame MV data as input (see *Section B.3* of *Appendix B* for information on how the motion vector data was extracted), ***VAM_extract*** yields VAM values for each P-frame of the sequence. Since this approach to visual activity quantification operates purely on compressed bitstream data, it should exhibit excellent computational efficiency. The following section evaluates its effectiveness for the prescribed task.

5.1.5.2. Effectiveness

Fig. 5.19 presents video images extracted from the three primary camera views of a training-corpus rugby-video, which correspond to shots deemed to exhibit a level of motion activity typically characteristic of the views concerned. For each case both predicted and reference frames are presented. Given the temporal interval between the predicted/reference images and the object distance, it is clear that for the global-view case the motion between the predicted and reference frames is very slight. In the zoomed-in view the objects are visibly larger. Hence, in this case the activity between

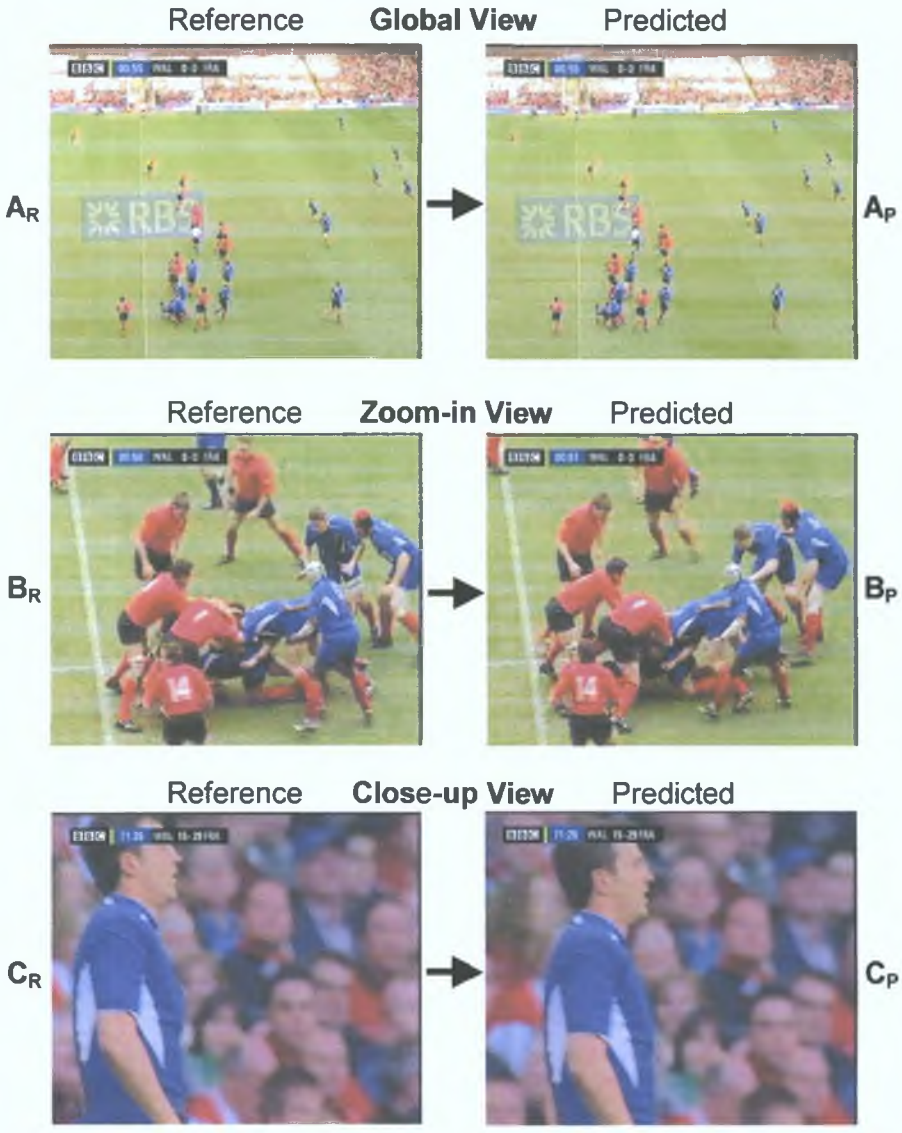


Fig. 5.19. Reference and predicted frames extracted from the three standard views of a training-corpus rugby-video sequence.

predicted and reference frames is somewhat more discernable. However, it is evident that amongst all three cases, it is the near-field close-up view that exhibits the most pronounced object displacement between frames. MVs were extracted from the P-frames in each case. This evidence was then used as input to the ***VAM_extract*** tool and the corresponding data is tabulated in **Table 5.6**. From this data it is evident that, as expected, the number of i-macroblocks used to encode the predicted images increases as the level of camera zoom increases, from global view to close-up view. However, more significantly, the number of 'zero' length MVs is considerably lower for the motion in the near-field close-up view than that of the other two cases. Consequently, the overall NZMVC, and hence VAM, for this view is substantially higher than that of the others.

Table 5 6 VAM_extract critical data for three views of rugby sequence

Case	# mblks	# Non-Zero p-mblks	# i-mblks	NZMVC	VAM
A_R-A_P	396	16	2	18	0 0454
B_R-B_P	396	62	4	66	0 1666
C_R-C_P	396	316	19	335	0 8459

Fig 5 20 presents a similar analysis concerning a training-corpus Gaelic football-video Again, predicted and reference frame video images pertaining to dynamic content from the three primary camera views are presented As in the previous illustration, in the global view case the motion visible between the predicted and reference frames is relatively slight, in the zoomed-in view it is slightly more discernable, and in the near-field close-up it is most pronounced As before, MVs were extracted from the predicted frames in each case Similarly, this evidence was used as input to the ***VAM_extract*** tool and the corresponding data is tabulated in **Table 5 7** From this data it is again evident that the number of i-macroblocks used to encode the predicted images increases as the level of camera zoom increases Also, it is similarly apparent that the number of 'zero' length MVs is greatly lower for the motion in the near-field close-up view than that of the other two cases Correspondingly, the VAM of this view is substantially higher than that of the others

From these two illustrations it has been illustrated how the visual activity is quantified Furthermore, the effectiveness of the chosen 'zero-threshold' in discriminating the vigorous motion of close-up views from the relatively more subtle motion of other camera views has been demonstrated

Table 5 7 VAM_extract critical data for three views of Gaelic football sequence

Case	# mblks	# Non-Zero p-mblks	# i-mblks	NZMVC	VAM
D_R-D_P	396	33	0	33	0 0833
E_R-E_P	396	153	6	159	0 4015
F_R-F_P	396	322	22	344	0 8686

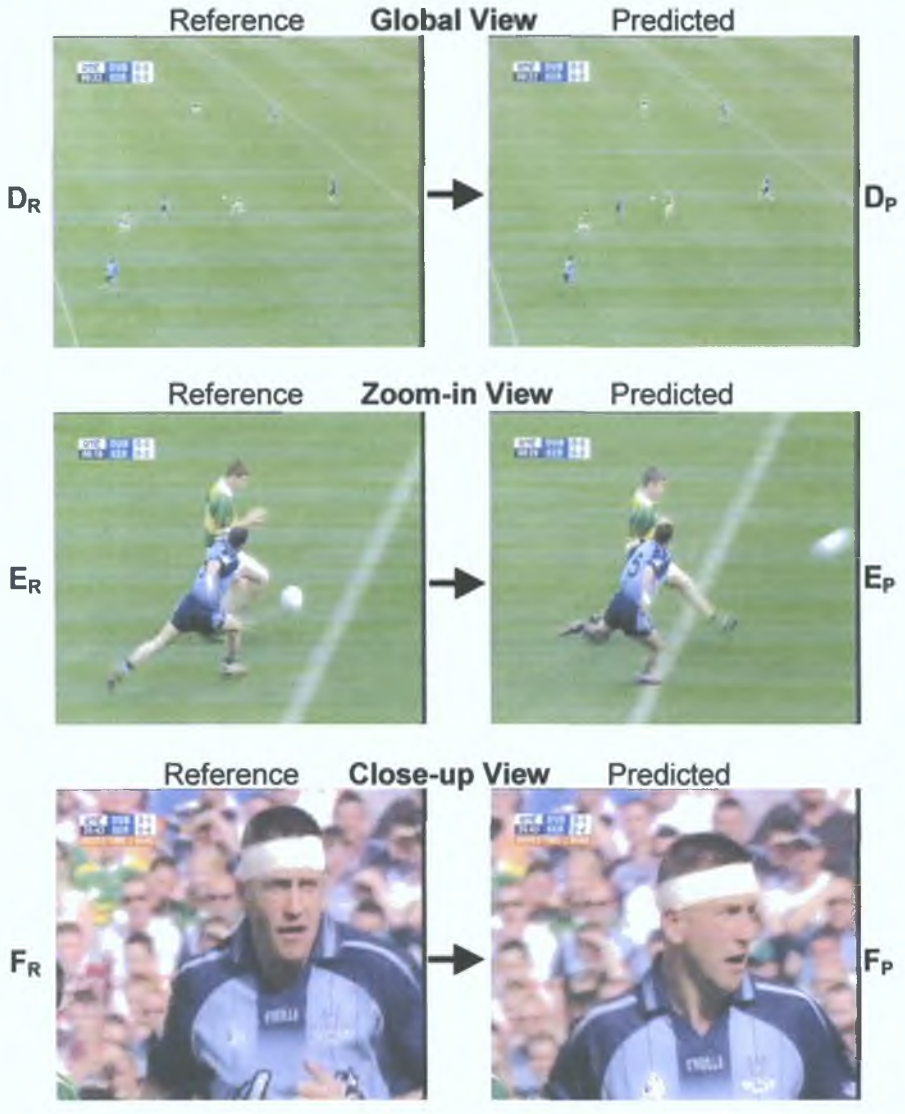


Fig. 5.20. Reference and predicted frames from three standard views of a training-corpus Gaelic-football video.

5.1.6. CF6: Field-Line Orientation Detection (θ)

In *Section 4.4.6*, an approach was proposed for the detection of the orientations of the most prominent field-lines in FSV images, which was based on the exploitation of pixel luminance/hue data, Roberts edge data, and Hough line space data.

5.1.6.1. Implementation & Parameter Settings

As described, in *Section 4.4.6.2*, it is required to select a hue tolerance η for the field pixel candidate (FPC) segmentation procedure. Towards selecting an appropriate value for the η multiple playing field grass samples were extracted in equal proportions from the

field-sport genres of the training-corpus. Following a calculation of respective values of the mode pixel hue value occurring for each broadcast, ψ, corresponding figures for grass pixel recall were generated for varying values of η. **Fig 5.21** illustrates the averaged results of this analysis. From this plot it is evident that the average grass pixel detection reaches maximum recall prior to when $\eta = 20^{\circ}$. Hence this was deemed a suitable tolerance value for the extraction of FPCs. **Fig. 5.22** presents a video image from a training-corpus soccer-video. The value of ψ was determined for the corresponding sequence, and using the derived hue tolerance value, the FPCs were detected for this image as shown.

As alluded to in *Section 4.4.6.2*, it is desirable to filter the FPCs in order to suppress elements of noise. In the FPC segmentation process, a binary image pixel map is yielded, where binary-1 represents a FPC, and binary-0 otherwise. In terms of filtering these FPC segmentation masks, given the CIF image resolution used, it was proposed that such be filtered using a 2-D [5x5] sliding window, which performs an erosion process as follows. For each binary pixel bit (**b**), its filtered equivalent (**b'**), corresponds to the combined product of itself and all the other pixel bits contained within its surrounding [5x5] window, as shown in **(5.9)**.

$$b'_{x,y} = \prod_{i=x-2}^{x+2} \prod_{j=y-2}^{y+2} b_{i,j} \qquad \textbf{(5.9)}$$

This operation has the effect of suppressing positive FPC bits that are not wholly enclosed by positive neighbours to the degree defined by the window size. For a more detailed illustration of this process see *Appendix C*.

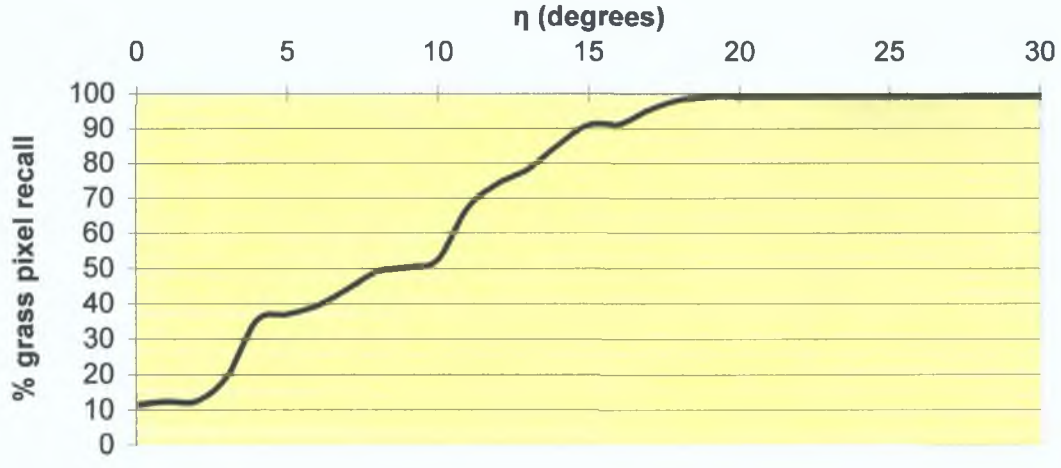


Fig. 5.21. Average grass pixel recall against η for training-corpus investigation.

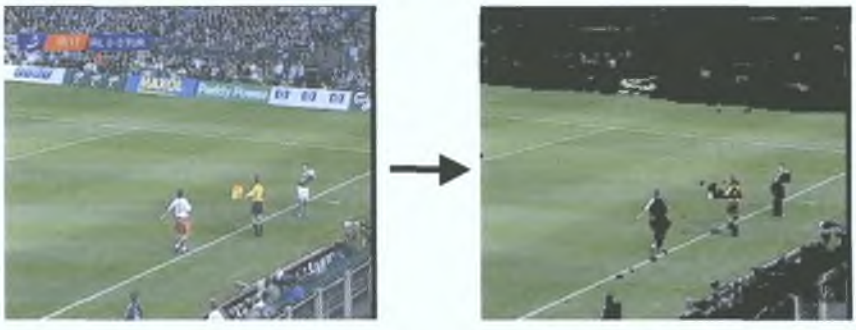


Fig. 5.22. Soccer-video image illustrating the segmentation of FPCs.

The FPC erosion process as described was applied to the FPC segmentation presented in Fig. 5.22, and the resulting output, i.e. the ***refined field pixel candidates*** (**RFPCs**) are illustrated in **Fig. 5.23**. In this example it is evident that following the erosion filtering procedure, noisy (non-grass) FPC pockets have been suppressed, while the majority of the true grass pixels have been retained. The only side effect of this process is that the frontier of the segmented field object is also slightly eroded. However, for a suitably sized window (such as the one used) this shrinkage should be negligible compared to the object size.

Following the luminance thresholding procedure described in *Section 4.4.6.3*, and the extraction of edges via the Roberts method, the Hough Line Transform was then applied to the edge detected binarised luminance RFPCs. The specific settings were a line angle step size ($\boldsymbol{\theta}$) of 1°, and normal ($\mathbf{d}$) length quantisation of 180 levels for $[0 \leq d \leq d_{max}]$ (where d_{max} was computed as ≈ 454 for the CIF resolution images used). From the resulting Hough space lattice values of the images, the angles of the most prominent field lines were discerned as those corresponding to the lines with the highest Hough space intersection tallies.

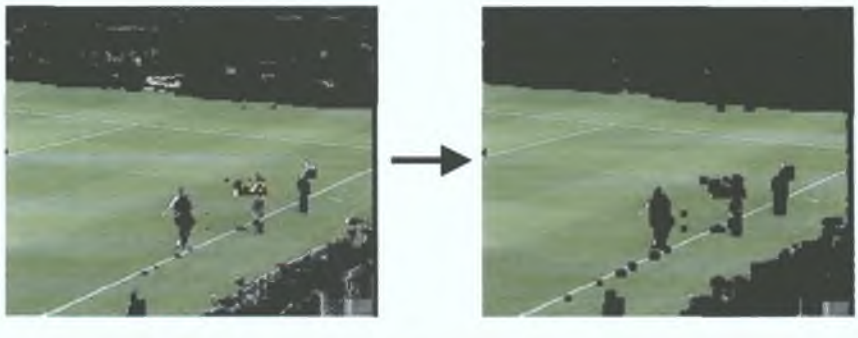


Fig. 5.23. Detected FPCs and FPC erosion yielding RFPCs.

To implement the above procedures a software tool called ***FieldLineOrientExtract*** was designed and built in the C programming language Given an image to be analysed, ***FieldLineOrientExtract*** exploits the appropriate signal-level data as described (see *Appendix B* for details) in detecting the angle of its most prominent field-line

5 1 6 2 Effectiveness

Fig 5 24 illustrates each processing stage involved in executing ***FieldLineOrientExtract*** on an extracted training corpus rugby-video image From this figure it is evident that the FPCs are reliably extracted, and the RFPCs satisfactorily suppress non-grass FPC pockets, while maintaining the majority of the playing field FPCs Using the adaptive broadcast-dependent threshold the luminance component of the RFPCs were binarised appropriately, such that the brighter playing field pixels (including the field-lines), were isolated from those constituting the darker grass

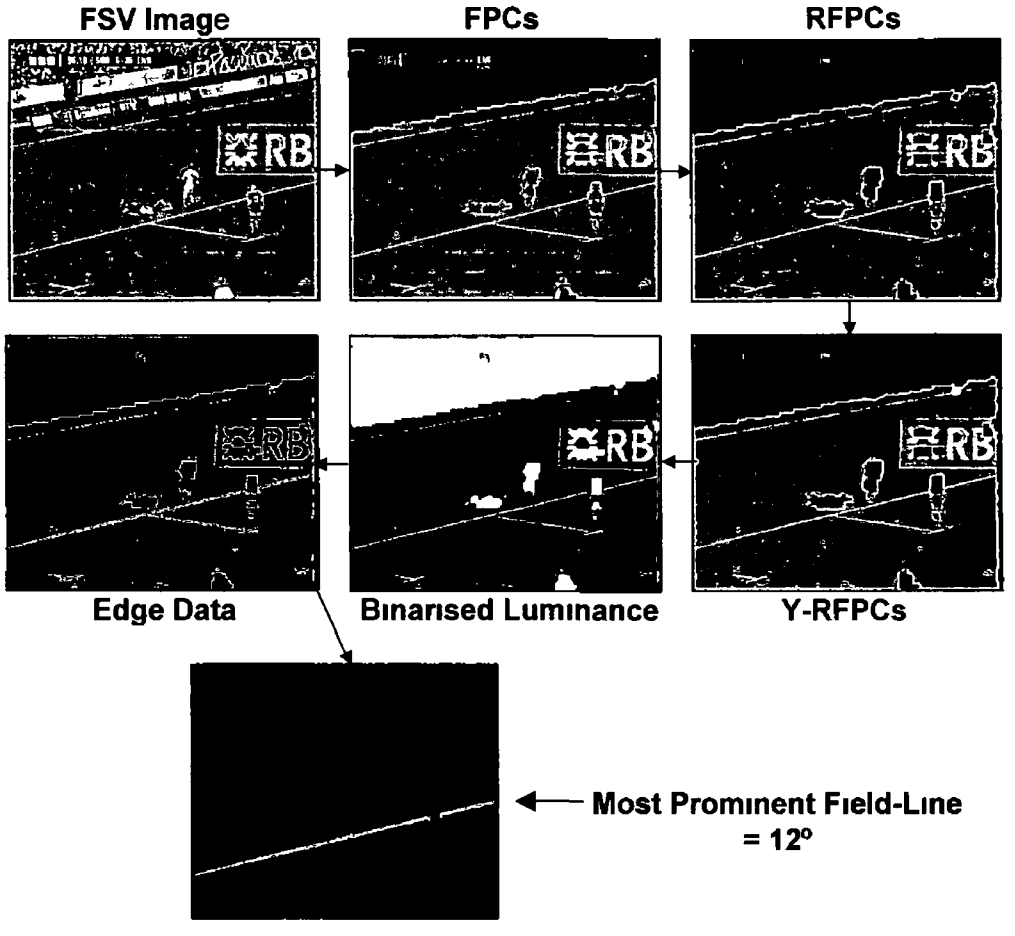


Fig 5 24 Extraction of most prominent field-line from rugby video image

Following the extraction of edges, and subsequently Hough line space data, the most prominent field-line was detected for this image as shown, by locating the highest intersection tally in the Hough space lattice as described. From the corresponding Hough space lattice angle index it was determined that this line has an orientation of 12° from the horizontal.

Fig. 5.25 illustrates the stages involved in similarly processing an extracted training corpus hurling-video image. Again, it is evident that the FPCs are reliably extracted, and that the RFPCs suppress many of the non-grass FPC pockets, while maintaining the majority of playing field FPCs. In this case it was determined from the corresponding angle index the detected line has an orientation of 17° from the horizontal.

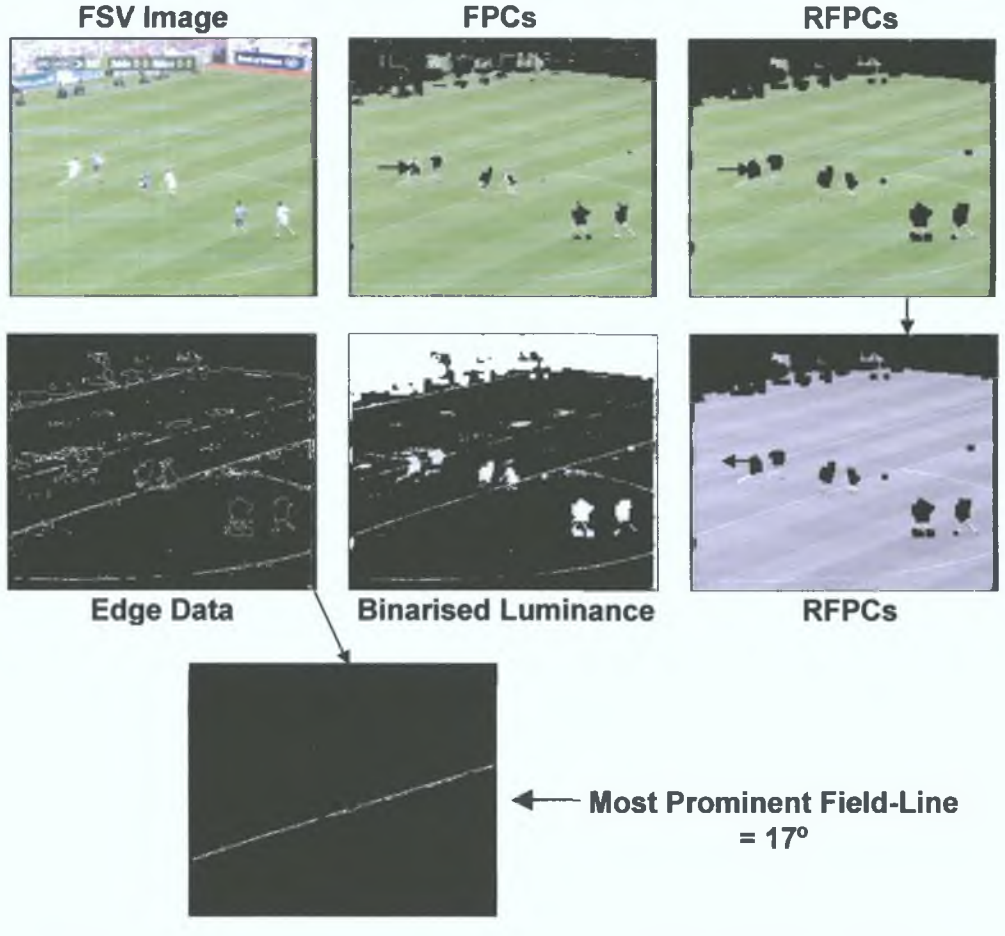


Fig. 5.25. Extraction of most prominent field-line from hurling-video image.

It has been demonstrated that the approach developed provides for the accurate extraction of the most prominent field-lines for the illustrations presented Furthermore, the line-angles suggested by the analysis concur with that of a manual verification On this basis, it is concluded that this scheme provides for the reliable extraction of this feature for FSV content

5.2. Implementation Of Shot Cut Detection

As described in *Section 4 5*, it was decided to employ an algorithm developed externally from this work [79] to implement the process of the detection of shot boundary transitions (cuts) In terms of its deployment within the implementation of the scheme herein, the algorithm settings were exactly as described in [79] Justification for reusing these settings is provided in *Appendix A*, where it is shown that it provides for very reliable detection of hard shot cuts in field-sports video content, which as explained in *Section 4 5*, constitute the large majority of the shot transitions

5.3. Implementation Of Pre-Processing Filter

As described in *Section 4 6*, the proposed pre-processing filter stage is comprised as a combination of two independent mechanisms, i e ad-break detection, and close-up-based shot rejection The ad-break detection scheme used is that developed externally to this thesis As explained, the scheme is biased towards precision, and in terms of its deployment in this work, the algorithm settings were exactly as described in [80] Recall that in terms of the close-up based shot filtering process it was proposed that for a given shot **i**, the maximum CuC exhibited by any of its respective reaction-phase seek window (RPSW) images ($[\mathbf{CuC}_{MAX}]_{RPSW_i}$), be compared *to* some threshold T_{CuC} towards determining whether the shot should be retained or rejected - see **(4 10)** Based on the following reasoning, in terms of implementing this procedure for the MPEG-1 data corpus in this work, it is proposed that probing at the I-frame level should be sufficient

As mentioned in *Section 3 6 2 6*, to combat the effects of error propagation in digital video, the group of pictures (GOP) structure must be have limited length For example, in MPEG-1 video the GOP length is typically restricted to between 10-18 frames Considering nominal MPEG framerate (25fps), this corresponds to an I-frame occurrence at least every 0 4s - 0 72s (i e a sub-second I-frame frequency) It was

required to compare this to shot length, and to this end, an investigation into the shot durations of the training corpus content was performed. **Fig. 5.26** presents the (logarithmic) average distribution of training-corpus shot durations. From this data it is evident that at least 99.9% of all observed shots exhibited a duration exceeding 1.0s – see dashed lines. Therefore, it follows that the vast majority of training-corpus shots contain at least one I-frame. On this basis, it is maintained that a sufficient resolution for the probing of RPSW close-up sequences corresponds to the I-frame level.

Experiments were then performed on the training-corpus such that an appropriate value for $\mathbf{T_{CuC}}$ be defined. **Fig. 5.27** illustrates how the proportion of retained training-corpus SUE-shots varies with the value of this threshold. From this data it is clear that as expected, when $T_{CuC} = 0$, the vast majority of all SUE-shots are retained irrespective of their associated RPSW I-frame CuC values. However as T_{CuC} increases, i.e. as the condition threshold becomes more stringent, the number of retained SUE-shots decreases accordingly. In accordance with the manually determined ideal, i.e. 98% of training-corpus SUE-shots were manually found to be followed by a close-up view (see *Section 4.2.2.1*), the optimum value for the threshold was proposed as the maximum value that provides for at least 98% SUE-shot retention. From the figure (see dashed lines) it was determined that a value corresponding to approximately that given in **(5.10)** provides this level of SUE-shot retention, and hence this is the T_{CuC} value employed in the implementation.

$$T_{CuC} = 0.08 \qquad \textbf{(5.10)}$$

Fig. 5.26. The (logarithmic) distribution of average training-corpus shot lengths.

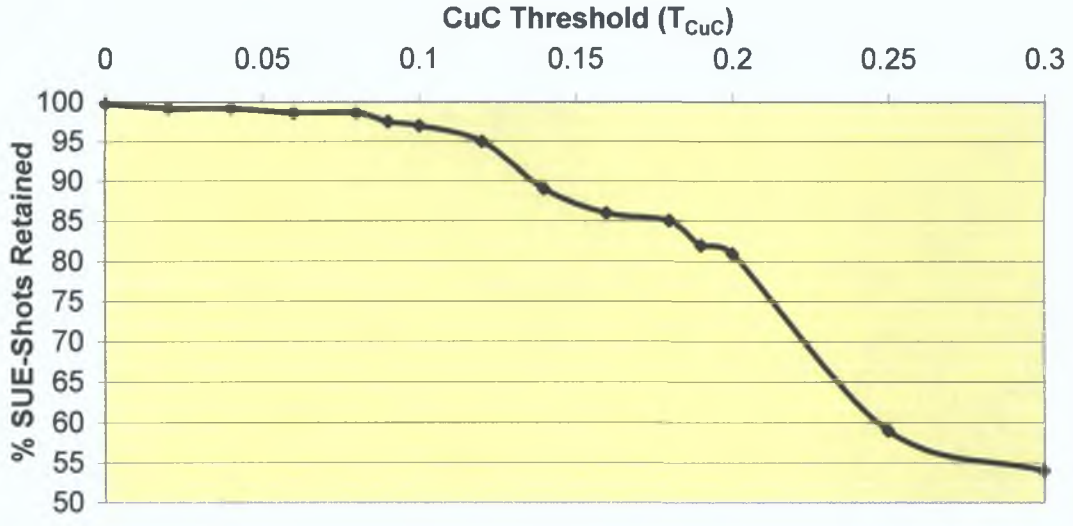


Fig. 5.27. Variance of SUE-shot retention with CuC threshold for training corpus.

It was noticed that, in relation to the assessment of the effectiveness of the close-up model presented in *Section 5.1.1.2*, the value in **(5.10)** is closer to the noise floor of the non-close-up images than it is to that of the true close-up views illustrated. It is proposed that this reflects the point made therein regarding the fact that the model is based on close-up images of a very well defined nature, and that the relatively low valued threshold thus reflects the large variance in the nature of close-up views from that of the ideal.

In terms of quantifying the individual and combined accuracy to which both of these preprocessor mechanisms realise their objectives, the filtering performances of such will be assessed as part of the overall presentation of experimental results.

5.4. Implementation Of Shot-Level Aggregation

As explained in *Section 4.7*, the shot-level aggregation stage concerns the process whereby the frame-level CFs are aggregated such that they constitute the vector component coefficients (VCCs) of an overall shot feature vector (SFV), which then forms the input for a higher-level pattern analysis phase in realizing the SUE-shot detection task. Given that CF1 evidence is exploited at the pre-processing stage, the shot-level aggregation stage is only concerned with CFs2-6. In terms of implementing this process, while the extraction of CF5 evidence is at the P-frame level (see *Section*

5 1 5 1), the extraction level of the remaining frame-level CFs is again chosen to be at the I-frame level (i e using the same justification as that given in *Section 5 3* above) Given this, the following subsections describe the specific implementation of the shot-level aggregation process for each CF (VCC) concerned

5 4.1 CF2 To VCC_1

The implementation of the methodology for the extraction of crowd image confidence (CIC) values from FSV images (i e CF2) was illustrated in *Section 5 1 2* On this basis, given a FSV, CIC values are calculated for each I-frame of the sequence using the tool ***CrowdConfExtract*** This yields feature dataset $[CF_2]$ for the sequence, as shown in **(5 11)**

$$[CF_2] = \{CIC\}_{I-Frames} \quad \textbf{(5 11)}$$

To maximally represent the likelihood that a given shot exhibits a reaction-phase crowd image instance, $\mathbf{VCC_1}$ is defined as the maximum I-frame CIC value found within its post-SEB RPSW That is, for a shot i, amongst the I-frames found within $RPSW_i$, $\mathbf{VCC_1}'$ is computed as shown in **(5 12)**

$$VCC_1' = \max [CF_2]_{RPSW_i} \quad \textbf{(5 12)}$$

5 4.2. CF3 To VCC_2

The implementation process for the extraction of (mean-filtered) speech-band audio levels (SBALs) at the video-frame level (i e CF3) was outlined in *Section 5 1 3* On this basis, for a given FSV, SBALs are calculated for each I-frame of the sequence using the tool ***SpeechBandEnergyExtract.*** This yields feature dataset $[CF_3]$ for the sequence, as shown in **(5 13)** However, to address the potentially sporadic variance of the mean audio signal levels across multiple broadcasts, the values of the $[CF_3]$ datasets are normalized to lie within the interval [0,1] for each case

$$[CF_3] = \{SBAL\}_{I-Frames} \quad \textbf{(5 13)}$$

To maximally represent the reaction-phase intensity of this feature for a given shot, $\mathbf{VCC_2}$ is defined as the maximum level found within its RPSW That is, for shot i, amongst the levels located within $RPSW_i$, $\mathbf{VCC_2}'$ is computed as shown in **(5 14)**

$$VCC_2^i = \max [CF_3]_{RPSW_i} \quad \textbf{(5 14)}$$

5.4.3. CF4 To VCC_3

The implementation for the extraction of scoreboard suppression confidence from FSV images (i e CF4) in the form of mode-variance measures (MVMs) was presented in *Section 5 1 4* On this basis, for a given FSV, MVM values are calculated for each I-frame of the sequence using the tool ***ScrbrdMVMextract*** This yields feature dataset [CF_4] for the sequence, as shown in **(5 15)**

$$[CF_4] = \{MVM\}_{I-frames} \quad \textbf{(5 15)}$$

Again, to maximally represent whether a given shot exhibits a reaction-phase scoreboard suppression instance, $\mathbf{VCC_3}$ is defined as the maximum I-frame MVM value found within its RPSW That is, for shot i, amongst the I-frames found within $RPSW_i$, $\mathbf{VCC_3^i}$ is computed as shown in **(5 16)**

$$VCC_3^i = \max [CF_4]_{RPSW_i} \quad \textbf{(5 16)}$$

5.4.4. CF5 To VCC_4

The procedures implemented for the extraction of visual activity measures (VAMs) from the P-frames of a FSV sequence (i e CF5) were outlined in *Section 5 1 5* On this basis, given a FSV, VAM values are calculated for each P-frame of the sequence using the tool ***VAMextract*** This then yields feature dataset [CF_5] for the sequence, as shown in **(5 17)**

$$[CF_5] = \{VAM\}_{P-frames} \quad \textbf{(5 17)}$$

As described in *Section 4 2 2 2*, the intense near-field visual activity associated with the SUE reaction-phase segments is, in the main, due to the prevalence of close-up views of celebrating players However, recall that also recognized as having an effect in increasing the post-SUE levels of this feature are the zoomed-in/close-up views typically used in the subsequent action replay segments, and the video effects sometimes used to delimit their multiple viewing angles Given this, in terms of maximizing the potential SUE discrimination, for the shot-level aggregation of [CF_5] evidence, it was considered desirable to quantify the extent of near-field visual activity recurrence within the RPSW, rather than probing for unique maximum instances To this end, for each shot of a FSV

sequence, the number of P-frames with VAM measures exceeding that of the sequence mean level is determined within its RPSW, and $\mathbf{VCC_4}$ is then set to a (normalized) value representing this P-frame count That is, for a shot i, $\mathbf{VCC_4}'$ is computed as shown in **(5 18)**

$$VCC_4' = \#[CF_5]_{RPSW_i} \geq \overline{[CF_5]} \qquad \textbf{(5 18)}$$

5.4.5. CF6 To $\mathbf{VCC_5}$

The procedures implemented for the extraction of the orientations (θ) of the most prominent field-lines from FSV images (i e CF6) were described in *Section 5 1 6* On this basis, for a given FSV, values for θ are calculated for each I-frame of the sequence using the tool ***FieldLineOrientExtract***. These angles then yield feature set $[CF_6]$ for the sequence, as shown in **(5 19)**

$$[CF_6] = \{\theta\}_{I-Frames} \qquad \textbf{(5 19)}$$

As described in *Section 4 4 6*, it is required to exploit field-line orientation evidence towards quantifying the confidence that a given shot culminates with the camera focused on action situated in the end-zone region of the playing field

As outlined in *Section 4 4 6 1*, due to the routine use of global-views in capturing dynamic FSV action, field end-zone perspectives are characterized by the most prominent field-lines exhibiting angles within a specific interval (see Fig 4 12) To enumerate this interval, an investigation was performed, in which end-zone field-lines (as illustrated in Fig 4 12) were extracted from the training-corpus and were analysed manually The average distribution of the line orientations is presented in **Fig 5 28**, and from this graph it is evident that, only a negligible number of the field-line orientations mapped outside the interval **[5° - 25°]** In exploiting this characteristic towards the said objective, for a given shot, as its shot-end I-frame field-line orientations are found to lie within the key range, its corresponding $\mathbf{VCC_5}$ value should increase accordingly To realize this, $\mathbf{VCC_5}$ is set to a value representing the number of I-frames in a shot that exhibit θ in the critical range, where the contribution of each I-frame is weighted such that those nearest the shot-end boundary have most influence Given the framerate/GOP structure employed in the data corpus, the weighting function chosen (δ) was that based on the decreasing exponential given in **(5 20)**, which for illustration is plotted in **Fig 5 29** On the basis of this function, for a given shot with α I-frames,

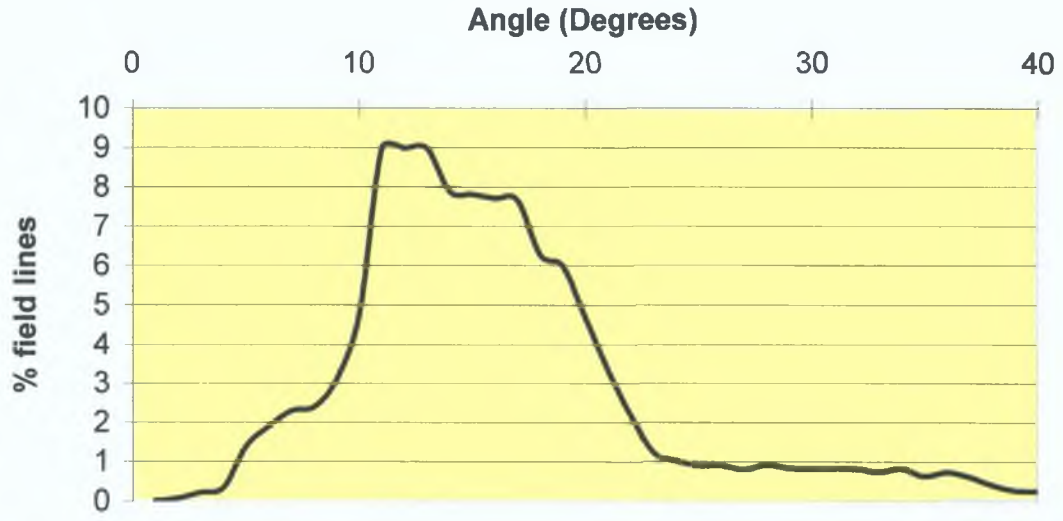


Fig. 5.28. Distribution of field-line orientations for field end-zone images extracted from training corpus.

starting at its SEB and working backwards, each encountered I-frame (i.e. n = 1, 2, … α), is assigned an associated weight, $\delta(x_n)$, which, as indicated in Fig. 5.29, will quickly decrease with increasing distance from the SEB. The $\mathbf{VCC_5}$ value of the shot is then computed as the (averaged) cumulative value of the weights of the I-frames that have θ

$$\delta(x_n) = \frac{1}{\exp(x_n)} \; : \; x_n = \{0, 0.15, 0.3, 0.45....\}, \forall n = 1, 2, 3,... \qquad \textbf{(5.20)}$$

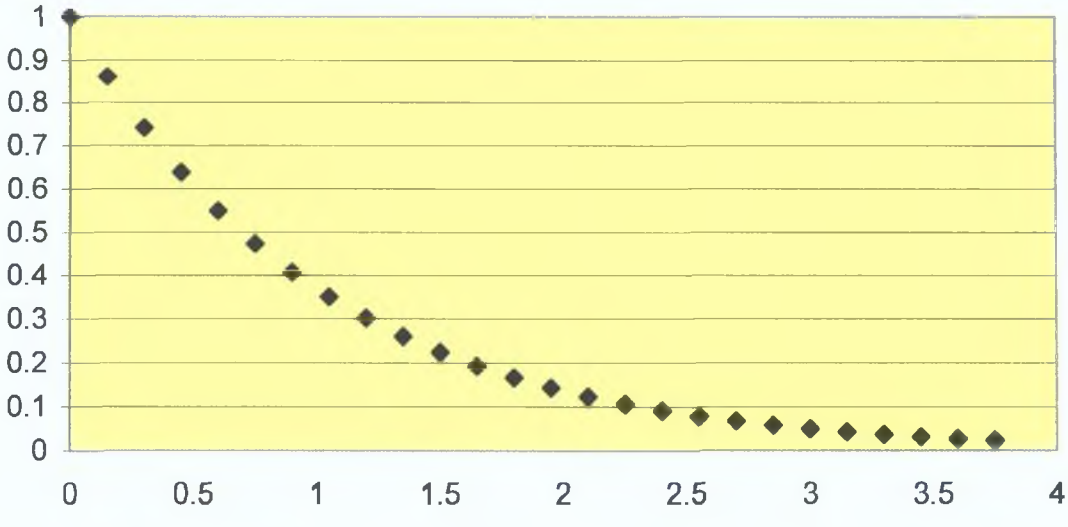


Fig. 5.29. Decreasing exponential function for the weighting of I-frame influence for VCC_5.

within the critical range That is, for a shot i with α_i I-frames, $\mathbf{VCC_5'}$ is computed as shown in **(5 21)**, where n indexes $[CF_6]$ in a backwards direction from SEB_i

$$VCC_5' = \frac{\sum \delta(x_n)}{n} \quad for\ all\ n\ (\leq \alpha_i)\ with\ \ 5° \leq [CF_6]_n \leq 25° \qquad \textbf{(5 21)}$$

5.5. Overview

In *Section 5 1*, it was described how the extraction methodologies of each CF were implemented It was then described in *Section 5 3*, how the exploitation of CF1 evidence at the preprocessor stage was implemented Next, the implementation procedures for the shot-level aggregation processes of CFs2-6 were described in *Section 5 4* To provide an overview of the specifics of the implementation, **Table 5 8** provides a list of all thresholds and conventions employed, coupled with a commentary on each in terms of their invariance or otherwise to different video scenarios and/or application constraints For additional clarity, an overview of the six frame-level critical features proposed, the signal-level data upon which their extraction methodologies are based, and a description of their corresponding shot-level exploitation/aggregation is provided in **Table 5 9** As mentioned earlier, in terms of quantifying the accuracy to which the exploitation of CF1 evidence as a pre-processing filter realises its objective, such will be assessed during the presentation of the experimental results Whereas, an investigation into the discriminatory capabilities of each individual VCC in terms of training data SUE-shot discernment is explored in *Section 6 2*, in the context of choosing an appropriate pattern classification technique

5.6. Chapter Summary

In this chapter, it was described how each element of the hypothesis for the summarisation solution proposed in *Chapter 4* was implemented Specifically, the implementation and parameter settings for the proposed extraction methodologies for each frame-level critical feature were described Furthermore, the effectiveness of each implemented CF extractor was assessed It was then described how the pre-processing filter was implemented in terms of selecting an appropriate threshold for the close-up-based shot retention condition Finally, it was described how evidence pertaining to the

remaining critical features is aggregated towards generating the critical shot feature vectors.

Table 5.8. List of system thresholds/conventions.

Threshold/Convention	Comments
Reaction-phase seek window (RPSW = 24s). (*Section 4.2.2*)	Determined by inspection across five diverse field-sports-video genres. Biased towards maximising recall.
ROE for close-up image model. (*Section 5.1.1.1*)	Derived empirically. Invariant to image resolution and/or video format.
Hue tolerance for close-up image model, $\xi = 10°$, (*Section 5.1.1.1*).	Derived empirically. Given that the levels/range of hue are consistent for any level of YUV quantisation (i.e. 0°-360°), this threshold should provide good results for any video scenario.
ROI for crowd image model. (*Section 5.1.2.1*)	Invariant to video format and/or image resolution.
Subband selection for speech-band model (i.e. subbands 2-7) (*Section 5.1.3.1*)	Rooted in the nature of the psychoacoustic model used in MPEG-1 Layer-II audio, which decomposes the audio spectrum into 32 equal subbands. Recalibration (and possibly redesign) required for alternative encoded audio formats.
1-D temporal sliding window for mean-filtering of video-frame speech-band audio levels (0.5s = 13 frames) (*Section 5.1.3.1*)	Recalibration required for video framerates differing from the standard 25fps rate characterising the corpus used.
Minimum AC-DCT coefficient count for potential scoreboard pixel blocks (= 10). (*Section 5.1.4.1*)	Invariant to any video format based on DCT encoding of [8x8] pixel blocks, e.g. MPEG, H.26x. However, recalibration of threshold possibly required for bitrate constraints lower/higher that those characterising the corpus used (i.e. from which the current value was derived).
Number of Potential Scoreboard Blocks (= 48) (*Section 5.1.4.1*)	Empirical average for the corpus used. Based on the decomposition of video images into [8x8] pixel blocks (e.g. MPEG, H.26x). Recalibration required for scenarios with non-CIF video image resolution.
Contrast enhancement warping. (*Section 5.1.4.1*)	Recalibration required for luminance quantisation levels differing from that characterising the corpus used (i.e. $0 \leq Y \leq 255$)
'Zero-threshold' (Z = 45) for non-zero motion vector count. (*Section 5.1.5.1*)	Derived empirically. Recalibration possibly required for non-CIF video image resolution, and/or where alternate (non-standard) motion estimation techniques/constraints are employed.
Hue tolerance for field-pixel candidate extraction, $\eta = 20°$, (*Section 5.1.6.1*).	Derived empirically. Given that the levels/range of hue are consistent for any level of YUV quantisation (i.e. 0°-360°), this threshold should provide good results for any video scenario.

2-D spatial sliding window for field-pixel erosion (= [5x5] pixels) (*Section 5 1 6 1*)	Best-fit window chosen for image sizes used Recalibration required for non-CIF video image resolution
Pre-processor close-up image threshold (Γ_{CuC} = 0 08) (*Section 5 3*)	Best-fit threshold determined by experimentation Invariant to different video scenarios
Critical angles for field lines (5°-25°) (*Section 5 4 5*)	Determined by inspection Invariant to video image resolution/format
Function (δ) for weighting [CF_6] towards the shot-end (*Section 5 4 5*)	Recalibration required for video framerates and/or GOP structures differing from those characterising the corpus used

Table 5 9 List of the six frame-level critical features, the signal-level data upon which their extraction methodologies are based, and a description of their corresponding shot-level exploitation/aggregation

Frame-Level Critical Features	Shot-Level Exploitation/Aggregation
CF1 Close-up image confidence (CuC) Derived based on pixel hue data	Used at the preprocessing stage Analysis of I-frame CuC values within the RPSW
CF2 Crowd image confidence (CIC) Derived based on edge data	VCC_1 max I-frame CIC within the RPSW
CF3 Speech-band audio level (SBAL) Quantified based on subband scalefactor data	VCC_2 max I-frame SBAL within the RPSW
CF4 Scoreboard suppression confidence (MVM) Derived based on DCT coefficients and pixel luminance data	VCC_3 max I-frame MVM within the RPSW
CF5 Visual activity measure (VAM) Quantified based on motion vector (+ macroblock type) data	VCC_4 (normalised) number of P-frames with VAMs exceeding that of the sequence mean level within the RPSW
CF6 Field-line orientation (θ) Extracted based on pixel hue/luminance/edge data and Hough line space data	VCC_5 (averaged) summation of shot-end biased I-frame weights for I-frames with θ in the key range

Chapter 6

Pattern Classification: A Support Vector Solution

The previous chapters dealt with describing the proposed hypothesis and the specifics of its implementation Therein, it was illustrated how CF1 evidence is exploited at the preprocessor stage, and how evidence pertaining to the remaining CFs (i e CFs2-6) is processed towards generating shot feature vectors (SFVs) This chapter addresses the issues regarding the task of SFV pattern analysis for the score-update episode shot (SUE-shot) classification process Firstly, the motivation for employing a machine-learning scheme is outlined, followed by an introduction to the general topic of machine learning and the various approaches employed Then, coupled with an exploration of the feature space of training-corpus SFV data, arguments for favouring a support vector solution are proposed

6.1. Machine-Learning

6 1 1 Motivation

As proposed in *Section 4 7*, the SFVs are to form the basis of the SUE detection approach That is, given the described hypothesis, it is anticipated that a perceptible discrepancy should exist between the SFV attributes for SUE-shots and those of non-SUE-shots However, the methodology of the mechanism used to reliably identify these pattern discrepancies remains to be addressed Broadly speaking, common approaches to data classification fall into one of two categories, i e rule-based heuristic schemes, and machine-learning solutions While for many cases rule-based approaches to data

classification have been shown to provide successful results, they tend not to be generic, and, in general, yield systems that are less robust Furthermore, it has been found that a mathematical prototype of a solution is sometimes unavailable, rendering classical programming methods ineffective in solving many of the problems encountered in scientific study [83] Moreover, even if a conventional algorithmic solution can be found, it may be sometimes so complex that the computation required may exceed the bounds of practicability Greater availability of both data and computational power has spurred the migration away from rule-based and manually specified models, towards statistical-based data-driven models Hence, given these issues, it was decided to employ a machine learning approach for the implementation of the data pattern classification task of this work

6 1.2. Machine-Learning Theory

Machine-learning involves the learning of a solution by programming computers to use sample data and/or past experience [84] It is most effective in cases where we cannot directly write a computer program, i e the program is too difficult to program by hand, but example data is available For example, consider the problem of handwritten character recognition Using a traditional algorithm methodology, it is considered extremely difficult, if at all possible, to design a computer program that can reliably identify e g the letter 'X' from an image However, there exist diverse instances of such within various handwritten alphabets, which may be coupled with a manually annotated ground-truth These examples, and their pattern consistencies, could thus conceivably form the basis of a learning approach towards the generation of a statistical-based solution Ideally, a number of both true and false 'X' examples (plus their associated ground-truth) are input to the ***learning machine*** (**LM**) On the strength of these bipolar examples, the LM aims to learn to recognize the general characteristics of the letter 'X', towards being able to reliably identify it amongst a pool of other data From this example it is clear to see why the machine-learning technique has been compared to how an infant becomes trained at reading, i e by being continually exposed to examples

6.1.3 Approaches to Machine-Learning

The subject of machine learning may be divided into two broad areas, i e unsupervised and supervised learning ***Unsupervised learning*** concerns data processing applications such as density estimation and clustering In this scenario, no training data input is

provided Instead, given a feature data space, this methodology is focused on discerning patterns within it and/or tangible relationships between the individual data points In ***supervised learning***, an annotated training data set is presented as input to the LM, as in the case of the described example above Then, across all input examples, the LM aims to infer the general correlation between the input data and their corresponding input annotation ***class*** In terms of the development of this work, the motivation (justification) for choosing a supervised learning approach was discussed in *Section 1 5 2 5* That is, it is proposed that the training-corpus content be exploited in a supervised manner towards the generation of a learned SUE-shot model, and the effectiveness of that model be then evaluated on the test-corpus content

6.1.4 Supervised Learning

The area of supervised machine learning may itself be divided into three subsections In ***binary classification***, the required classification output is a binary decision, i e a test data point is deemed either positive or negative In ***multi-class classification*** it is required that the test data be pigeonholed into a predefined finite number of categories Finally, in ***regression***, the input data annotations are real valued numbers as opposed to a categorical class, whereby the LM aims to learn the correlation between these and their associated input data Correspondingly, the classification outputs are real valued numbers also, representing a prediction based on the input provided It is clear that in terms of obtaining a supervised machine-learned solution, the task of this thesis is concerned with the former aspect, i e the binary classification of FSV shots into SUE or non-SUE categories

6.1.5 Machine-Learning Terminology

6 1 5 1 The Target & Decision Functions

In the supervised learning scenario as described above, the input data/class couplets epitomize a functional relationship The basic function upon which this relationship is based represents what a LM aims to learn by example, and is hence termed the ***target function*** By examining a set of training data, the LM generates an approximation to the target function called the ***decision function***, which is typically selected from a prescribed set of nominated functions called ***hypotheses***

6.1.5.2 Capacity, Consistency, Generalisation & Overfitting

The concept of LM ***capacity*** is defined as its capability of learning the target function of any given training set without error. Ostensibly, the capacity of a LM relates to a quantification of its generic adaptability. For example, an infinite capacity LM should exhibit the ability to learn the relationship between any set of input data/class couplets, irrespective of how they are labeled. Furthermore, if the learned decision function exactly matches the target function of a given a data set, it is said to be ***consistent***.

However, it should be noted that high LM capacity is not necessarily conducive to good classification performance, and it is actually not uncommon for the converse to be true. In fact, the overall performance of a LM is more effectively gauged by its ***generalization*** performance, which is the ability of its learned decision function to accurately classify data points that were not observed in the training set. For example, a LM may exhibit the ability to accurately learn every intricacy of the data points in a given training set, but then make very inaccurate decisions on those of an unrelated set. Such a LM is said to exhibit an unsatisfactory generalization performance, in that it essentially learns 'by rote' the idiosyncrasies of the training set, and then gets confused when confronted with unseen data. This is known as ***overfitting*** and it arises due to the fact that in order to be consistent with the training set, the decision function becomes overly complex. Undoubtedly, if a low LM capacity is maintained the problem of overfitting should not be significant. However, this creates a *catch-22* scenario since a low capacity LM might tend to disregard many of the critical details of the target function. Therefore, selecting the decision function with suitable capacity is a sensitive trade-off.

6.1.5.3 Risk Of Error

As explained above, the aim of a LM is to discern the target function. Given a training set and a learned decision function, the ***empirical risk of error*** corresponds to the number of training set points that would be classified incorrectly by the decision function when applied. However, as explained, the real challenge is to select the hypothesis that maximally reduces the risk of error in the classification of a test point, which corresponds to optimization of its overall generalization performance. Clearly this actual risk of error cannot be determined since it requires knowledge of the unknown probability distribution from which the data are drawn. Nonetheless, recently there have been significant developments in ***structural risk minimization*** (***SRM***) theory [85],

which is a methodology that aims to control the capacity of a learning machine at the same time as minimizing the empirical risk

6 1.6. Approaches to Supervised Machine Learning

Given a data set, the objective of a LM is to be able to correctly categorize the examples into their appropriate classes based on the characteristics of their respective input data This task is known as ***pattern classification*** and in the case of supervised learning, is based on determining the best decision hypothesis **h**, from some hypothesis space **H**, given the observed training data, **D** That is, we are interested in the probability that **h** holds given **D**, i e **P(h | D)** This is called the ***posterior probability*** of **h**, because it reflects the confidence that **h** holds after we have seen the training data **D** There exists several different approaches to evaluating **P(h | D)** and these may be broadly divided into two main types, i e generative and discriminative modelling

6 1 6 1 Generative Modeling

In the ***generative*** approach to pattern classification, the classes are described by modeling their structure, i e their generative statistical model [86] That is, the underlying class behaviours are expressed as random stochastic processes [87], and from these models, the posterior distribution of the labels is derived or estimated via Bayes' formula Specifically, **P(h)** is known as the ***prior probability*** of **h** and denotes the initial probability that **h** holds, before we have observed the training data **P(D)** denotes the prior probability that training data **D** will be observed, i e the probability of **D** given no knowledge about which hypothesis holds Thus, **P(D | h)** denotes the probability of observing data **D** given some world in which hypothesis **h** holds In modeling such attributes, generative approaches generate estimates of posterior probability **P(h | D)** via invoking the Bayes' rule, as shown in **(6 1)**, in which **P(h | D)** increases with **P(h)** and with **P(D | h)** Bayes' theorem is the corner stone of generative methods because it provides a way to estimate the posterior probability **P(h | D)** from the prior probability **P(h)**, together with **P(D)** and **P(D | h)**

$$P(h \mid D) = \frac{P(D \mid h)\, P(h)}{P(D)} \qquad \textbf{(6 1)}$$

Overall, generative models allow for measures of uncertainty, ambiguity, and therefore generalizations [87] In addition, they tend to be efficient in handling large amounts of

data, and are hence most conducive to modeling time-series data [88] Popular schemes include Naive Bayes, Gaussian (Mixtures), Hidden Markov Models, Bayesian Networks, etc

6 1 6 2 Discriminative Models

Algorithms that model the posterior probability **P(h | D)** directly, or alternatively learn the mapping from inputs to the class labels towards generating a confidence score (i e **g(h | D)**), are known as ***discriminative*** models That is, in contrast to the generative approach, discriminative schemes make no attempt to model the underlying distributions (class densities) [86] Instead they are only interested in optimizing a mapping from inputs to outputs Therefore, in realizing pattern classification objectives, all modeling and computational resources are exclusively focused on directly estimating this decision rule (boundary), and hence typically provide superior performance in doing so Common discriminative approaches include K- Nearest Neighbour, Support Vector Machines, Neural Networks, etc While these schemes are anatomically diverse, they exhibit a common characteristic in that, towards finding the exact decision hypothesis that minimizes classification errors on the training data, each aims to predict the class label directly based on the feature representation [89]

6 1 6 3 Generative Vs Discriminative

The relative advantages and disadvantages of the two supervised approaches has been a recurring source of debate in the field of machine-learning to date, resulting in a variety of studies on the subject being published in the literature For example, in [89] it is argued that if the training data is sparse, a generative approach is most appropriate, since using a discriminative scheme in this scenario may lead to overfitting problems Correspondingly in [87] the author claims that generative schemes are most applicable when there is a lot of uncertainty and there is not enough data to train against Furthermore, in [90] it is maintained that discriminative schemes lack the elegance of generative models, are troublesome since they require hands-on tweaking (e g penalty functions, regularization, and kernel functions), and that the relationship between variables are not explicit or visualizable, i e they are 'black-boxes' However, in both [86] and [89] the authors assert that the generative approach to modeling the subject classes is usually an unnecessarily more difficult problem than solving the classification problem directly Moreover, in [89] it is claimed that discriminative classifiers tend to be

generally more effective, since they directly optimize the classification accuracy, and thus exhibit precision superior to that of generative schemes Furthermore, in [86] it is stated that discriminative schemes tend to be more robust than generative models since less assumptions about the classes are made, and significantly, in [91] it is shown using empirical evidence, that discriminative models tend to exhibit lower asymptotic error as the training set size is increased One of the most comprehensive discourses on the debate is provided by Nallapati in [92], where it is proposed that discriminative models tend to be sensitive to noise in the training examples, whereas generative models are relatively impervious to data-noise and require very little training However, it is also argued that unlike many generative models, discriminative models typically make very few assumptions and, in a sense, let the data speak for itself, and this represents the primary motivation for why discriminative schemes have been preferred over traditional generative models in many machine-learning problems in the recent past

Overall, it seems to have been widely accepted that each of the two distinct approaches possess inherent qualities that tend to render them more effective in certain scenarios However, in particular, the exceptional classification performance of modern discriminative schemes has been emphasized by most contemporary studies, e g [86], [89], [91], [92], [93] On this basis, and further justified by the abundance of training data available in this context, it is proposed that this superior accuracy be exploited in applying a discriminatory-based machine-learning approach for the task of binary SFV classification However, as alluded to in [89], any discriminative-based scheme is wholly sensitive to the particular choice of features, and can only be as effective as the discriminatory performance of such So far in this analysis the features that constitute the SFVs have been aggregated heuristically, based on hypotheses inferred from training-corpus observations Therefore, to fully justify their deployment as part of a discriminatory-based classification approach, it is first necessary to explicitly evaluate their intrinsic SUE-shot discernment potential To this end, an exploration of the SFV space follows

6.2. Shot Feature Vector Space Analysis

Recall that each shot is tagged with its own five-dimensional SFV, which exhibits the form shown in **(6 2)**, where **c** (the example class) is a positive/negative flag indicating whether or not the referenced shot is an SUE-shot

$$[c]\ [VCC_1, VCC_2, VCC_3, VCC_4, VCC_5] \qquad \textbf{(6 2)}$$

As explained in *Section 5 4*, the individual vector component coefficients (VCCs) are characterized as follows VCC_1, VCC_2, and VCC_3 correspond to the maximum intensity of reaction-phase seek window (RPSW) crowd image confidence, speech-band audio level, and scoreboard suppression confidence, respectively VCC_4 quantifies the extent of post-shot near-field visual activity, and VCC_5 represents the confidence that the shot culminates in activity located in the field end-zone To gauge the overall SUE discrimination potential of this model, it is desirable to examine the relative resultant vector positions in the SFV-space for both SUE-shot (positive) and non-SUE-shot (negative) examples In this space, a first-rate discriminatory performance should result in a well-defined clustering of the positive and negative points into two distinct groups However, the SFV-space is of dimension 5, and therefore, without resorting to some form of Principal Component Analysis, is not easily conducive to illustration Nonetheless, the scheme is formulated in anticipation of the two data classes being separable on the basis of absolute VCC intensity, i e positive class SFVs should generally exhibit higher valued VCC values than those of the negative class Therefore, it is anticipated that the overall discriminatory potential of the SFV model may be sufficiently inferred from the trends exhibited by the individual VCC component values The following sections explore this concept

6 2 1. 1-D Vector Component Coefficient Exploration

SFVs were extracted for each shot of the multi-genre training-corpus, where each extracted SFV instance is known as a ***training point*** (**TP**) As outlined in Table 1 3, across all genres, the training-corpus consists of 883 SUEs, the locations of which were manually annotated Given this, the SFVs of the SUE-shots (i e positive TPs) were labeled as class **+1**, while the remainder (i e negative TPs) were labeled as class **−1** For example, **(6 3)** presents a positive training point (**PTP**) and a negative TP (**NTP**) as extracted from the training corpus

$$\begin{aligned} &PTP \quad [+1]\ [0\ 138989, 0\ 512867, 0\ 995904, 0\ 133215, 0\ 942561] \\ &NTP \quad [-1]\ [0\ 073177, 0\ 933813, 0\ 495906, 0\ 898297, 0\ 556816] \end{aligned} \qquad \textbf{(6 3)}$$

Note that in this case most of the individual VCC values of the PTP outweigh those of the NTP, which, as described above, represents the basis for the anticipated 5-D SFV

separability of the two classes. However, as mentioned, it is not trivial to illustrate this, and therefore the discriminatory trends of the constituent VCC values are illustrated individually, and their usefulness postulated on that basis.

VCC_1 values were extracted from the SFVs of all 883 SUE-shots (PTPs) of the training-corpus. For comparison, these values were also extracted from the SFVs of 883 randomly chosen training-corpus NTPs, and both sets are plotted in **Fig. 6.1**. From this plot it is evident that, as anticipated, the two classes are inseparable on the basis of this feature alone. Recalling that VCC_1 corresponds to the maximum RP crowd image confidence, this is to be expected since (i) it is not every SUE that exhibits a crowd sequence in its subsequent RP, and (ii) every crowd sequence instance is not always preceded by a SUE-shot. However, it was previously shown that in many cases this premise does in fact hold, and this is reflected in the general PTP Vs NTP VCC1 trend in the figure. Specifically, it is evident from the plot that the PTPs exhibit a definite value bias in terms of VCC_1, compared to that of the NTPs, i.e. the majority of the PTPs tend to exhibit higher values than that of the NTPs, and vice-versa. Thus, while not solely providing for a clear-cut discrimination, the broad PTP/NTP trend divergence of this vector component should contribute significantly to the separation to be provided by the overall SFV.

Similarly, PTP Vs NTP plots for **VCC_2**, VCC3, VCC4, and VCC5, are presented in **Figs. 6.2**, **6.3**, **6.4**, and **6.5**, respectively. As in the previous example, the two classes are inseparable on the basis of the individual features alone, however, it is

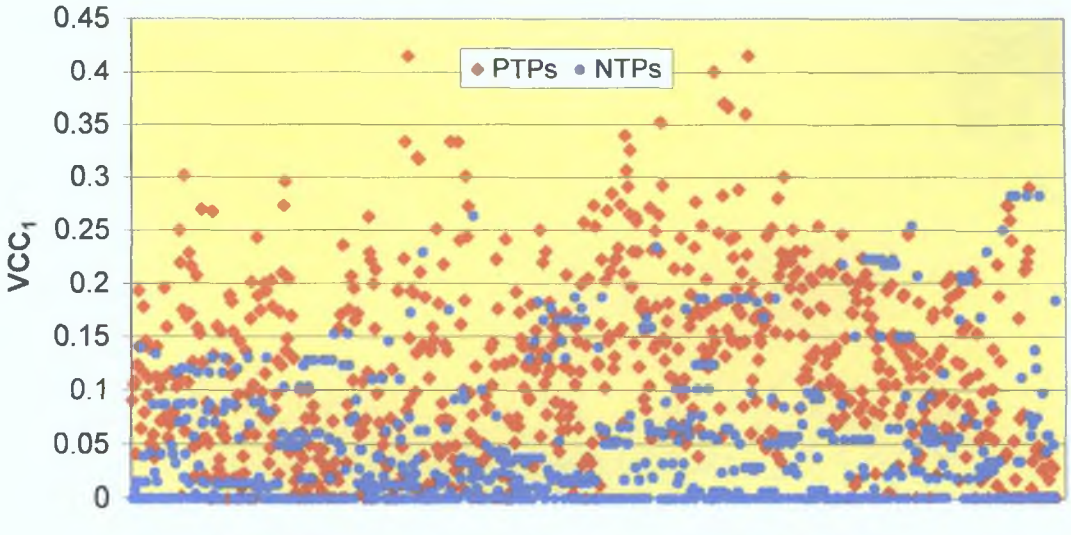


Fig. 6.1. VCC_1 values for training-corpus PTPs and NTPs.

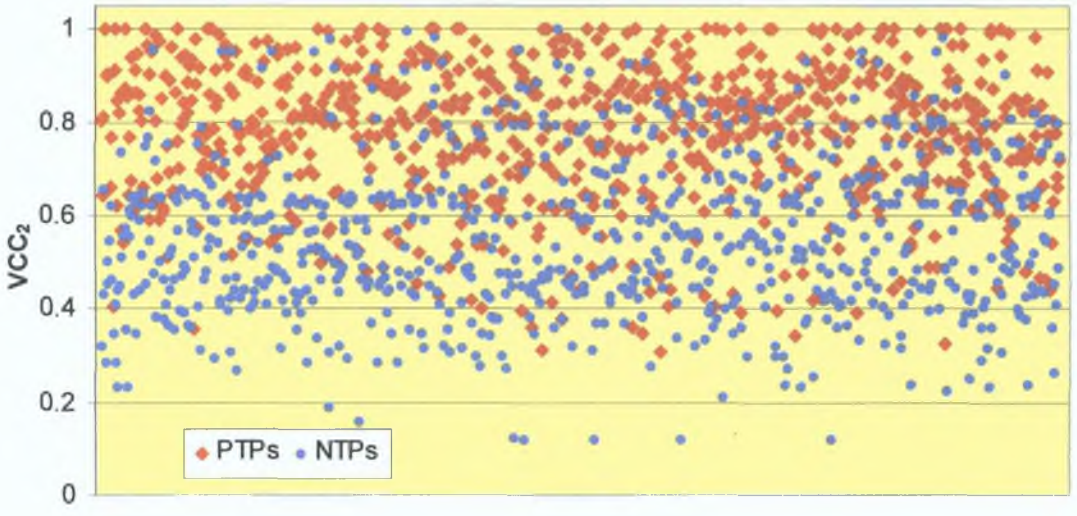


Fig. 6.2. VCC_2 values for training-corpus PTPs and NTPs.

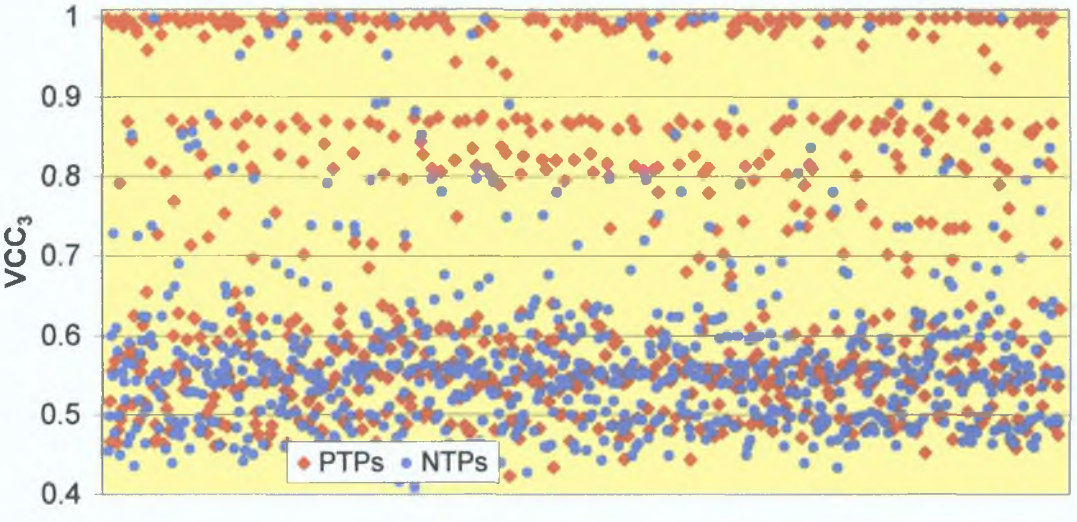


Fig. 6.3. VCC_3 values for training-corpus PTPs and NTPs.

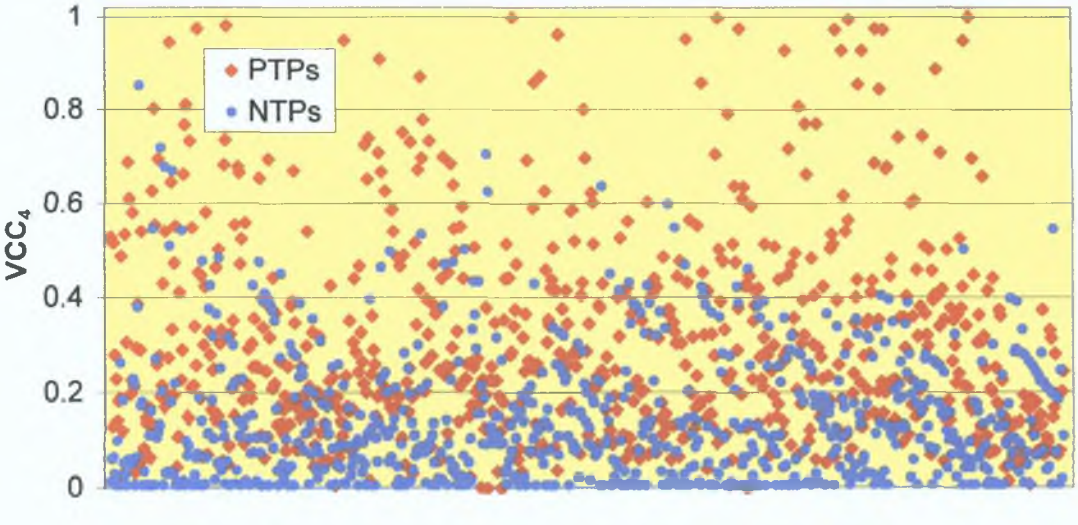


Fig. 6.4. VCC_4 values for training-corpus PTPs and NTPs.

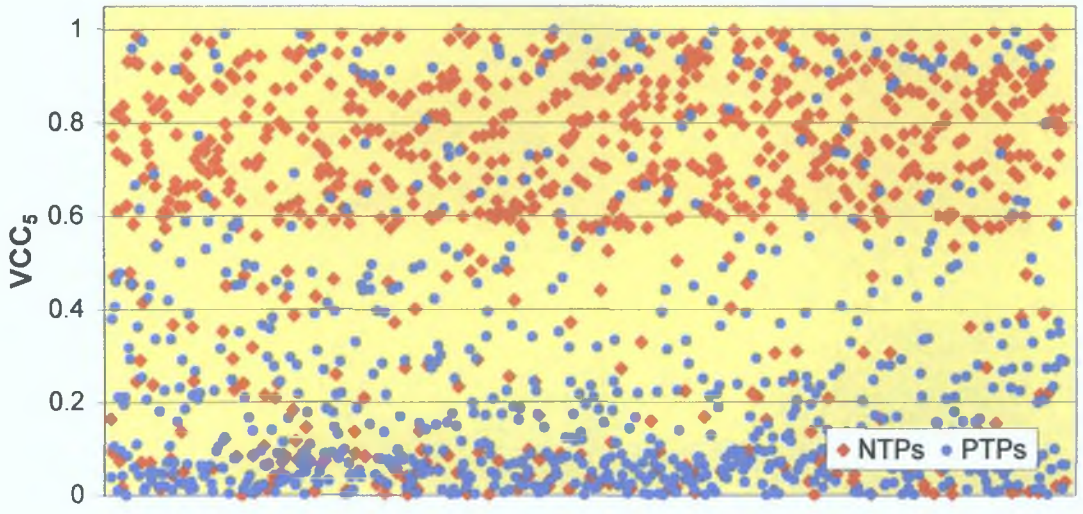


Fig. 6.5. VCC_5 values for training-corpus PTPs and NTPs.

again evident that the general trend of the values is that of an intensity bias towards the PTPs. That is, in each case (and in some more profusely than in others), more PTPs occupy the higher values than NTPs, and likewise, more NTPs occupy the lower values than PTPs. Hence, over the training-corpus data, the contribution of each individual VCC is shown to be constructive in the discrimination of PTPs and NTPs (i.e. in the detection of SUEs in FSV). It is important to acknowledge this prior to any further system development, such that in optimizing the model to yield the best results possible, any individual component that is shown to contribute destructively in the training phase may be either rectified or removed entirely from the system.

It is evident from these five figures that there exists a variance in PTP/NTP discrimination strength across the five individual features. Via a crude cross-comparison of the plots, it was observed that relatively strong PTP/NTP discrimination is provided by $\mathbf{VCC_1}$, $\mathbf{VCC_2}$, and $\mathbf{VCC_5}$, while slightly weaker (but nonetheless valuable) level of discrimination is given by $\mathbf{VCC_3}$ and $\mathbf{VCC_4}$.

6.2.2. 2-D Vector Component Coefficient Exploration

While it has been shown that on an individual basis the VCCs exhibit discriminatory trends in relation to PTPs and NTPs, in each case the two classes remain inseparable. Thus, to further bolster the justification of the SFV model, it is desirable to determine the extent of the improvement in PTP/NTP separability (if any) by combining the

VCCs in pairs in 2-D space. To this end, **Fig. 6.6** presents a plot of $\mathbf{VCC_1}$ against $\mathbf{VCC_3}$ for the same extracted PTPs and NTPs used previously. From this plot it is evident that, while still not wholly distinguishable, there is significant improvement in the separability of the two classes compared to that yielded by either of the components acting alone (c.f. Figs. 6.1 and 6.3). For purely illustrative purposes, a crude separating function is shown in the figure. Although, such a separator would be unsatisfactory in a practical scenario, it serves to demonstrate the improvement in the separability of the data.

Similarly, **Fig. 6.7** presents a plot of $\mathbf{VCC_2}$ against $\mathbf{VCC_4}$. From this plot it is likewise evident that the PTPs and NTPs are not fully separable. However, once again there is a clear substantial improvement in their differentiation compared to that generated by the either of the components alone (c.f. Figs. 6.2 and 6.4). Again, purely for example, a crude separating function is estimated as shown in the figure.

Hence, it has been shown by example that by combining the VCCs in 2-D pairs the overall separability of between PTPs and NTPs is improved. Therefore, given the improvement from 1-D to 2-D, it is anticipated that in 5-D SFV-space, i.e. on the basis of $\mathbf{VCC_1}$, $\mathbf{VCC_2}$, $\mathbf{VCC_3}$, $\mathbf{VCC_4}$, and $\mathbf{VCC_5}$ combined, the positive and negative examples should be largely separable. As alluded to above, given this, it is proposed that a discriminative-based classifier be employed to implement the separation (classification) task. However, it remains to be investigated which discriminative approach should be used such that the optimum performance may be attained.

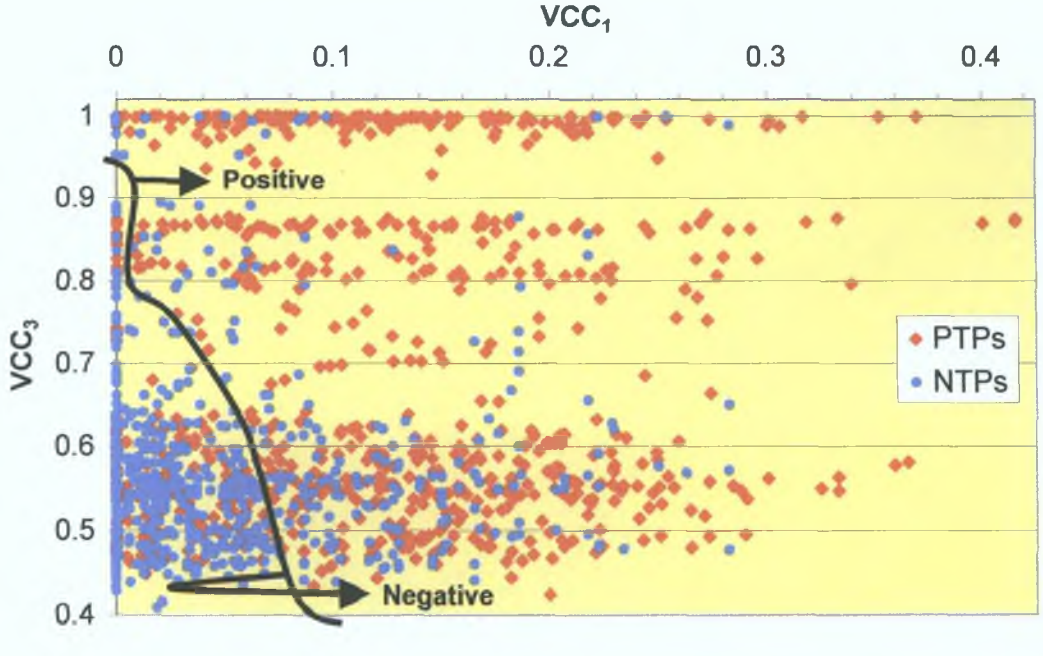


Fig. 6.6. VCC_1 Vs VCC_3 for training-corpus PTPs and NTPs.

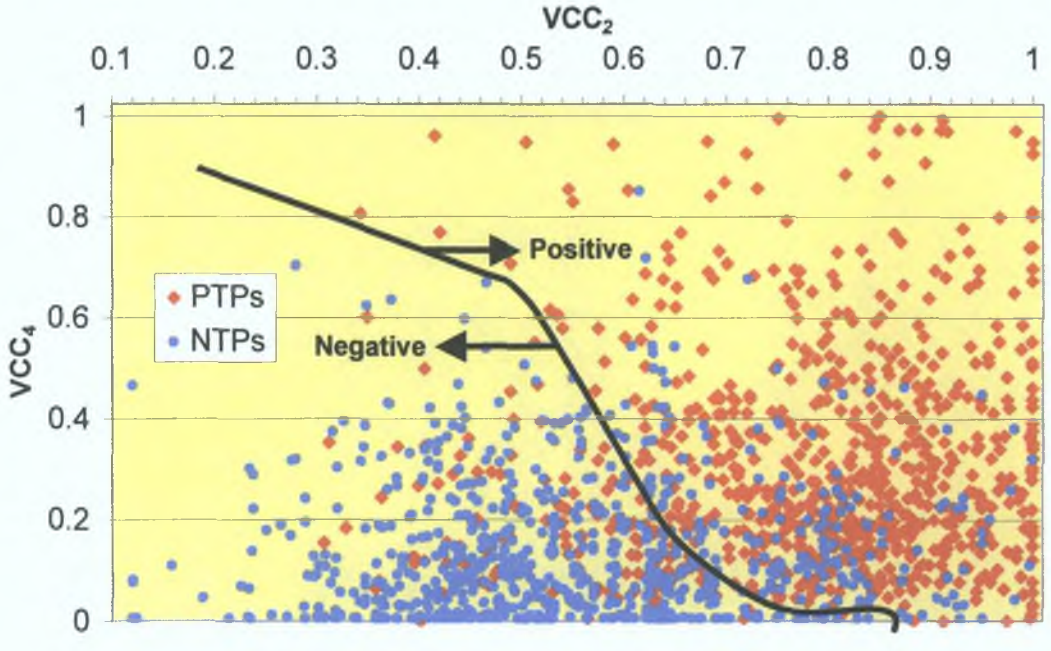


Fig. 6.7. VCC_2 Vs VCC_4 for training-corpus PTPs and NTPs.

6.3. Discriminative Pattern Classifiers

As described earlier, given an annotated data set, discriminative schemes approach the problem of pattern classification directly, in attempting to optimize the mapping from input data to output class. The following sections briefly outline three of the most common and effective approaches (i.e. K- Nearest Neighbour, Support Vector Machines, and Neural Networks) [90], and outline how they compare against each other.

6.3.1. K- Nearest Neighbour

As described in [89], K- Nearest Neighbour (KNN) is an example of a retrieval based classifier, which is rooted in the straightforward application of basic well-established similarity techniques. In the KNN scheme all training examples are stored, then given a test example, the technique finds the k examples (i.e. neighbour examples) that are most similar (via some metric, e.g. dot product) to the test point. The class that is most common to these neighbour examples is then assigned to the test query, i.e. the neighbours vote for the class. The scheme can be improved by considering the relative distance of the neighbours, e.g. a closer neighbour has more influence on the outcome.

Critical elements of this scheme are the feature representations, and the distance metric used in the similarity check

6.3.2. Neural Networks

The development of Neural Network (NN) classifiers is rooted in the exploitation of the new neurological discoveries of the 20th century In employing NNs for data classification, it is assumed that the target function to be discerned is a non-linear function that can be represented by a layered system of interconnected nodes mapping input data values to output classes, i e a neural network infrastructure By training the NN on a known training set the decision function hypothesis is then found by varying the weights governing each node connection until a specific error metric (i e the empirical risk) is minimized A comprehensive discourse on NN technology for supervised pattern classification may be found in [94]

6 3 3 Support Vector Machines

As described in [93], Support Vector Machines (SVMs) are essentially binary classifiers representing a relatively new approach to pattern classification developed from the theory of structural risk minimization [95], which was mentioned in *Section 6 1 5 3* above Basically, SVMs assume the target function is a non-linear function that can be represented by a linear classifier supplemented by a kernel function The decision function in SVMs is given by the hyper-plane that separates the two classes of training examples with the largest margin [96], and is found by minimizing the classification errors on the training examples It is expected that the larger the margin, the better the generalization performance of the classifier The hyper-plane is in a higher dimensional space called kernel space and is mapped from the feature space The mapping is done through kernel functions that facilitate operation in the input feature-space while providing the ability to compute inner products in the kernel space The key idea in mapping to a higher space is that, in a sufficiently high dimension, data from two categories can always be separated by a hyper-plane [92] A comprehensive discourse on SVMs (and other kernel-based learning methods) may be found via [97]

6 3.4. Comparison Of Discriminative Classifiers

The main advantages of the KNN approach are that (i) no training is needed, (ii) the

scheme can be applied to any distance measure and feature representation, and (iii) it is empirically effective However the scheme is disadvantaged by two innate characteristics the high time complexity needed to find the nearest neighbour, and the imprecision when the number of examples is small, which may be often true in high-dimensional feature spaces [89] Moreover, in [98], the authors compare the performances of KNN and SVM in realizing a conventional audio segmentation task Therein, the SVM approach was shown to significantly outperform KNN, both in terms of classification accuracy and computation performance

NNs have traditionally been the most widely used discriminative classification approach, and have been shown to be very effective in a wide range of scenarios [94], provided the structure of the NN is appropriate in each case [89] However, as alluded to in [89], the NN approach suffers from two main problems, i e it is hard to interpret the trained classifier, and there is typically no guidance available on the choice of the NN architecture

Within the more contemporary literature on the subject, arguments for the superiority of SVMs seem to be in abundance For example, in [89] it is stated that SVMs are the most theoretically well founded of all classifiers and guarantee a certain amount of generalization ability Furthermore, in [98] it is claimed that once trained, the computation in a SVM depends on a usually small number of supporting vectors and is fast In [93] it is stated that SVM has an outstanding ability as a binary splitter and that the classification results are known to yield a better generalization performance compared with other classifiers In [99] it is argued that one of the main advantages of SVMs is that it is most robust to noisy data In [100] it is argued that SVMs have a greater ability to generalize in comparison to other statistical classification methods, and this is qualified by showing how they outperform a variety of other classification methods within a speech recognition context Thus, while SVMs are not without weaknesses (e g the training time tends not scale well with the size of the training data, and an appropriate kernel design is required [89]), it was considered desirable to investigate why they are quickly becoming the most championed of the discriminative classifiers

Above all, the justification for this stems from the SVM formulation, which uniquely embodies the structural risk minimization (SRM) principle [95] As mentioned in *Section 6 1 5 3*, SRM minimises an upper bound on the expected risk, as opposed to merely the empirical risk, via a simultaneous control on capacity In contrast traditional

NNs are solely rooted in the principle of ***empirical risk minimisation*** (**ERM**), which minimises the error on the training set However, SRM has been shown to be superior to ERM [101], and therefore on this basis, it is proposed that SVMs are equipped with a greater ability to generalise [102], which is the ultimate goal in statistical learning That is, compared to ERM based NNs, it is argued that SRM driven SVMs tend to yield a better learned decision function, with less overfitting, which approximates the target function more closely [103]

Furthermore in [103] it is proposed that because NNs use gradient descent search, they can sometimes converge to local minima In other words, the classification model that a NN finds might not be the best classifier In contrast, due to their sound mathematical formulation, SVMs always achieve the global solution [103] In addition, it is proposed in [103] that the NN learning process requires training with the data set repeatedly over many times to better learn the hypothesis function, i e the more times they get trained, the better they learn Thus it tends to take more time to have a good NN working model than an SVM equivalent, and there is no precise way to tell how much training is required [103]

However, both SVMs and NNs have their own drawbacks For example, both suffer a decline in performance as the dimension and the quantity of the data inputs increase Also, whereas NNs rely heavily on the structure of the networks, i e the choice of an appropriate number of hidden layers, the success of a SVM depends on how well the chosen kernel functions work to create a non-linear boundary in the input space for separating data Nonetheless, overall it seems that while the debate is not quite concluded (e g there is a significant result showing, both in theory and in practice, that NNs still work better in regression learning tasks [103]), given the above arguments, it is not difficult to comprehend why SVMs are fast becoming the more relied upon scheme for many data discrimination tasks

Finally, but significantly, in [83] the author reports on a specific investigation into quantifying the effectiveness of SVMs in video segmentation applications Therein, it is not only concluded that such are applicable to video-based classification scenarios, but are shown to yield excellent accuracy in the tasks realized

Thus, based on the latter evidence and that of all the aforementioned arguments, it is proposed to employ an SVM-based approach in realising the positive/negative SFV discrimination task A more comprehensive introduction to

SVMs may be found in *Appendix D*, while a description of the actual SVM implementation used (i e **SVM**^{light}) may be found in *Appendix E*

6.4. Chapter Summary

In this chapter, the topic of pattern classification was introduced in relation to the task of analysing and classifying the shot feature vectors (SFVs) Given the motivation for a machine learning approach, a description of the two main areas of this subject was presented, i e supervised and unsupervised learning, where the classification task of this thesis is concerned only with the former Following an analysis of the arguments advocating the various approaches to supervised machine-learning, it was proposed that a discriminative-based approach be employed To further justify this, an exploration of the training data shot-feature vector components was performed, in which it was shown that each component contributes constructively in discriminating between the positive and negative data points On this basis, it was postulated that the quasi-separability observed for low-dimensional SFV component combinations should be consistent and improve as all 5 components are combined in true SFV space Given this, the three most commonly advocated discriminative classifiers were discussed, and for the reasons outlined, a Support Vector Machine (SVM) implementation was favoured

Chapter 7

Experiments & Summarization Performance

In *Chapter 4* it was proposed how a generic solution for the summarization of field-sports-video (FSV) may be realized based on the detection of critical features indicating the score-update episodes (SUEs) In *Chapter 5* it was described how this proposed hypothesis was implemented The motivation for employing a Support Vector Machine (SVM) approach to realize the pattern learning/classification processes of the scheme was then outlined in *Chapter 6*, with an introduction to the specific SVM implementation being provided in *Appendix E* Given the hypothesis, the implementation, and the proposed classification approach, this chapter describes the details of the training and testing phases of the experiments performed, followed by a comprehensive discussion and evaluation of the results obtained in terms of the summarization task

7.1. Training-Phase

7.1 1. Training Data

The shot-boundary detection algorithm [79] was executed on the entire training-corpus From this, 68508 shot transitions were detected The corresponding shot feature vectors (SFVs) for all 68509 shots were then extracted exactly as outlined previously in *Section 5 4* These 68509 SFV training-points constitute the SVM training-phase input As explained, the SFVs associated with each SUE-shot correspond to positive training points (PTPs), while those of the remaining shots constitute negative training points (NTPs) Given the set of correctly labeled training data, the SVM attempts to learn the

correlation between input feature data and the corresponding binary classes It is described in *Appendix D* how an SVM will learn the hypothesis that should yield the optimum generalization performance, i e the hypothesis that should produce the best results in classifying the test points of an unseen test-corpus

7.1 2. Outlier Filtering

Recall, the SFV training points have the form shown in **(7 1),** whereby the relative intensities of the vector component coefficients (VCCs) provide the overall probability of whether a given point is of positive or negative class (c), i e whether or not its associated shot exhibits a high probability of being an SUE-shot

$$[c]\ [VCC_1, VCC_2, VCC_3, VCC_4, VCC_5] \qquad \textbf{(7 1)}$$

From above the training dataset consists of 68509 examples Recall that this is comprised of 883 PTPs, and thus 67626 (= 68509-883) NTPs However, it is not unfeasible for outliers to occur in the training data, i e points that exhibit an inconsistency between the SFV class and the bias implied by the feature data Retaining these inconsistent examples within the input training data would tend to have an adverse effect on the learning performance of SVM [83] Hence it is desirable to have them removed in advance While data outliers are not always easily discerned (hence the reason for the machine learning approach in the first place!), many may be obvious, and both the SVM optimisation and the resultant SVM performance should benefit from the removal of these

Given that the SFV class probability is rooted in relative VCC intensity, it is proposed that the attribute of SFV magnitude should provide a reliable basis for outlier identification To investigate this, SFV magnitudes for all 883 PTPs, as well as those for a corresponding illustrative subset of the 67626 NTPs, are plotted in **Fig 7 1** While the two classes are clearly inseparable in this representation, it is evident that several PTPs/NTPs exhibit SFV magnitudes that tend to belie their known class That is, while most of the PTPs exhibit high valued SFV magnitudes, some are found to exhibit low values that fall within the range generally occupied by the NTPs Likewise, some NTPs fall within the range generally occupied by the PTPs Such examples may be considered as inconsistent data outliers within the training set However, given this compromised representation it is considered tolerable to remove only the most conspicuous of these, since it is imperative that the decision surface characteristics in the true SFV training

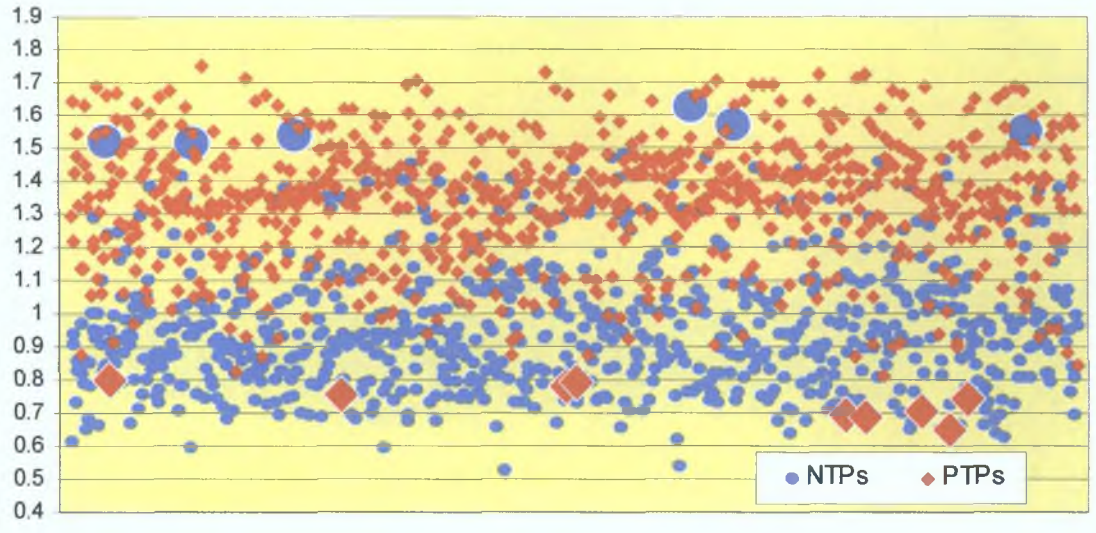


Fig. 7.1. Plot of training-corpus PTP/NTP shot feature vector magnitudes.

data representation remain unaltered. Hence, on the basis of the orientations observed, it is proposed that PTPs with SFV magnitudes less than 0.8 are removed from the training set, as well as NTPs with SFV magnitudes greater than 1.5 – such points are indicated in the figure by enlarged markers. In applying this filter, the PTPs of the training set are reduced in number from 883 to 874, while the NTPs are reduced from 67626 to 66987.

Training data outlier removal operations tend to be a common practice in many discriminative pattern classification schemes, e.g. [104]-[107]. In accordance with the results presented in these works, it is anticipated that following this operation the SVM should be able to better estimate the intrinsic target function of the training data. An additional benefit of the outlier removal procedure is that the number of training examples is reduced, which in turn should reduce the computation (and hence time) required for training.

7.1.3. SVM Cost-Factor

Following the outlier filtering process the number of data points in the training set is reduced from 68509 to 67861, i.e. 874 PTPs plus 66987 NTPs. Hence, even following the outlier removal process, a large imbalance remains in the numbers of retained PTPs/NTPs. This is problematic, since when faced with disproportionate datasets, the performance of SVM drops significantly [108]. That is, deterioration occurs when the

magnitude of the noise in the dominant class outweighs the total number of the minor class examples When this phenomenon arises, the minor class examples may be indiscernible from the noisy dominant class examples, and therefore the optimal hyperplane determined by the SVM (see *Appendix D*) will typically classify all members of the training set as dominant class examples [109] A popular approach towards solving this problem is to bias the classifier so that it pays more attention to the minor examples For SVMs, this can be achieved by increasing the error penalty **C** associated with misclassifying the minor class, relative to that of the dominant class In many SVM implementations, including the chosen implementation (SVM^{light}), this is achieved by setting a user-defined parameter known as the ***cost-factor***, **j**, which dictates the extent to which training errors in positive examples should outweigh those of negative examples, i e it allows adjustment of the cost of false positives Vs cost of false negatives Clearly for a perfectly balanced dataset the cost-factor should be set to unity, however given the disproportion evident in the dataset used here, the appropriate cost-factor is defined as the ratio of PTPs to NTPs for the outlier reduced training set, as shown in **(7 2)**

$$j = \frac{\#PTPs}{\#NTPs} = \frac{874}{66987} = 0\,013047 \qquad \textbf{(7 2)}$$

7.1.4. SVM Kernel Function

As outlined in Appendix D, in using SVMs, it is required to specify a ***kernel function*** such that it can handle non-linear data separation However, the SVM formulation does not include criteria to select a kernel function that will yield the best performance [110] Moreover, it is a commonly held argument that there exists no theoretical basis on which such a decision may be made [111] However, the three most well-studied and commonly used SVM kernel functions are described in *Section D 4 1* (*Appendix D*) - corresponding to polynomial, radial basis, and sigmoidal functions Hence, it was proposed that the relative performances of each be compared, and the best performing implementation be chosen on that basis

As explained in *Section 6 1 5 2*, in tandem with training set classification accuracy, the performance of a learning machine is also critically gauged by the ability of its learned decision function to accurately classify data points that are not observable in the training set, i e its generalization performance Recall that exhibiting good

generalisation corresponds to maintaining a low learning capacity, which in turn is directly related to a quantity known as the VC dimension of the machine (see *Appendix D*) That is, it is critical that the VC dimension be controlled (minimized) in addition to keeping the number of training data misclassifications as low as possible As explained in *Section D 1*, while the VC dimension cannot always be calculated, it is generally possible to calculate its upper bound

Given the above, it was proposed that kernel performance comparisons be performed based on the following critical criteria, (i) the number of training set points misclassified (i e the empirical risk of error), and (ii) the estimated upper bound on the VC dimension (i e the capacity/overfitting indicator) To this end, with γ set as calculated above, three distinct SVM classifiers were trained on the outlier-reduced training dataset using each of the abovementioned kernel functions As a point of reference, a linear SVM classifier was trained also Note, in each case it was left to **SVM**^{light} to (i) determine the default error penalty **C**, and (ii) define the default kernel parameters **Table 7 1** presents values for each implementation, representing the number of training set misclassifications (expressed as a percentage of the overall dataset), as well the estimated upper bounds on the VC dimension of each classifier (calculated by **SVM**^{light}) From this data it is evident that the optimal hyperplane found using a radial basis function (RBF) kernel outperformed the others both in terms of offering a lower empirical risk, as well as a lower estimated VC dimension upper bound It was thus concluded that this kernel represents the most favourable mapping for the problem domain herein, and was thus the implementation employed

The formulation of the RBF kernel is that as shown in **(7 3)**, where γ is a user-defined parameter that is specific to this kernel function

$$K(x,y) = \exp(\gamma\|x-y\|^2) \qquad \textbf{(7 3)}$$

Table 7 1 Estimates of training set errors and upper VC dimension bound for four different kernel functions

Kernel Function	% Training Data Misclassified	Upper Bound on VC Dimension
Linear	15%	230
Polynomial	14%	226
Radial Basis	8%	123
Sigmoidal	49%	∞

To determine the optimum value of γ for the problem domain, it was varied across a range of values, and then using the same criterion as above, an optimum value was selected. Specifically, with **j** set as before, a range of RBF driven SVM classifiers were trained on the outlier-reduced dataset, while γ was varied. **Fig 7.2** illustrates how the percentage of training set misclassifications varies with γ. From this data it is evident that a global minimum occurs for $\mathbf{1.2} \leq \gamma \leq \mathbf{1.3}$, yielding error performances that slightly improve upon that generated by the **SVM**^{light} default γ value of 1.0 (illustrated). Also indicated (numerically) in the figure is the variance in the estimated bound on the VC dimension. Given that this aspect was found not to alter significantly with γ, for the forthcoming RBF-based SVM classification experiments, γ was chosen as the median of the abovementioned range yielding minimum training set misclassifications, i.e. $\gamma = \mathbf{1.25}$.

7.1.5. Error Penalty Variance

As explained in *Section D.3.1* (*Appendix D*), the error penalty **C** is a user-defined SVM parameter, which determines the relative significance of training errors compared to the width of the margin in the objective function to be optimized. Consequently, there exists a tangible relationship between the chosen value of **C** and the overall SVM performance. In effect, variation of **C** during the training phase allows the user to tune the classification, such that an increase in **C** should improve precision at the expense of recall (and conversely a decrease in **C** should yield higher recall at the expense of

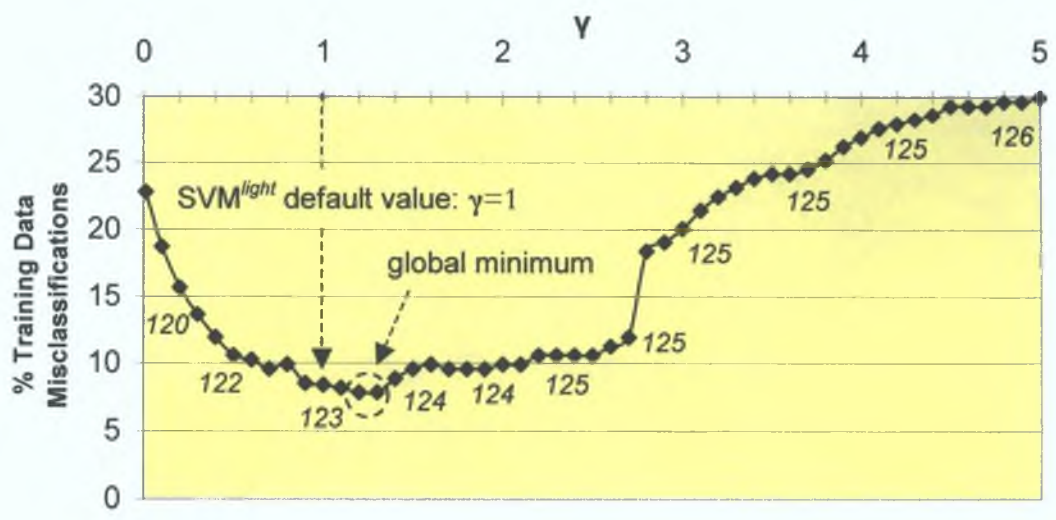


Fig. 7.2. Variance of training data misclassifications (and VC dimension bound) with γ value of RBF kernel Support Vector Machine.

precision) While the **SVM**^{light} algorithm tends to perform well in setting an appropriate default value for **C** given an input data set [112], to provide an indication of the range of possible results obtainable for the scheme, it was proposed that a set of SVM classifiers be trained for a variety of error penalty values To this end, with $\jmath$ and γ set as outlined above, RBF driven SVM classifiers were trained on the outlier-removed dataset, whilst **C** was varied though a critical range Coupled with the test-corpus data, this resultant set of classifiers formed the basis for the testing phase

7.2. Testing Phase

7.2.1. Test Data

Using the set of trained classifiers described above, the testing phase involves the classification of data constituting an unseen corpus As described in Table 1 4, similar to but distinct from the training-corpus, the test-corpus consists of a further 90-hours of field-sport content, encompassing 850 SUEs in total From the audiovisual streams of the test-corpus content the required signal-level data was mined, from this the frame-level CFs were extracted, and subsequent SFVs were then generated as described earlier Once again, the SFVs constitute the SVM input, but in the testing phase the classes are not input, and the SVM is charged with estimating them using its learned hypothesis However, as a preamble to this procedure it is desirable for the test-corpus content to be preprocessed as described in *Section 5 3*, such that some of the groundwork in disregarding non-SUE-containing shots may be achieved prior to SFV pattern analysis

7.2 2. Pre-Processor Filtering

The task of the preprocessing filter is to reduce the probing domain of the SFV pattern classification stage Recall that two distinct procedures implement this task The first of these concerns the detection of advertisement breaks using the algorithm described in [80] Secondly, a shot-filtering process is performed based on the stipulation that for a shot to be retained, it must exhibit an I-frame close-up image instance within its reaction-phase seek window (RPSW) As outlined in *Section 5 3*, this is verified for each shot in question by comparing the close-up confidence (CuC) values for its RPSW I-frames with the empirically determined threshold value, $\mathbf{T_{CuC} = 0\ 08}$

The test-corpus content was preprocessed in this manner Based on a manual

annotation of SUE-shot locations, **Table 7 2** provides resultant values for (i) ***content rejection ratio*** (**CRR**), which denotes the percentage of test-corpus content rejected by the preprocessor, and (ii) ***SUE retention ratio*** (***SUERR***), which indicates the corresponding percentage of SUE-shots retained by this process For increased transparency, these values are broken down across the individual field-sport genres that constitute the corpus From this data it is evident that across all constituent genres, the preprocessor performed effectively in both the rejection of non-SUE content and the retention of the vast majority of SUE-shot incidences For example, consider its performance on the Gaelic football content alone From Table 7 2, following the ad-break detection process, 8 5% of this content is listed for rejection, while 100% of all Gaelic football SUE-shots are retained In parallel, following the close-up shot-filter, 41 7% of the content is listed for rejection, while a corresponding 97 3% of SUE-shots are retained Combining the two processes yields an overall preprocessor performance of 46 6% content rejection for 97 3% SUE-shot retention (i e there exists some overlap between the rejection periods determined by the two independent processes – a trait that is evident across all test-corpus genres) Recalling that the average duration for a Gaelic football broadcast is 91-minutes (see Table 1 2), via the preprocessor alone, on average, over 42-minutes of this content is rejected, while retaining 49-minutes, which includes over 97% of the games' SUEs Some genres exhibit better preprocessor performance than others For instance, in the case of hockey, although 52 3% of content was rejected, only 93 5% of its SUEs were retained, which significantly differs from the 100% ideal However, taken as a whole, it was determined that, on average, 48 3% of all test-corpus content was rejected, while 95 9% of all included SUE-shots were retained These statistics suggest that, albeit for a nominal penalty in SUERR, the preprocessor performs well in the rejection of non-SUE content

Table 7 2 Percentage ratios for content rejection and SUE retention following the preprocessing of test corpus content

	Ad-break Removal		Close-Up Filter		Combined	
GENRE	**CRR**	**SUERR**	**CRR**	**SUERR**	**CRR**	**SUERR**
Soccer	4 4%	100%	55 3%	95 3%	56 0%	95 3%
Gaelic Football	8 5%	100%	41 7%	97 3%	46 6%	97 3%
Rugby	5 2%	100%	36 2%	97 1%	38 0%	97 1%
Hurling	8 7%	100%	42 0%	96 4%	48 7%	96 4%
Hockey	5 8%	100%	50 6%	93 5%	52 3%	93 5%
				Average	**48 3%**	**95 9%**

7.2.3 Shot Classification

Given that 48 3% of the test-corpus content was rejected at the preprocessor filtering stage, it was thus required to detect the retained 95 9% of SUE-shots amongst the remaining 51 7% of content representing the probing domain of the SFV pattern analysis phase As described above, during the training phase, multiple SVM classifiers were generated by varying the error penalty, **C**, through a critical range Using this set of classifiers, each learned hypothesis was executed on the SFV data of the retained content In doing so, each individual shot was assigned a decision class, i e positive or negative, based on its corresponding SFV characteristics As before, a positive decision class indicates that, on the basis of the SFV attributes and the decision function in operation, the given shot is likely to be an SUE-shot The following section describes and evaluates the effectiveness of this process in generating summarized output

7.3. Summarization Performance

Following the execution of the abovementioned procedures, for each trained SVM (from the set generated by varying **C**), the test-corpus content was processed as described By comparing the positive shot classification decisions with those of a manually generated test-corpus annotation (ground truth), and by determining the ideal levels of content rejection, the summarization performances obtained for each SVM instance were determined As described, the set of classifiers were learned from the training-corpus data taken as a whole, and then applied to the test-corpus content as a whole However, for increased transparency, the results of the test-corpus summarization performances are broken down across its individual constituent sports genres

7.3.1. Rugby-Video

Described in Table 1 4, the rugby-video portion of the test-corpus encompasses 167 SUEs It was manually determined that the corresponding 167 SUE-shots constitute 3 7% of the total rugby-video test content Therefore, the ideal summarization performance for this particular genre corresponds to the retrieval of all 167 of these true-positive test-points, coupled with the rejection of the remaining 96 3% of the content To quantify the performance of the scheme in realizing this task, CRR (a true-

negative/false-positive performance statistic) and SUERR (a true-positive/false-negative performance statistic) statistics were computed as before. Moreover, both were estimated for the decisions made by each trained SVM classifier as **C** was varied through its prescribed range. **Fig. 7.3** presents a combined plot of CRR/SUERR against **C**, where also shown are the ideal summary performance values of such, i.e. 96.3% CRR and 100% SUERR. Recall that the levels of CRR/SUERR following the preprocessing phase are 38.0% and 97.1%, respectively (see Table 7.2). These preprocessor values form the point of entry in the graph (on the y-axis) and are indicated with the symbol 'X'. Beginning with C=0.02, once the corresponding SVM was applied the SUERR level dropped only very slightly from its preprocessor level, while the CRR level immediately increased by about 8%. Subsequently, as **C** was incrementally increased from this point (i.e. for each corresponding learned SVM), the resulting classification performance varied according to the graphs as shown. That is, following an initial period of stability, the CRR level progressively increased from its preprocessor level towards its ideal level, while simultaneously the SUERR value gradually decreased away from its own entry level (and thus diverged from its own ideal level). Ultimately, above and beyond C≈1.15 the two CRR/SUERR statistics saturate at approximate levels of 77% and 67% respectively. Hence, in terms of summarization, the general trend observed was that as the value of **C** was increased, the level of content rejection increased but at the expense of event retrieval.

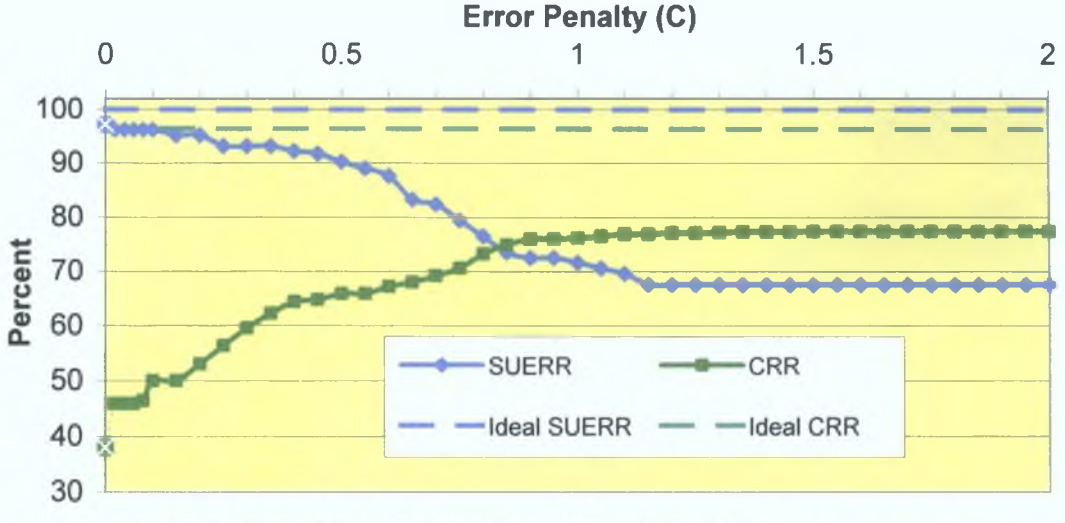


Fig. 7.3. Plot of SUERR/CRR Vs C for rugby-video test content.

7.3.2. Soccer-Video

Recall from Table 1.4, that the soccer-video portion of the test-corpus encompasses 56 SUEs. Similarly to above, it was manually determined that the corresponding 56 SUE-shots represent 1.7% of the total soccer-video test content. Hence, in this case the ideal summarization performance corresponds to 100% SUERR coupled with 98.3% CRR. Again, to quantify the performance of the scheme in realizing this, CRR/SUERR statistics were estimated for the decisions made by each SVM classifier as **C** was varied through its prescribed range. As in the previous case, **Fig. 7.4** presents a combined plot of CRR/SUERR against **C**, where again shown are the ideal summary performance values. In this case the CRR/SUERR preprocessing levels are at 56.0% and 95.3% respectively (Table 7.2). Following the application of the initial (C=0.02) SVM, the CRR level immediately increased by about 4%, while the SUERR level is maintained at the at the preprocessor value. Again, as **C** is increased from this point the two statistics vary as shown. That is, similar to the rugby-video scenario, following short periods of stability, the CRR is increased towards its ideal level, while the SUERR diverges away from its ideal level. In this case for C>≈1.3 the two statistics encounter saturation corresponding to CRR/SUERR of approximately 84% and 66%, respectively.

7.3.3. Hurling, Hockey, & Gaelic Football-Video

Similar summarisation statistics were generated for the remaining test-corpus genres analyzed, i.e. hurling, hockey, and Gaelic football, and **Figs. 7.5**, **7.6**, and **7.7** present

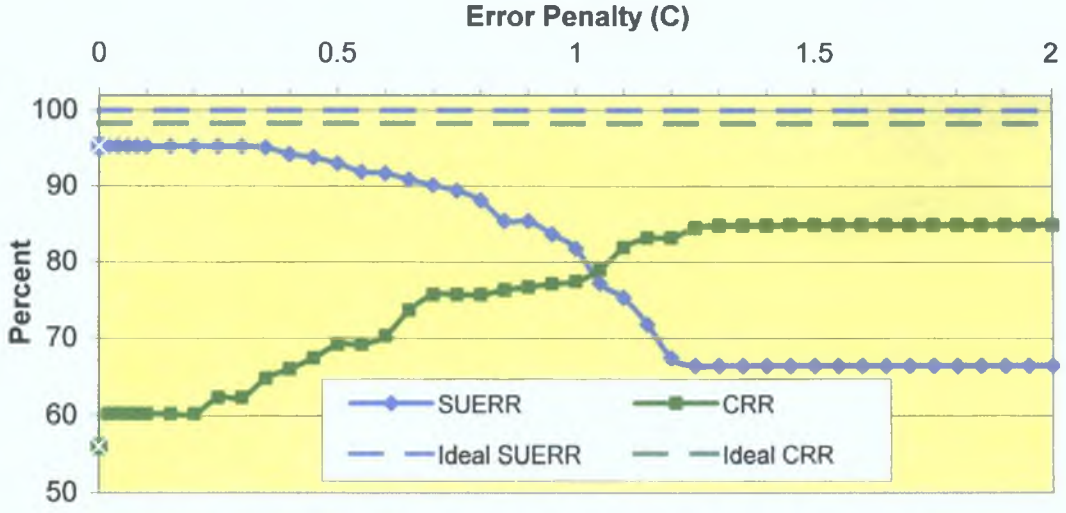


Fig. 7.4. Plot of SUERR/CRR Vs **C** for soccer-video test content.

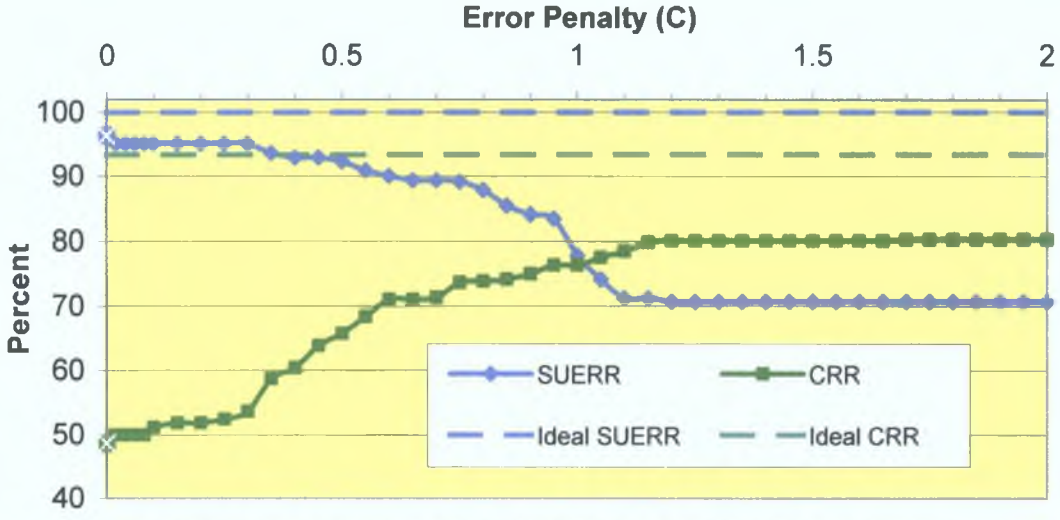


Fig. 7.5. Plot of SUERR/CRR Vs **C** for hurling-video test content.

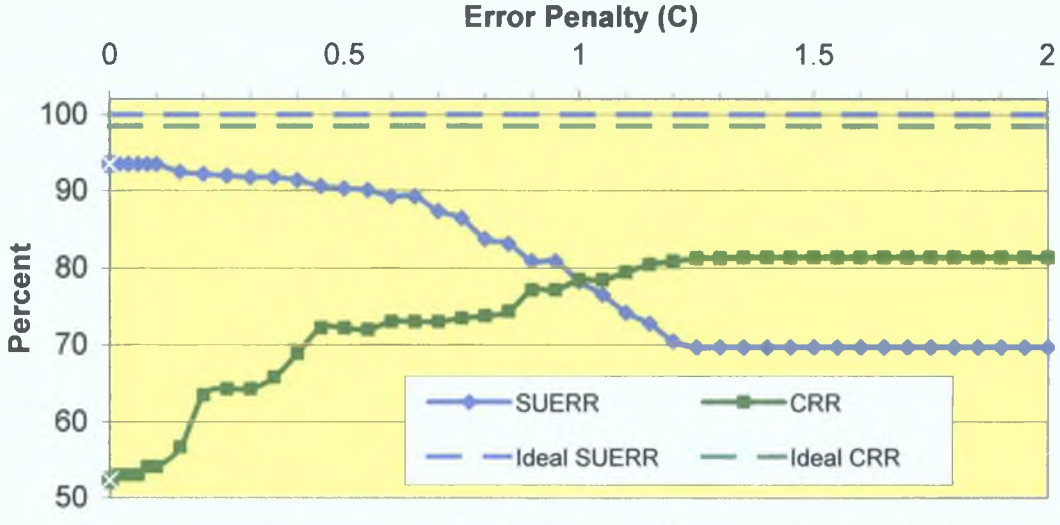


Fig. 7.6. Plot of SUERR/CRR Vs **C** for hockey-video test content.

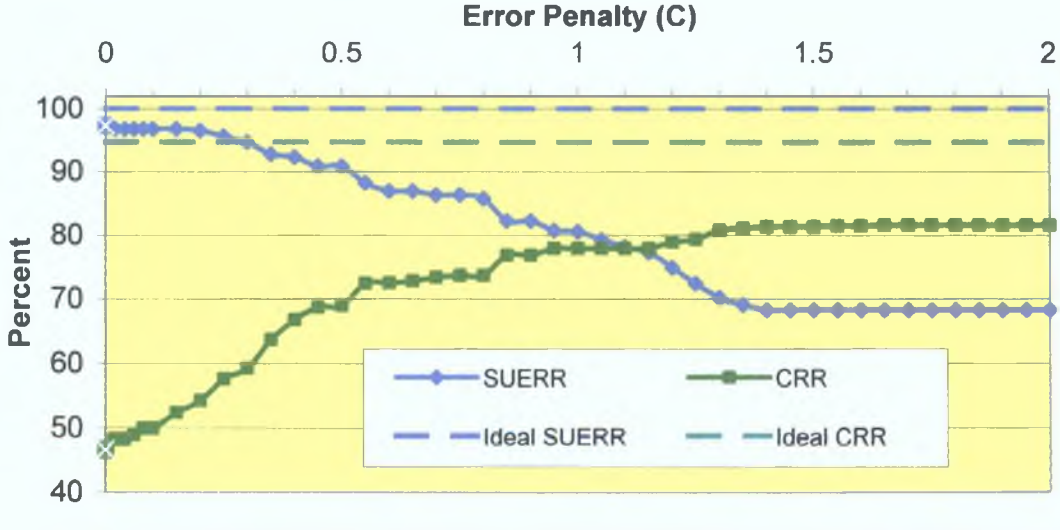


Fig. 7.7. Plot of SUERR/CRR Vs **C** for Gaelic football-video test content.

analogous graphs for these respective cases The relative proportions of content pertaining to SUE-shots were manually determined as 6 6% for the hurling content, 1 5% for the hockey content, and 5 3% for the Gaelic football content Thus, as before the corresponding CRR/SUERR levels for the ideal summarization performance are included in the figures for each case Across all three graphs it is clear that in each case the trends exhibited reflect those previously observed, i e as the value of **C** is increased, the content rejection level increases at the expense of event retrieval towards a saturation point It was observed that the individual saturation points correspond to C>≈1 2 for the hurling-video scenario, C>≈1 3 for the hockey-video scenario, and C>≈1 4 for Gaelic football case

In the above analysis it has been demonstrated how a variation in **C** during the training-phase provides for a variance in the test-phase trade-off between the critical summarization statistics of SUE retrieval and content rejection Moreover, given the adjustment of **C**, the relative responses of each individual sports-genre to the corresponding common set of learned models have been illustrated However, it is desirable to perform a more detailed evaluation of the results obtained, including a cross-comparison of the individual genre performances, such that the overall merits of the scheme are illustrated

7.4. Performance Analysis

7.4.1. Misclassifications

Following a manual test-corpus investigation, the explicit causes for SFV misclassifications were found to be diverse However, it was determined that, as expected, in each case the underlying reasons were related in some way or another to the breakdown of the SUE model as learned For example, it was established that occasionally the SFVs pertaining to some of the positive test-points did not exhibit characteristics consistent with those dictated by the SUE-shot model Hence, if these instances were not rejected at the preprocessing stage, they tended to be misclassified at the pattern classification stage These phenomena, i e ***false-negative*** classifications, are the basis for the non-ideal SUERR values observed in the results above, and their existence indicates a slight SUE retrieval deficiency in the model Furthermore, it was found that some negative test-points tended to exhibit positively biased SFV characteristics in terms of the model definition Such points tended to be misclassified

as positive, thus yielding ***false-positive*** decisions. These phenomena are the basis for the non-ideal CRR values observed in the results, and their occurrence suggests that, in addition to the aforementioned retrieval deficiency, the precision aspect of the scheme is somewhat lacking also.

7.4.2. Optimum Performances & Cross-Genre Evaluation

To better evaluate the results obtained, **Figs. 7.8, 7.9, 7.10, 7.11,** and **7.12** present Cartesian plots of CRR against SUERR for the various values of **C**, for the rugby, soccer, hurling, hockey, and Gaelic football-video test-corpus genres respectively. Once again, the corresponding ideal values are illustrated in each case, however, in CRR/SUERR space these values intersect, and therefore the ideal solutions are represented by unique points. These ideal points are represented in the figures by the symbol '+', while as before the preprocessing level entry points are marked by the symbol 'X'. From the graphs it is evident that as the value of **C** varies the CRR/SUERR curves vary in proximity to their ideal points, i.e. the performance responses vary in proximity to their ideal solutions. On this basis, it is proposed that for each genre, the position of its optimum performance (i.e. its optimum **C**) may be estimated geometrically by determining Euclidean distances (δ) between its ideal point and the points that define its CRR/SUERR curve, and then determining the shortest distance (δ_{opt}). The Euclidean distance is defined in **(7.4)**.

$$\delta = \sqrt{(x_I - x)^2 + (y_I - y)^2} \qquad \textbf{(7.4)}$$

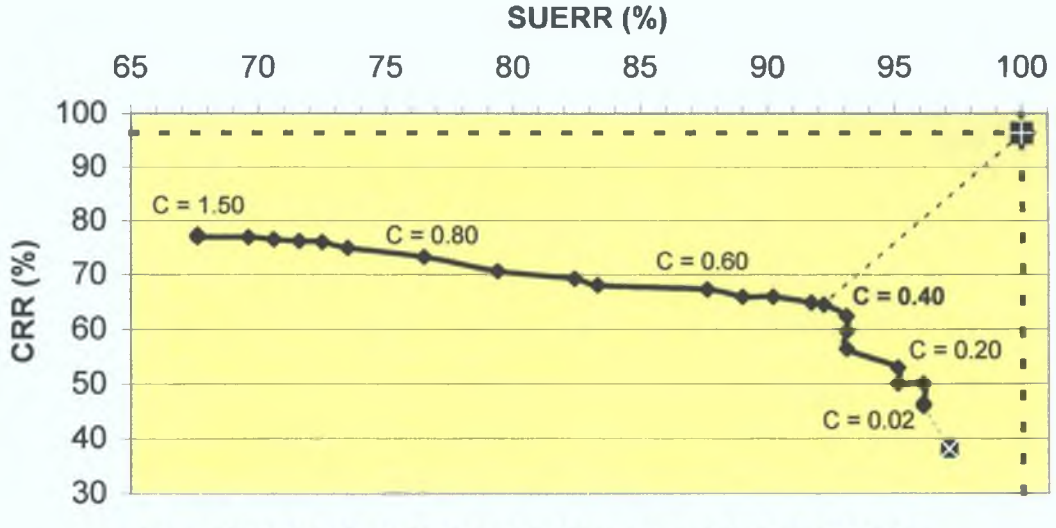


Fig. 7.8. CRR Vs SUERR for $0.02 \leq \mathbf{C} \leq 0.2$ in rugby-video.

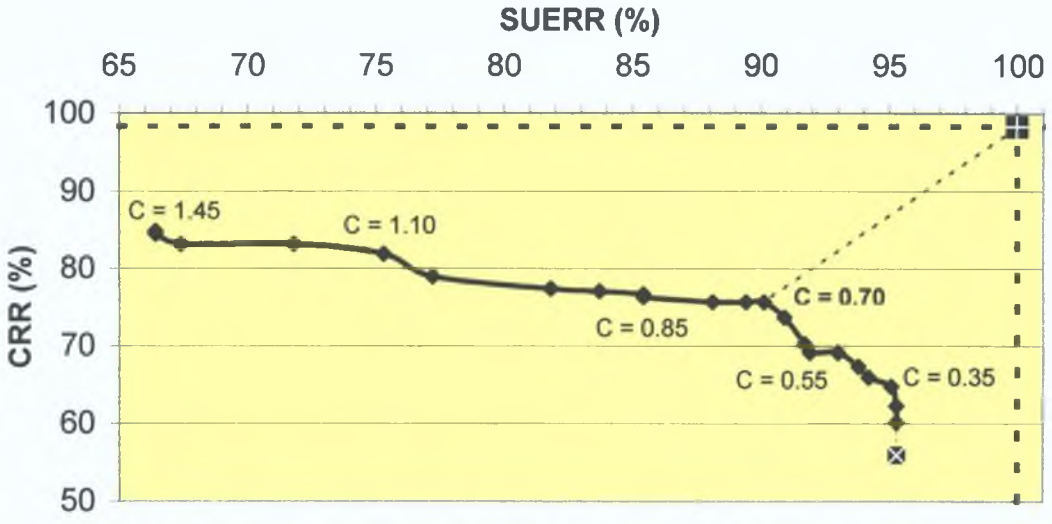


Fig. 7.9. CRR Vs SUERR for $0.02 \leq \mathbf{C} \leq 0.2$ in soccer-video.

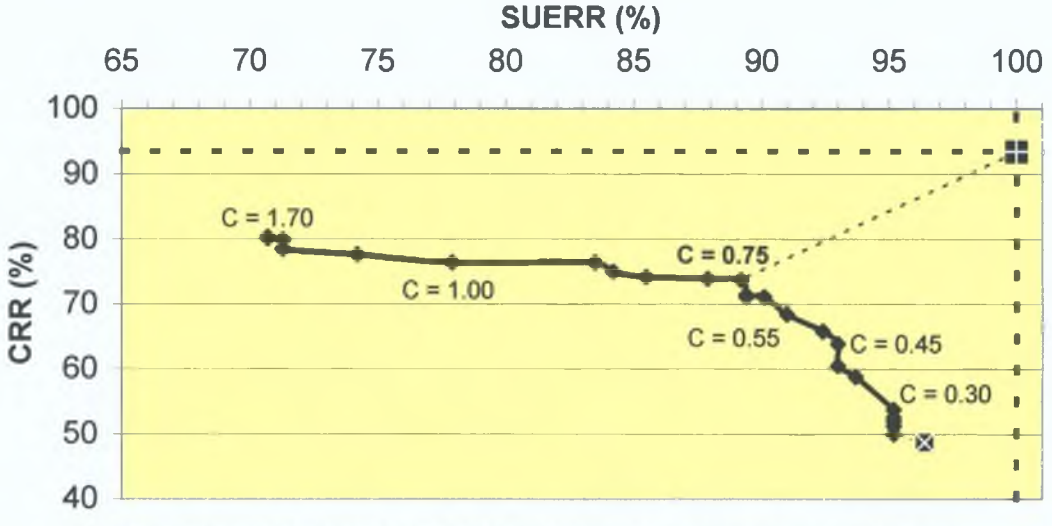


Fig. 7.10. CRR Vs SUERR for $0.02 \leq \mathbf{C} \leq 0.2$ in hurling-video.

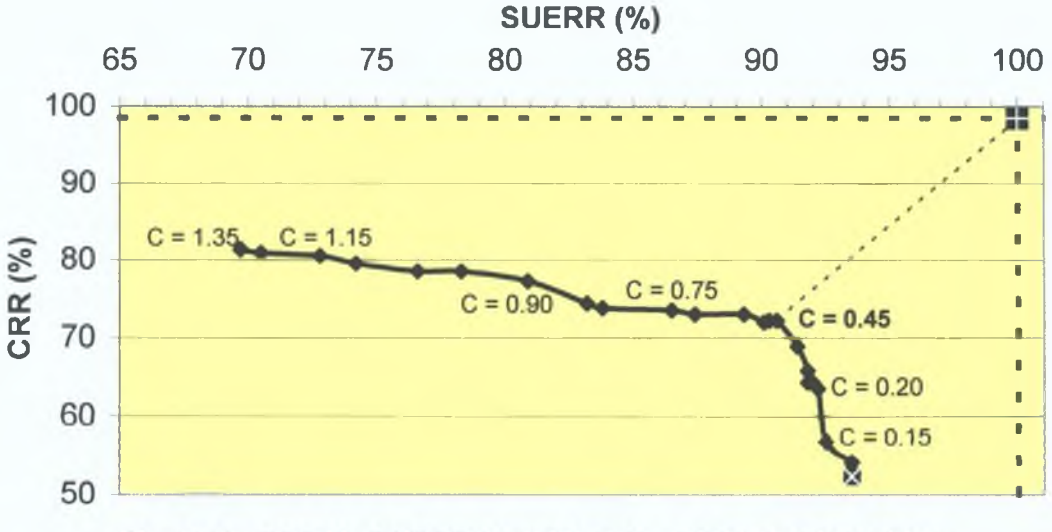


Fig. 7.11. CRR Vs SUERR for $0.02 \leq \mathbf{C} \leq 0.2$ in hockey-video.

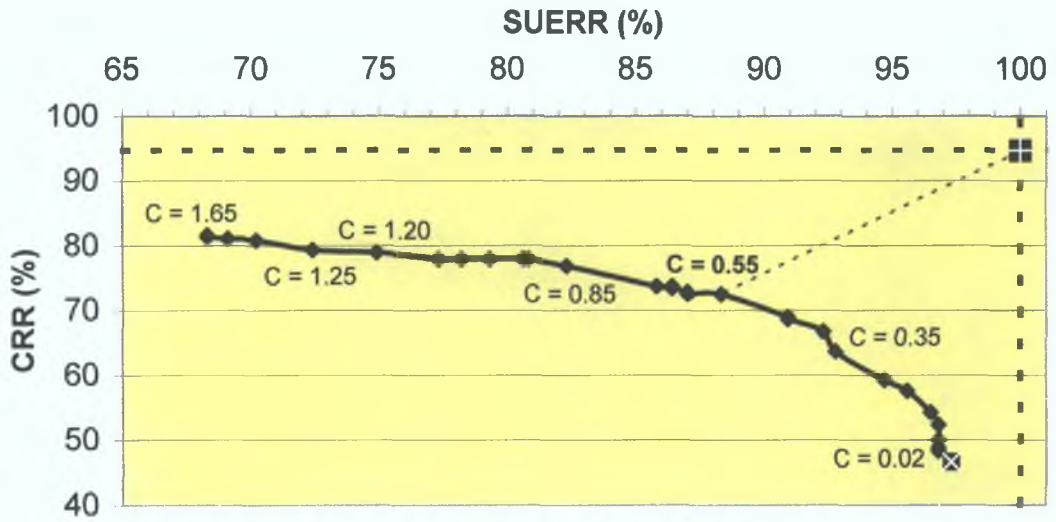


Fig. 7.12. CRR Vs SUERR for $0.02 \leq \mathbf{C} \leq 0.2$ in Gaelic football-video.

Within each of the above graphs, the shortest δ lengths (δ_{opt}) were calculated and are illustrated. For each case these distances, plus their corresponding optimum values of **C**, CRR, and SUERR, are recorded in **Table 7.3**. For each genre, the smaller the δ_{opt} value, the closer the performance comes to realizing the ideal. Furthermore, for all tested genres, the values of δ_{opt} themselves should provide a relatable basis for a cross-comparison of the individual genre responses to the set of classifiers. From the data in Table 7.3 it is evident that the hurling-video scenario exhibits the lowest δ_{opt} value, followed by the soccer-video scenario, then the Gaelic football-video, hockey-video, and rugby-video cases respectively. Hence, of each optimum performances attained, that of the hurling-video context is closer to its ideal than that of any other, and is thus ostensibly the best responding genre of the five. Likewise, that of rugby-video is further from its ideal than that of any of the other genres, and by the same token, it represents the least well responding genre.

Table 7.3. Values for δ_{opt}, and corresponding optimum C, CRR, and SUERR, for each analyzed test-corpus sports genre.

Test-Corpus Genre	δ_{opt}	C	CRR	SUERR
Rugby	32.8	0.4	64.4%	92.2%
Soccer	24.6	0.7	75.7%	90.1%
Hurling	22.4	0.75	73.7%	89.2%
Hockey	27.9	0.45	72.2%	90.6%
Gaelic Football	25.1	0.55	72.5%	88.3%

The performance variances observed across the different genres suggest that for the same set of classifiers, model violations (i e false-positive/negative incidences) were more prolific in one genre than another While the reasons for these variances may be numerous, following a closer investigation of the content, it is postulated that one such reason may be rooted in the underlying pace of the respective field-sport games concerned That is, it is evident from the results obtained that the genres concerning faster paced games, i e hurling and soccer, outperform the others On the contrary, it is the slowest paced game, i e rugby, which is the least well performing of the genres Whereas, the more ambiguously paced games of hockey and Gaelic football exhibit performances between the two extremes Following a manual examination of their respective content, it was observed that the faster paced games tend to contain more live action, i e less play breaks, than the slower games Therefore, the video structure in the faster games, i e hurling and soccer, tends to be more defined, i e there tends to be less scope for contextual content On the contrary, broadcasts of a relatively slower paced game such as rugby, tend to be less restricted in this respect, and tend to have a higher amount of background content e g close-up shots, crowd shots, replays, etc Hence, the slower paced games tend to exhibit a relatively higher sporadic abundance of the features critical to the SUE model deployed, and as a consequence their genres are relatively more challenging in terms of SUE discernment on that basis Given these observations it is postulated that it is the genre dependant trait of game pace primarily accounts for the respective performance variances observed in the results illustrated

7.4.3. Optimum Error Penalty Values

It is proposed that the above supposition is to some extent corroborated by the respective values of **C** that are required to yield the optimum performance responses for each genre Table 7 3 presents these values Recall that the error penalty **C** determines the relative significance of training errors compared to the width of the SVM margin in the objective function to be optimized That is, a higher error penalty limits the number of training errors tolerated by the SVM It was noted that the optimum performance of the least well performing genre (i e rugby), is obtained at an error penalty value of C = 0 4, and those of the relatively faster paced games (i e hurling and soccer) are yielded at C = 0 75 and C = 0 7, respectively Thus, the overall trait is that the optimum error penalties in the faster paced games are shown to be higher than those of the slower paced games It is postulated that this trend could be a reflection of the relative

separability of the test-content of the respective genres, probably attributable to game pace as discussed above For example, assuming the rugby-video test content exhibits the least separable test-data of the genres, this might account for the relatively lower error penalty (i e the 'softer margin') that realizes its optimum performance That is, assuming that many of the positive and negative test-points tend to be in relatively close proximity (i e overlapping), a better response may be obtained when applying a decision function that reflects the more general trends of the training data (low C) to that which is more fitted to the training data (high C), such that most may be still correctly classified In contrast, assuming that the test content of the faster paced games is more objectively separable as discussed, this would account for the affordability of a higher error penalty in realizing its optimum response That is, given a clear separation of the test data, a decision function fitted to the training data may provide no more or less accuracy on the test examples than the more generalized decision function, thus suggesting why it is possible for the optimum response to be realized by a high value of the error penalty C

Overall, from these results it is evident that the optimum performances in each genre are obtained for error penalty values lying within the range $0\,4 \leq C \leq 0\,75$ Therefore, if desired, by choosing an appropriate value for **C**, the scheme response may be tailored (tuned) towards realizing the optimum performance for any of the five particular field-sport genres analyzed

7.4.4 Global Optimum Error Penalty

To gauge the overall performance of the scheme in terms of global SUE retrieval across all five analyzed genres, **Fig 7 13** presents a CRR/SUERR plot, for the results of the classification of the test-corpus content taken as a whole In this case, the mean preprocessing levels correspond to 95 9% SUERR and 48 3% CRR (see Table 7 2), and it was determined that combined, all 850 SUE-shots account for 3 8% of the 90-hour test-corpus content, thus rendering an ideal content rejection ratio value of 96 2% As illustrated, this ideal CRR level again intersects the 100% SUERR level in yielding a unique ideal point in CRR/SUERR space As before, the shortest Euclidean distance from this ideal point to those constituting the CRR/SUERR curve was determined, and based on this metric the optimum value of the error penalty for this global scenario was determined to be **C = 0 5** At this point the statistics correspond to 68 4% CRR and 91 3% SUERR, approximately Hence, this optimum performance provides for the

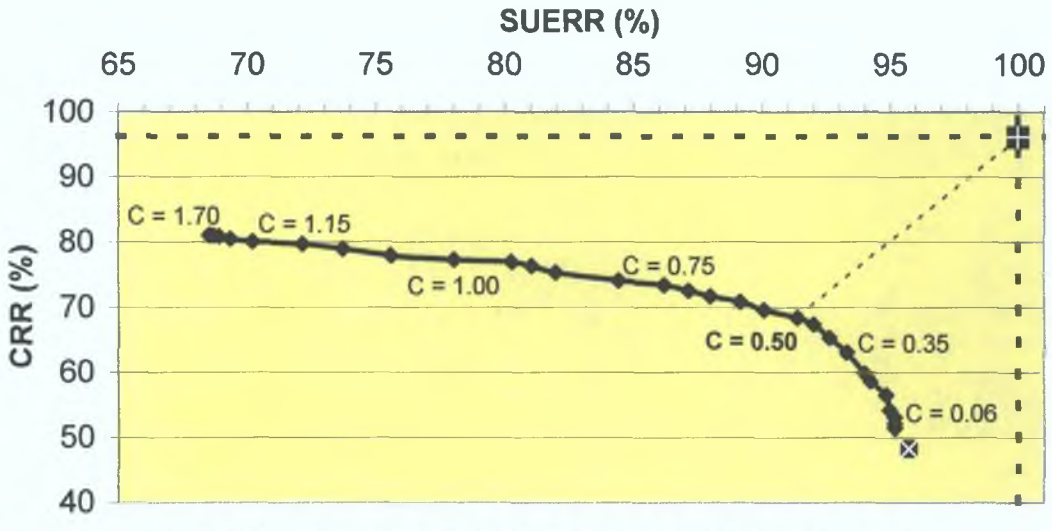


Fig. 7.13. Global CRR Vs SUERR plot for $0.02 \leq \mathbf{C} \leq 0.2$.

summarization of a FSV down to, on average, 31.6% of its original broadcast length, where the summary includes at least 91% of all SUEs.

7.4.5. Practical Performance Optima

In the above evaluation the optimum performances for the individual test-corpus genres were determined, and Table 7.3 presents these in terms of respective CRR/SUERR statistics along with the corresponding values of **C**. However, these optima were discerned analytically, and while they have been effectively exploited in realizing a cross-genre response comparison, they may not represent the most sensible performance levels attainable in terms of a practical scenario. For instance, from this data it is evident that in all cases the analytically discerned optimum SUERR levels lie in and around the 90% mark. However, these levels may not suffice in an application where, in terms of the CRR/SUERR trade-off, event retrieval is deemed paramount. That is, it may be required to bias the classification (as long as the preprocessor limits permit so) further towards the SUERR ideal (at the expense of CRR), by manually choosing a more appropriate value of **C**. To provide an indication of what is attainable for each test-corpus genre, and to complement the results presented in Table 7.3, **Table 7.4** presents for each case (determined from Figs. 7.3, 7.4, 7.5, 7.6, and 7.7), the highest CRR level corresponding to the maximum achievable SUERR level short of the preprocessor limit.

Table 7 4 Maximum SUERR levels achievable for each genre before reaching those of the preprocessor limit Also shown are corresponding values of C and CRR

Test-Corpus Genre	CRR	SUERR	C
Rugby	50 0%	96 1%	0 10
Soccer	64 8%	95 1%	0 35
Hurling	53 7%	95 2%	0 30
Hockey	63 5%	92 2%	0 20
Gaelic Football	52 4%	96 8%	0 15

7.5. Performance Evaluation

7.5.1 Performance Accuracy

As declared at the outset of this thesis, in the field of sports-video technology development, the ideal target for a genre-independent scheme is for its solution to be capable of yielding consistent performances across multiple genres, with accuracy comparable to that offered by a genre-specific approach Given this challenge in terms of field-sports-video, it is desirable to ascertain to what extent the solution attained by the scheme developed achieves this objective

7 5 1 1 Overview & General Conclusions

As shown in *Section 7 4*, in terms of the FSV case study undertaken, the generic scheme developed in this thesis provides for relatively consistent performance levels in automatic summarization across five distinct FSV genres It is argued that the five sports analyzed present a good diversity both in game nature and game pace Given this, it is concluded that the initial aspect of the ideal target (i e a consistent performance across multiple genres) has been successfully achieved for the case study undertaken

In *Section 7 4 4*, it was shown that the global optimum summarization performance (i e 68 4% CRR and 91 3% SUERR) provides for the summarization of a FSV to, on average, 31 6% of its original length, where the summary includes at least nine-out-of-ten of the entire game's SUEs Given that the ideal average CRR is 96 2%, this corresponds to the retrieval of 27 8% (= 96 2% - 68 4%) of non-SUE-shots As discussed previously, the retrieval of superfluous content in addition to the true SUE-shots corresponds to false-positive misclassifications Clearly, it is desirable to minimize the retrieval of this content However, as illustrated by the CRR/SUERR graphs in

Section 7.4, while it is possible to reduce this towards its ideal level by making the overall retrieval more selective, this also results in an increase in false-negative classifications, i.e. a reduction in the true SUE retrieval performance of the scheme. Therefore, it is concluded that, as is the case in many retrieval schemes, the false-positives are a byproduct of the system that simply have to be tolerated. However, recall that false-positive classifications arise from circumstances where abnormal negative test-points exhibit positively biased characteristics in terms of their shot-feature vector (SFV) representations. Given that each SFV component conveys an innate level of content excitation (i.e. visual activity, audio activity, etc.), based on a manual investigation it was found that while in the strictest sense they constitute retrieval errors, the false-positive episodes tend to exhibit content of a high significance level. For instance, it was found that it was not uncommon for the critical feature excitations indicating SUE-shots to temporally supersede them, and be sustained throughout shots constituting subsequent content, e.g. the reaction-phase segments. That is, given a detected SUE, it was found that now and again some of its reaction-phase content was also assigned positively biased SFVs, and thus retrieved as an add-on to the preceding SUE-shot. Again, while this behaviour is erroneous in the strictest sense, since some users may find the rapid presentation of SUE-shots in isolation visually disturbing and/or incomprehensible, it is arguable that the retrieval of such content may be regarded as valuable in terms of conveying the contextual perspective of its corresponding SUE. Hence, in circumstances where longer summaries are tolerable, the tagging of a small amount of contextual content to the detected events may be seen to constitute a beneficial by product of the summarization process. It was manually determined that the misclassification of reaction-phase content as described accounted for the large majority of non-ideal CRR results. Other sources of false-positive classification corresponded to episodes such as near misses, controversial incidences, etc. For instance, depending on their relative significance in the game, such episodes were found to sometimes exhibit critical feature excitation on a par with that of SUE incidences, thus leading to their mistaken retrieval. However, in circumstances where the conditions on the summary length are not strict, there once again exists an argument for suggesting that the inclusion of these events in a generated summary may be considered favorable.

In short, the video summaries that are achievable with this multi-genre scheme have been shown to consistently encompass the large majority of the narrative-critical events. Although, a consequence of this high retrieval performance is the inclusion of

some extra non-narrative-critical content, on the basis of the above reasoning, it is concluded that the superfluous content additionally retrieved tends to concern that of at least quasi-significance, and/or is typically constructive in conveying the contexts of the detected events

7.5.1.2 Comparative Performance

Given the above, it is shown that the performances obtained via the scheme developed provides for a favourable solution to the summarization task However, as described in *Chapter 2*, there exists a variety of previously established alternative schemes for semantic sports-video content analysis (of both genre-specific and genre-independent methodologies) that also declare successful results within their respective domains Therefore, to fully expound the merits of the scheme developed, it is desirable to cross-compare its performance with those professed in the alternatives However, it is recognized that the conclusions of any such comparison would be compromised by the fact that there is no correlation whatsoever between the data corpuses from which the respective sets of experimental results were drawn Moreover, between these previous works and the work undertaken herein, the only common sports-genre analysed was that of soccer-video, i e no prior account of analysis of rugby, hurling, hockey, or Gaelic football-video was discovered in the research of the state-of-the-art While clear implication of this fact is that the scheme developed is inherently novel in this respect, it clearly rules out the prospect of performance comparisons for any other genre except for soccer-video Nonetheless, to provide at least some indication of the relative performances of the alternative schemes compared with that provided by the scheme herein, it was considered desirable to perform a cross-comparison of their respective soccer-video analysis performances

From *Chapter 2* it was concluded that there were sixteen previous works incorporating some aspect of soccer-video analysis However, amongst these it was evident that even within this restricted domain there tended to be a considerable variance in the specific task definitions, rendering a performance comparison unfeasible in many cases That is, in contrast to the work of this thesis, few of the alternative works explicitly outline the task of score-update episode (goal) shot retrieval towards soccer-video summarization as a specific scheme objective Rather, in the majority of cases this task tends to be addressed implicitly, i e within the remit of a more nonspecific highlight detection objective For example, the works of [15], [29], [42], [47], and [48]

provide results for soccer-video analysis that correspond to their accuracy in the combined retrieval of all of the most effervescent moments in such content (i e the retrieval of highlights including, but not restricted to the goals) Many of the remaining works present results that correspond to altogether different soccer-video analysis tasks including object identification and tracking [14, 22], soccer-video mosaicing [19, 14], high-level structure segmentation [20, 24, 49], shot-view classification [45], and placed kick detection [23] In fact, it was found that only three of the sixteen specified schemes, i e [21], [36], and [27], declare explicit results corresponding to the recognition of goal incidences

The first of these [21] proposes a genre-specific methodology for detecting a wide range of semantic events in soccer content However, as described in *Chapter 2*, the scheme is entirely dependent on the availability and accuracy of player/ball position knowledge, and moreover, assumes this is on hand (i e it is suggested that this knowledge may be inferred from a tracking system that interprets signals emitted by transponders attached to the players and ball during the game) Furthermore, the system also requires some level of manual input corresponding to certain referee decisions, e g start/stop of each period, etc Therefore, while excellent results are reported for the specific task of goal recognition, it is concluded that comparing this scheme with that developed by this author would be unproductive since, in contrast, it does not relate to a fully automated approach The second of these works [36] proposes a multi-modal soccer-specific framework for the task of goal recognition However, while the scheme reports high accuracy in this task, it is only evaluated on a 2 5-hour test-corpus encompassing six goals, i e merely six positive test-points Hence, while the results of this scheme may be considered encouraging, given this relatively small test corpus, a more comprehensive evaluation is warranted before concrete conclusions can be drawn

In contrast to [21] and [36], the goal recognition technique in [27] does relate to a complete and fully automated approach, and the results are drawn from a relatively sizeable test-suite (i e 11 5-hours of content encompassing 30 goals) Therefore, a comparison of its performance (hereafter known as the 'external scheme') with that generated for the soccer-video aspect of the scheme herein (hereafter known as the 'internal scheme') was considered However, while the authors of this soccer-specific external scheme present the accuracy of their approach in flagging the occurrence of goal incidences, unlike the internal scheme, they do not provide results indicating the extent to which their scheme can extract the individual goal shots in isolation towards

generating summarized versions of the content Nonetheless, given that the authors advocate goal detection recall rate as the most important performance quantifier [27], and have accordingly optimised their solution towards this, a cross-comparison of this statistic with that of the optimum SUERR performance attainable in the internal scheme remains valuable To this end, recall that for the internal scheme, the optimum value of the error penalty for the soccer-video scenario was determined to be C = 0 7, which yielded corresponding optimum SUERR of **90 1%** (for 75 7% CRR) In the external scheme, although tested on a comparatively smaller test-corpus (i e 11 5-hours compared to the approximately 28-hours of soccer-video analyzed in the internal scheme), it was reported that amongst the 30 goals encompassed within its test-corpus, 27 were detected accurately (with 32 false-positive detections) This corresponds to a comparative SUERR of **90 0%** Therefore, at the professed optimum performance points of both schemes, the SUERR of the internal scheme effectively matches that of the goal recall rate of the external scheme (for some level of false-classification in both cases) Hence, in terms of the most vital performance quantifier, the two schemes may be considered to provide reasonably comparable performances However, the soccer-video summarization task of the internal scheme represents a single component of a wider FSV summarization remit That is, unlike the genre-specific external scheme, the internal scheme has been shown to yield relatively consistent performances generically across a range of other sports genres, including rugby, hurling, hockey, and Gaelic football Based on these results, and their relative proximity of their optimum responses to the genre-specific recall benchmark of [27], it is concluded that in terms of the task of developing a generic approach, the developed scheme represents an excellent approximation to the ideal target

In completing the discourse on scheme accuracy, it remains to be discussed by what means the performances levels of the approach may be further improved in terms of a discussion on relevant potential future research This topic will be addressed in the subsequent chapter

7 5 1 3 Generalization Performance

The system performance in terms of accuracy and consistency has been shown to be favourable when trained and tested across the sports genres cited As a final performance evaluation criterion, it was considered desirable to ascertain how the scheme would perform given a FSV genre that was not represented in the training data

(i.e. an ***unseen*** FSV genre, the characteristics of which have not been exploited in the learned SUE model). In light of the deficiency of significant amounts of content relating to FSV genres not already represented, it was decided to synthesize such a scenario using the content to hand, and then postulate on that basis. Specifically, it was proposed that the system be retrained using data corresponding to four of the five genres originally represented, and then tested explicitly on the unseen genre. It was anticipated that an approximation of the systems generalization ability could then be inferred by a comparison of the performances attained for the seen/unseen training scenarios. However, the elimination of certain training data in the training-phase could have detrimental effects on the learned SUE model, to the extent that the seen/unseen performance comparison effectively relates to two non-alike schemes. Hence, it was decided that the data pertaining to the genre with ostensibly the least contribution to the training set be disregarded, and thus nominated as the unseen genre in the test-phase. Recall that of the 883 positive training points (PTPs) constituting the training-corpus, those corresponding to the hockey-video genre represented the least proportion (see Table 1.3). Given this, SVM classifiers were trained as before but this time using only the soccer/rugby/hurling/Gaelic football-video training dataset as input. The resulting classifiers were then utilized in testing the response of the (unseen) hockey-video test content to the learned models. As before, CRR/SUERR summarization statistics were generated, and **Fig. 7.14** presents the corresponding graphs according to their variation

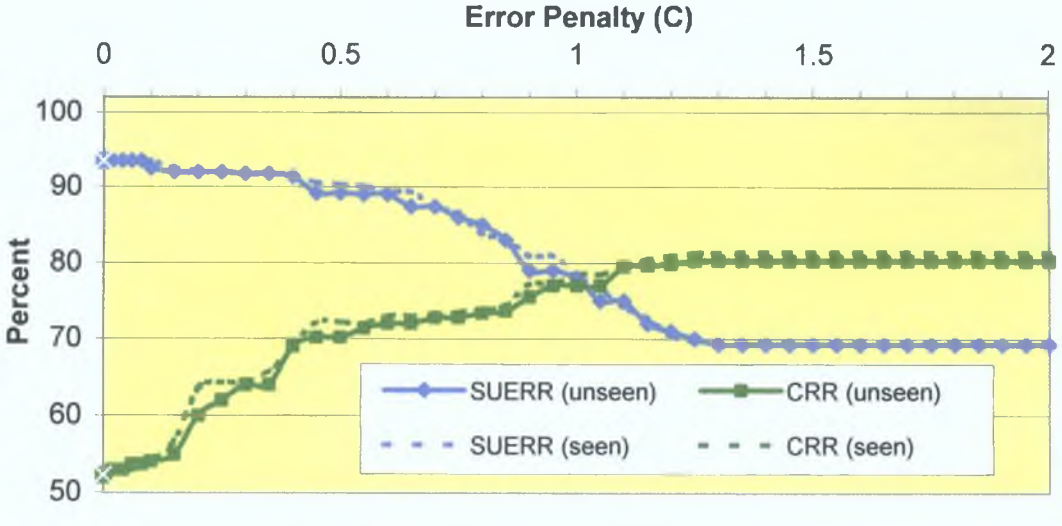


Fig. 7.14. Comparison of hockey-video summarization performance for seen/unseen training scenarios.

with **C** Also illustrated in this figure are the equivalent graphs corresponding to the hockey-video seen training scenario (reproduced from Fig 7 6) By comparing the trends, it is evident that the performance exhibited in the unseen training scenario is relatively consistent with that of the seen case previously observed That is, while the solution developed in the seen scenario may be judged to slightly outperform that of the unseen case, the differences between the two statistics at any given point is generally minor On the basis of this observation, it is concluded that the scheme developed generalizes well across the FSV genres represented Given this, and justified by the wide-ranging characteristics of the genres already represented, it is assumed by extension that the solution should generalize well across unseen FSV genres, i e provide satisfactory results consistent with those previously observed for any sports-video satisfying the defining criterion specified in *Section 4 1 1*

7.5.2 Speed Performance

A critical aspect of any real system is its speed/efficiency response, and hence it was considered desirable to investigate this in relation to the performance of the scheme developed herein Details of this investigation may be found in *Appendix F*

7.6. Chapter Summary

In this chapter a detailed description of the training/testing phases of the experiments performed was presented This was then supplemented by a comprehensive evaluation of the results obtained in terms of the summarization task Firstly, the issues critical to the SVM training procedure were discussed and appropriately configured, so that the SVM could learn to the best of its ability the underlying decision function of the training data In testing the scheme, the performance of the preprocessing filter was described, in terms of its ability to reject irrelevant content prior to the pattern classification stage Combining this with the SVM-driven SUE-shot detection process, the overall summarization performance of the scheme was then analysed, which included (i) a description of how the retrieval varies with the choice of the error penalty, (ii) a cross-genre performance evaluation (including how the system response may be tailored to realize the optimum performance for each genre), (iii) a postulation for the reasons for inter-genre performance variance, and (iv) a declaration of both global and practical performance optima Given this, a detailed evaluation of the performance was provided,

including a comparison of the scheme accuracy to that of a state-of-the-art equivalent scheme

Chapter 8

Thesis Synopsis, Conclusions, & Future Work

This thesis introduces novel technology towards finding a generic solution to the problem of extracting automatically generated summaries from fields-sports-video (FSV) content Via the evaluation presented in the previous chapter, it was shown that the technology developed provides for a successful realization of this objective Following a brief thesis synopsis, this chapter provides a discussion on the conclusions drawn, and the potential future work aspects of this research, with respect to the scheme developed herein, and in reference to the field of sports-video analysis in general

8.1 Thesis Synopsis

In the opening chapter the motivation for the problem of video summarization was introduced, followed by a discussion on the more specialized area of sports-video highlighting, with particular reference to the amenability of such content towards event detection-based summarization Given the dichotomy in approach methodologies for this topic, the research objective to be targeted in this thesis was then formally introduced, i e the development of a generic solution for event detection-based summarization in FSV Following this, the proposed realization approach was outlined

In the second chapter a synopsis of the current state-of-the-art technology for sports-video analysis was provided From this discourse it was evident that the majority of the approaches in the literature concern that of a genre-specific methodology

However, many generic frameworks are also described, in which multiple sports genres are analysed based upon a common hypothesis The limitations of the state-of-the-art were then discussed

In the third chapter, background knowledge concerning the principles of digital video was provided Specifically, the topics of colour space models, video structure modeling, and data encoding/compression applied to digital video, were introduced Finally an overview description of the video-encoding standard critical to the work of this thesis (i e MPEG-1) was provided

In the fourth chapter a generic hypothesis for event detection-based summarization in FSV was proposed Specifically, following a description of the features deemed both necessary and sufficient in characterizing this particular supergenre, features that were deemed to reliably indicate score-update episodes (SUEs) were then inferred from a training-corpus investigation On this basis, it was proposed that the detection of the prevalence/intensity of these critical features (CFs) should provide a reliable basis for the detection of SUE-shots, towards a favourable summarization solution Following an introduction to the algorithm proposed for the shot boundary detection task (the description of which is expounded in *Appendix A*), it was proposed that evidence pertaining to one of these CFs be exploited in the development of a shot-level pre-preprocessing filter (in conjunction with an externally developed advertisement detection algorithm) It was then proposed that the remaining CF evidence be aggregated towards the extraction of shot-level semantics, the patterns of which should constitute a reliable basis the detection of SUE-shots from the (preprocessed) content

In the fifth chapter the details regarding the implementation of the hypothesis proposed in *Chapter 4* were described To some extent, the implementation approach reflected the nature of the content representation used, i e MPEG-1, and utilized the set of signal-level data extraction tools described in *Appendix B*

In the sixth chapter, potential avenues for the task of pattern classification were explored Given the motivation for a discriminative machine-learning approach, three of the most commonly advocated discriminative classifiers were compared On the basis of the opinions conveyed in the literature, Support Vector Machine (SVM) technology was favoured (a comprehensive overview of which is provided in *Appendix D*, with *Appendix E* providing a description of the actual implementation used)

In the seventh and preceding chapter, a detailed description of the experiments undertaken in this thesis was provided At the outset, the issues critical to the SVM training-phase were discussed, and it was then described how the SVM was trained In testing the scheme, the performance of the preprocessing filter in rejecting irrelevant content was illustrated Combining this with the SVM-driven SUE-shot detection process, it was shown via an evaluation stage that the overall scheme provides high SUE retrieval and content rejection statistics in terms of the summarization task A description of the speed response was provided in *Appendix F*

In completion of this thesis, the outstanding issues that remain to be addressed in this final chapter relate to the conclusions that may be drawn from the results obtained in terms of the research objectives targeted, and also a discussion on the potential future work aspects

8.2. Conclusions

The research objectives of this thesis were explicitly stated at the outset of this thesis, i e in *Section 1 5* of *Chapter 1* In short, the objective was to develop a generic solution for event-detection based summarization of field-sports-video, whereby the attained solution provides consistent performances across the various sports genres that constitute this supergenre (see Table 1 1) Furthermore, it was stated that the performances should exhibit accuracy that rivals that of the genre-specific equivalent solutions On the basis of the following reasoning, it is concluded that these objectives were met successfully

It was shown via the analysis presented in *Section 7 4*, that the scheme developed provided for a relatively consistent level of summarization performance across five distinct field-sports-video genres, and critically, in *Section 7 5 1 3*, that this performance generalizes well across unseen FSV genres As outlined in *Section 7 5 1 1*, on average, the scheme provides for summarization down to approximately 30% of the original input video, where the summaries generated include over 90% of the score-update episodes (SUEs) As explained in *Section 7 5 1 2*, this is a first-class SUE recall rate that is comparable with that of a state-of-the-art, genre-specific equivalent scheme It was also acknowledged that although the performance does not provide for ideal levels of summary reduction (i e there is a certain level of false positive SUE retrieval), it is argued in *Section 7 5 1 1* that the inclusion of such content may be of interest to the

user, on the basis that while the false retrievals do not relate to scoring events, they do tend to correspond to exciting moments

The final significant point is that the successful realization of the objective of this work represents a novel instance of where a meaningful boundary has been put on a generic solution to the problem of sports-video summarization In doing so it is argued that a significant improvement has been made on the prior art, and on that basis this thesis represents a significant contribution to the field

8.3. Furthering The Scheme Developed

In this section, potential future research that relates exclusively to the scheme developed are proposed, such that its overall performance both in terms of increased accuracy and/or implementation efficiency might be improved

8.3.1. Further Critical Features

The FSV summarisation scheme devised is rooted in the detection of SUEs, based on the extraction and aggregation of evidence pertaining to six critical audiovisual features As described, one of these features is employed in the pre-processing stage, while the remainder constitute the basis for the pattern classification phase of the scheme In *Chapter 4* it is explained how these six features were chosen following a manual investigation of the training-corpus, the aim of which was to establish which features might be potentially indicative of SUEs Following this investigation, the effectiveness of each chosen feature in providing such indication was determined, thus justifying their selection for the model However, it is recognised that FSV sequences exhibit additional critical features that could ostensibly contribute constructively towards furthering the performance accuracy in the detection of SUEs To this end, the following sections discuss the preliminary investigations undertaken into these features, and explain the issues that have so far not been exploited, but which could be targeted as future work

8.3.1.1 Identification Of Digital Video Effect Transitions

As explained in *Section 4.5*, due to the typically high-tempo nature of FSV content, during the live-action segments the broadcast director has little chance to utilize shot transition types other than abrupt shot-cuts However, during a break in the play, he/she typically exploits the chance to use (digital) video effects (DVEs) in constituting

such, i e dissolves, wipes, and morphs As explained, the moments immediately following SUEs typify such breaks in play and therefore it is not uncommon for DVE-transitions to be used in these periods (e g delimiting the reaction-phase shots and/or the multiple viewing angles of subsequent replay segments) Hence, if all transition types could be reliably detected, and furthermore, if discrimination between cuts and DVEs could be achieved, given that the latter are typically prevalent following SUEs, their identification might contribute to further improving the SUE detection accuracy of the scheme Conventionally, the problem of detection, and moreover, the identification of shot transitions other than hard-cuts is considered a challenging task in the field of digital video processing The topic remains a very active area in the field [113], and the more contemporary literature suggests that there has been some considerable progress made For example, Lienhart [114] proposes a technique that claims reliable dissolve detection, and significantly, Naci *et al* [115] propose a scheme in which it is asserted that reliable discrimination between the various types of shot transition is possible Thus, a potential future work task might involve developing/sourcing a scheme that improves upon that used herein [79] in providing for the reliable detection and identification of DVEs, and once finalized, applying it as described to gauge any positive effect such evidence may have on the performance accuracy of the scheme

8 3 1 2 Scoreboard Text Recognition

Recall that one of the six critical features exploited in the SUE detection hypothesis relates to the update of the scoreboard graphics That is, it was shown that following a SUE it was typical for the scoreboard to be temporarily suppressed during its update procedure In fact, it was determined from the training-corpus that this phenomenon was observed in at least 61% of the SUEs observed (see *Section 4 2 2 4*) Given a detected scoreboard, a technique for the detection of this scoreboard suppression was proposed based on a mode-template-differencing methodology, and the accuracy of this was illustrated for a variety of scoreboard formats However, for the remaining 39% of cases, in which the scoreboard update procedure occurs on screen, the aforementioned technique will be unsuccessful in detecting the scoreboard activity Therefore, a potential future work task involves a rectification of this situation, i e the development of a scheme where both on-screen and off-screen scoreboard updates are flagged in the system to an equal degree

Given a detected scoreboard, the desired ideal would be to develop a scheme that can reliably flag a change in its numerals that indicate the score tally If such a technology was realizable, SUE locations may then be indicated irrespective of whether or not the actual updating procedure occurred off-screen It is proposed that the first task in developing such an algorithm would be to detect the characters constituting the text that comprises the detected scoreboard graphic To this end, it is proposed that an optical character recogniser (OCR) might be used However, following a preliminary investigation into the attributes of extracted training corpus scoreboards it was found that, due to the typically small size of the graphic within the images, and given the image resolution used (i e CIF), the text-characters typically emerged blocky or 'pixelated' Furthermore, as is to be expected, there seemed to be problems induced by the spatial compression employed in the encoding of the images That is, the sharp edges of the scoreboard text, such as those required to convey the contrast between the foreground/background, tended to be softened or blurred by the compression algorithm used This is a common consequence of most spatial compression algorithms, which tend to 'step' sharp edges by introducing an intermediate pixel value, between the two edge extremes These two phenomena are problematic for the character recognition process, and it is thus concluded that prior to developing an OCR-orientated technology for extracting text-based semantics from FSV scoreboards, these two issues would have to be given due consideration

However, given that these issues may be overcome, another factor that would tend to hamper the recognition of the text characters is the transparency of the scoreboards, which was alluded to in *Section 5 1 4 1* It was observed that it is not uncommon for FSV scoreboard graphics to exhibit some degree of transparency, a characteristic that is purposely employed to limit the occlusion disturbance to the viewer However, a consequence of this is that the luminance values of the scoreboard background planes are subject to transparency-noise, the effect of which is that the background luminance is typically unstable for a moving camera scenario It is anticipated that this phenomenon could have detrimental consequences for the luminance-based segmentation (binarisation) of the scoreboard text into foreground/background regions, which would be required prior to the application of an OCR Recall that to overcome the effects of transparency-noise in the development of the scoreboard suppression detection task, a contrast-enhancement step was introduced However, it is unanticipated that this remedy would be sufficient to overcome the

effects transparency may have on the text segmentation This represents another challenge that must be resolved prior to the realization of reliable OCR-based scoreboard text identification

Assuming the abovementioned issues may be overcome, i e given reliable scoreboard text recognition, the next step would involve detecting the numerals of interest (i e those representing the score tally) amongst the remainder of the text constituting the detected scoreboard graphic It is proposed that this could be achieved by simply exploiting the fact that at the outset of each game, the score tallies are set to zero, and therefore may be realized by employing some zero-detection mechanism, given the OCR output Once the zeroes are detected, a similar image differencing mechanism may be applied exclusively to their corresponding positions within the graphic, such that the minute changes in spatial pixel luminance associated with an on-screen tally update may be detected and then flagged to the system as described

In summary, it is concluded that there are several potentially problematic aspects associated with the challenge of on-screen scoreboard tally update detection, which have served to discourage its development in the scheme so far An investigation into how these maybe overcome represents a clear opportunity for future work

8 3 1 3 Commentator Vocal Pitch Tracking

Another of the six critical features exploited in the SUE detection hypothesis relates to audio energy Specifically, it was shown that following a SUE it was typical for the energy level of the audio track to be increased, particularly in the speech-band frequency range In fact, it was determined for the training corpus FSVs that, on average, 84% of all observed SUEs exhibited peak audio levels that exceeded corresponding broadcast mean levels (see *Section 4 2 2 3*) In exploiting this, a speech-band audio level tracking mechanism was developed based on the extraction of signal-level subband scalefactor evidence from the compressed domain This mechanism was shown to exhibit good performance in terms of SUE indication However, as well as a surge in the audio energy level of the commentator vocals during exciting moments, it was found that it was not uncommon for an increase in the commentator vocal pitch to be also perceived Thus, a further future work task might concern exploiting this characteristic in developing a mechanism for the tracking of commentator vocal pitch, such that the contribution of such evidence to furthering the accuracy of SUE detection may be gauged That is, it would be desirable to ascertain whether or not the addition of such

evidence to the system would succeed in enriching the knowledge already yielded by the speech-band audio level evidence, to such a degree as to justify its inclusion

Many reliable vocal pitch estimation techniques exist in the literature (e g [116, 117]), which ostensibly could be used to realise this task However, the drawback is that they all assume pure speech signal input, i e a speech signal that is relatively free from destructive background noise This is certainly not how the audio content dealt with in these experiments could be described, where the audio tracks are characterised by a commentator vocal signal mixed in with the ambient noise of the game environment, which is typically dominated by spectator-generated noise Therefore, it is concluded that most of the established pitch tracking algorithms would encounter severe difficulty in providing for reliable tracking of the pitch of the commentator speech Hence, this represents the biggest obstacle to be overcome in investigating the exploitation of this characteristic

One proposed means of overcoming this problem concerns an attempt to extract the speech signal from the audio track prior to applying the pitch-tracking algorithm That is, given a FSV audio track, if the vocal signal may be cleanly isolated from the noisy ensemble, it is assumed that the pitch-tracking algorithm should be able to accurately extract the required information This task comes under the ambit of a topic known as ***audio-source separation*** and is typically regarded as a very challenging aspect of the audio-processing field However a preliminary hypothesis into how this objective may be realised is proposed as follows

If it is a case that the commentator vocal signal is a mono signal, centre panned in a stereo pair, then by exploiting both this and the assumed stereo asymmetry of the background noise ensemble, it should be possible to subtract the vocal signal from the original stereo signal, i e leaving a remainder signal, which corresponds purely to the background noise sources That is, since centre panning corresponds to an equal representation of a given audio source in both channels of a stereo signal, subtracting the left and right channels from each other yields a resultant monaural signal in which the source components that are centre panned in the stereo field are removed Thus, assuming that the commentator vocals are centre panned, the resultant signal will not feature this, i e it will only contain the sources that exhibit asymmetry in the stereo field Assuming that the background noise audio sources are characterised in this way, i e that they are mixed asymmetrically in the stereo field (a common trend for stereo FSV audio tracks), the resulting signal will be purely representative of these Hence, it is proposed

that the spectrogram envelopes be estimated for both this resultant signal (purely representative of the background noise) and the monaural equivalent of the original signal (containing the background noise plus the vocal signal) By determining the differences between the two envelopes, it is proposed that the frequency components that correspond exclusively to the vocals in the original audio signal may be established Given this knowledge, it is proposed that the original signal could be then frequency filtered such that only these components are retained Thus, the signal resulting from this process would be expected to be highly representative of the pure vocal source, i e equivalent to the original signal excluding those frequency components corresponding to the noise sources It is anticipated that a pitch-tracking algorithm would be able to handle such a signal more successfully

8.3.2. Improving Speed Performance

Given the results of the speed performance analysis (provided in *Appendix F*), it is concluded that the developed scheme is clearly not yet optimized for high-speed performance application Therefore, a significant future work task concerns an investigation into how the underlying processes may be accelerated, towards improving this attribute Potential avenues for this are described in *Appendix G*

8.3 3 Scalable Output Functionality

Given the developed scheme, the further critical features proposed, and the potential scheme acceleration avenues, a more functional-level future work task relates to the post-processing system issue of output content scalability Given an optimised FSV summary generated by the scheme developed, which is comprised of a set of detected narrative-critical events (SUEs), a user may desire to have the option to further scale back the amount of content presented according to his/her demands This corresponds to having some method that allows discernment of which of the detected narrative-critical events constituting the summary are most significant, and by the same token which may be more expendable If a figure of significance (FOS) could somehow be determined for each detected event, an event hierarchy could then be formed, upon which the scalability may be based Two proposed criteria that might provide favourable FOS determination follow

Recall that SUE-shots are indicated by excitation in a set of critical features It was observed that, compared to more trivial SUE episodes, following a SUE of major

significance the player celebrations tend to be increasingly sustained in response Correspondingly, the critical feature excitations tend to recur or be sustained for a longer amount of time Hence it is suggested that, given a detected SUE, the duration of sustained/recurring critical feature excitation may be linked to the event significance Therefore, if a method of quantifying this could be developed, it would constitute a reliable basis for FOS assignment

In [34] the authors describe an approach for event detection and summarization in an American football-video context, in which, given a set of detected events, a significance hierarchy is proposed based purely on the corresponding audio energy levels observed for each event retrieved That is, the authors argue that, although many audiovisual features indicate events and are exploited in doing so, the real *acid test* for relative event significance are the corresponding noise levels observed, which are primarily attributed to the reactions of the spectators and/or commentator To some extent this argument is verified by the exploratory SFV component coefficient analysis undertaken for this scheme in *Section 6 2 1* Therein, it was shown that for the training-corpus content, of the five vector component coefficients constituting the SFVs, the component pertaining to the audio speech-band energy level feature (i e VCC_2) was one of the most discriminatory in terms of SUE-shot discernment On this basis, and motivated by the arguments put forward in [34], it is proposed that the tracking of audio levels suggests another criterion for deriving FOS values for detected events

As mentioned, scalable functionality is an attractive aspect of a video summarisation application, the realization of which would greatly enhance the implementation of this developed scheme Given the two proposed criteria for FOS derivation, a potential future work task thus concerns the development of such in determining which of the two provide the best performance in summary downscaling, i e ascertaining which of the two implements best, the trade-off between discarding potential SUEs as detected and retaining those of most significance Also of interest would be an investigation into the performance rendered by a combination of these two approaches for the said task

8.4 Furthering The Overall Field

In completion of this thesis it is required to discuss potential future work in terms of how the overall field of sports-video analysis may be further progressed

8 4.1. Further Supergenres: Towards A Complete Solution

Clearly, in terms of moving closer towards finding a complete solution for the problem of sports-video summarization, the approach advocated by this author is that presented in *Section 1 4* This point is expounded below

As discussed in *Section 1 3 2 1*, it is desirable to move away from genre-specific solutions However, as explained in *Section 1 4*, the principle difficulty pertaining to the development of a genre-independent solution to sports-video summarization concerns the conflict that exists between the event concept definition, and the required provision for generic applicability That is, given the event detection-based summarization task, it is ultimately unfeasible to suggest that there exists a unique solution that will operate successfully across all genres of sports-video Conceding this, it was then proposed that the overall sports-video domain be segmented and analysed, not at the genre-level, but at a higher 'supergenre' level, throughout which the event concepts and the general aspects of the games might be said to be consistent (see *Section 1 4*) Proposed supergenres were listed in Table 1 1 It was anticipated that by grouping characteristically linked sports-genres together in this way, a unique summarization solution might be obtainable for each supergenre, which exhibits accuracy comparable with that of a genre-specific approach

Given the successful realization of the objective of this thesis (i e developing a summarization solution for the field-sports supergenre), it is argued that this constitutes significant evidence that testifies to the validity and effectiveness of this proposed approach Moreover, it is the opinion of this author that the favourable results obtained for this case study should serve to motivate further exploitation of this approach, i e in terms of developing solutions for other supergenres, e g ring-sports, motor-sports, court-sports, etc Given a suite of supergenre solutions, clearly the next aspect of future work should then involve comparing the individual methodologies, towards establishing any potential commonality across them That is, if the solutions of two or more supergenre schemes were sufficiently alike (e g they exploited similar features in a somewhat similar manner), then it should be investigated as to what extent the performance accuracy can be maintained while merging the two solutions into one solution, which could potentially operate reliably across all the member genres of the two supergenres combined Merging supergenre solutions in this manner corresponds to the process of moving up the sports-video analysis 'value chain', i e the more supergenres that can be combined, the higher up the value chain we get, and hence the

closer we get to realizing the ultimate solution of having a unique scheme that can be applied to any sports-video genre

8 4.2. Common Forum

It was concluded that the generic scheme proposed in this thesis yields performances that approach that of the ideal target described In doing so, it is illustrated what is achievable via a certain methodology, and on this basis it is argued that the conclusions drawn provide a significant contribution to the field of sports-video analysis However, in *Chapter 2* numerous other sports-video processing works are also described, which to varying degrees also profess encouraging results with respect to both their own particular objectives and domains That is, from that of the earliest works ([13] - 1995), to those most recently documented ([49] - 2004), there has been almost ten years of development in the field of sports-video processing, characterized by an abundance of proposed schemes However, for the reasons outlined earlier, the suitability of these schemes towards both a cross-comparison of results and/or a combination of methodologies, is somewhat lacking It is argued that, while having been shown to operate reasonably successfully in their own right, the overall practical impact of these schemes in the field has been weakened by this fact Therefore, it is the opinion of this author that in order for the sports-video processing field to be further progressed towards finding accurate and robust solutions to the various challenges presented, a common forum is required, within which the following should be established, (i) a common baseline dataset, (ii) a set of specific task objectives, and (iii) a standardized results format It is anticipated that, given such regularization, the sports-video analysis field in general would benefit overall An example of such a forum is that of the ***Text REtrieval Conference*** (**TREC**) [118], which is essentially a research support convention set up for the field of text retrieval Initiated in 1992, TREC dictates the infrastructure necessary for the large-scale evaluation of various text retrieval methodologies Significantly, what started out in 2001 as a TREC-sponsored video 'track' devoted to research in automatic segmentation/indexing and content-based retrieval of digital video, became an independent evaluation (i e **TRECVID**) [119] in 2003 To summarize its mission statement, the goal of TRECVID is to encourage research in multimedia information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results However, apart from a task concerning the recognition of sports-video content, none of

TRECVIDs specific tasks have so far related to the task of extracting semantics from sports-video content

8.4 3 Alternative & Emerging Topics

As can be seen from the account of the related work given in *Chapter 2*, the most common task undertaken in the field of sports-video processing corresponds to that addressed herein, i e the problem of automatic highlight detection towards game summarization/abstraction However, according to some recent literature on the subject, i e [120], [121], and [122], there seems to be evidence of some alternative sports-video related tasks currently being targeted in the field, e g tactics and player performance analysis, augmented reality presentation, and referee assistance

Tactics analysis involves recognition of the tactics that teams or individual players have used, while performance analysis is concerned with appraising the performance of a team or a player, e g through analyzing their motion and activity in games Both aspects exhibit commonality in that they correspond to the task of rendering statistical data from the underlying games It is a commonly held argument that broadcasters are interested in such results for presenting sports video with additional statistical information Furthermore, it is envisaged that both coaches and players would be interested in performance knowledge in particular, as a basis for improving their play for later games

The topic of augmented reality presentation concerns the development of methodologies for enhancing the viewing experience of sports-games to the viewer One such aspect of this topic concerns 3D reconstruction technology, with a view to providing the viewer with images of unfolding events from arbitrary perspectives Another facet involves the overlaying of illustrations onto the original video images, the purpose of which is to aid the viewer in understanding the events as they unfold One example of such is the recent trend of superimposing virtual lines onto a soccer field to illustrate the extent to which a certain player was on/off-side Another aspect that comes under the banner of augmented presentation (but arguably not under that of viewer enhancement!) is the development of technologies for the optimized placement of visual advertisements within sports-video This subject, known as commercial value enhancement, is currently emerging to be a hot-topic in the field of sports-video processing

Real-time referee/umpire assistant schemes are targeted towards aiding (or

replacing) the manual game-rule judgment calls that must be made in any sport, generally when difficulty is encountered in doing so Such systems may utilize either dedicated electrical sensors, or they may be based on more adaptable real-time video analysis systems For example, in the major tennis tournaments (e g Wimbledon), a video-based system known as ***hawk-eye*** [123] uses dedicated cameras to accurately track the players and the ball, and is used to ascertain exactly where the tennis ball lands with respect to the fixed court line position This system has been shown to operate successfully not only in tennis, but also in a cricket scenario as well However, in the multitude of sports genres, many decisions that need to be made are either subjective and/or do not correlate to a fixed line/boundary, e g the off-side rule in soccer video In such cases, more complex and powerful video solutions are thus required

One notable aspect from these emerging applications is that critical to future development seems to be the principle of exploiting new sources for increased data acquisition, e g the development of referee assistance technology by employing electrical sensors towards retrieving data This echoes the work proposed in [21], whereby in the context of developing a player-tracking system for soccer video, the notion of attaching signal emitting transponders to both the players and ball to render such data to a tracking system was proposed

8.5 Chapter Summary

At the outset of this chapter a brief thesis synopsis was presented Following this, based on the result evaluation performed in *Chapter 7*, an account of the conclusions drawn was presented Next, a comprehensive outline of potential future work aspects was presented, both in terms of further developing the scheme developed, and furthering the area of sports-video analysis in general, including a discussion on the alternative topics/applications emerging in the field

Appendix A

Shot-Boundary Detection

As introduced in *Section 3.3.4*, the camera shot, which corresponds to the video resulting from a continuous, unbroken recording by a single video camera [54], is the basic syntactical unit of a video sequence. This appendix introduces the topic of shot-boundary detection, which concerns the task of analysing each frame of a digital video sequence with a view to determining whether or not they represent shot transitions. Following this introduction, the shot-boundary detection algorithm used in this work is described in detail, along with an appraisal of its performance on the field-sports-video (FSV) data corpus.

A.1. Shot Transitions

Shots may be delimited by a variety of boundary transition types. The most basic of these are ***shot-cuts***, which are sudden shot transitions that occur abruptly between two neighbouring frames. Video effects processing provides for other shot transition types such as fades, dissolves, wipes, morphs, etc. A ***fade*** is a gradual increase or decrease in brightness, i.e. either to, or from, a black frame. ***Dissolves*** are similar to fades except that they involve a temporary crossover of two adjacent shots (i.e. during the short intersection period, the images of the leading shot become gradually darker, while those of the following shot become gradually brighter, until the latter completely replaces the former). In shot-wipes, a moving edge frontier (or that of some geometric shape) is employed to erode the images of the current shot while revealing those of the next shot. Shot-morphing graphics correspond to the form-altering process where two (sets of) images are merged, transforming one into the other.

A.2. Approaches To Shot Boundary Detection

As explained in *Section 3.3.4*, at standard video framerates, (e.g. 25fps, 30fps), the images within a particular shot differ only very slightly from frame-to-frame. Hence, most approaches to the task of shot-cut detection are concerned with the quantification of the dissimilarity of consecutive video frames, in ascertaining whether or not they belong to a common shot. If the decision process implementing this procedure determines that two subsequent frames are sufficiently dissimilar, it is then concluded that a shot transition (cut) must have occurred, and a shot boundary is declared on the latter frame.

In [124] the authors outline how a thresholded sum-of-absolute-difference metric, operating on decompressed video pixel data, may be employed to achieve accurate shot-cut detection for generic video. Another pixel-level method, proposed in [125], involves the process of edge detection, whereby the intensity and position of edges in consecutive frames is used as an information source upon which a frame dissimilarity metric is built. A more computationally moderate approach is to use colour histograms to facilitate the generation of a frame dissimilarity metric. An example of such an approach may be found in [126]. Furthermore, in [127] the authors argue that the ***cosine dissimilarity measure*** (**CSM**) yields the best results for detecting histogram dissimilarity fluctuations. More advanced approaches attempt shot-cut detection in the compressed domain. For example in [128] the authors propose a link between the number of intra-coded macroblocks used to encode a P-frame, and its probability of representing a shot-cut. However, most significantly, in [129], the authors compare all of the abovementioned methods using an extensive and diverse video corpus. It is their conclusion that the most reliable methods are those rooted in the histogram-based techniques.

A.3. Cut_Detect

In 1999, a shot-cut detection tool, ***Cut_detect***, was designed, implemented, and tested with success on a diverse television broadcast video corpus by research colleagues O'Toole, Smeaton, Murphy, and Marlow [79]. It was proposed that the techniques underpinning this scheme be recycled towards facilitating reliable shot-cut detection for the work herein.

A 3 1 Description Of Cut_Detect

Cut_detect is a shot-cut detection algorithm for MPEG-1 video files The approach is based on the quantification of frame-to-frame dissimilarity, which is implemented via the generation of metrics relating to both histograms and statistical moments, for the colour components of each video image Based on these descriptors, the algorithm invokes a threefold thresholding mechanism to quantify the significance of dissimilarity between frames, towards the detection of abrupt shot cuts in the video

Initially the algorithm is charged with the task of the manipulation and decompression of an input MPEG-1 sequence, so that the generation of the frame descriptors may be facilitated To implement this procedure, the algorithm makes use of XIL library functionality [130] where appropriate Given an appropriate level of decompression, the first detection mechanism is invoked, which involves the analysis of colour histograms Specifically, three 64-bin histograms for each YC_bC_r component are generated for each video frame For two consecutive frames, their corresponding histograms are compared using the CSM If the measure indicates a sufficiently high degree of dissimilarity, then the latter frame is logged as a cut Failing this a shot-cut may be yet detected by the second mechanism, which concerns the analysis of statistical colour moments In this case, three colour moments for each YC_bC_r component are generated for each frame These relate to mean intensity, intensity variance, and intensity skew (i e nine discriminatory values in total for each frame - three moments for each of the three colour components) For two consecutive frames, their corresponding values are used in a difference equation, which calculates a resultant dissimilarity distance value Similarly, if this measure is sufficiently high, then the latter frame is recorded as a cut

Finally, if it turns out that the characteristics of a given a shot-cut are too subtle to trigger detection by either of the first two methods operating in isolation, it may yet be detected by a safety-net mechanism, which involves the combination of the first and second methods Explicitly, this final concept stipulates that while the dissimilarity measures generated in either case may be deemed insufficiently high to activate outright detection, if they however are both deemed moderately high at the same time, a shot-cut is declared

A.3.2 Implementation of Cut_detect

This section provides a step-by-step description of how ***Cut_detect*** performs its shot-

cut detection task, as outlined above As described above, to realize the MPEG decompression and video image manipulation tasks, the algorithm utilizes XIL library functionality where appropriate Furthermore, given a decompressed video sequence, the XIL library provides additional functionality for the efficient generation of video-frame colour histograms Again, descriptions of the procedures involved in realizing these XIL-related tasks may be found in [130]

In implementing the histogram-based cut-detection mechanism, three 64-bin colour histograms (one for each YC_bC_r component) are generated for each video image These are then concatenated to form an overall 192-element frame-representative vector The contrast between two such vectors is then quantified using the dissimilarity analogue of the CSM (DACSM), which is a standard dissimilarity metric for contrasting two vectors The DACSM is defined in **(A.1)**, where M and N are two vectors for comparison The formula returns a value between 0 0 and 1 0 according to their dissimilarity

$$DACSM = 1 - \frac{M \bullet N}{\|M\| \, \|N\|} \qquad \textbf{(A 1)}$$

In implementing the moments-based cut-detection mechanism, the three colour components (YC_bC_r) are analysed for each video image The mean (μ), variance (σ), and skew (η) are then calculated for each, using **(A 2)**, **(A 3)**, and **(A 4)**, respectively

$$\mu = \frac{\sum_{i=1}^{P} X_i}{P} \qquad \textbf{(A 2)}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{P} (X_i - \mu)^2}{P}} \qquad \textbf{(A 3)}$$

$$\eta = \sqrt[3]{\frac{\sum_{i=1}^{P} (X_i - \mu)^3}{P}} \qquad \textbf{(A.4)}$$

The contrast between two given frames is then quantified by using the nine (3 x YC_bC_r colour components) moments in a difference equation, which calculates a resultant dissimilarity distance measure (Δ) That is, for two video frames m and n, Δ is calculated as shown in **(A 5)** This difference equation calculates the superposition of the absolute

values of the frame-to-frame disparities, as described by the nine defined moments Hence, the value **Δ** should represent overall dissimilarity with high values, which for abrupt shot- cuts, should be discernible amongst other video data

$$\begin{aligned}\Delta &= \left|\mu_Y^m - \mu_Y^n\right| + \left|\sigma_Y^m - \sigma_Y^n\right| + \left|\eta_Y^m - \eta_Y^n\right| \\ &+ \left|\mu_{C_b}^m - \mu_{C_b}^n\right| + \left|\sigma_{C_b}^m - \sigma_{C_b}^n\right| + \left|\eta_{C_b}^m - \eta_{C_b}^n\right| \\ &+ \left|\mu_{C_r}^m - \mu_{C_r}^n\right| + \left|\sigma_{C_r}^m - \sigma_{C_r}^n\right| + \left|\eta_{C_r}^m - \eta_{C_r}^n\right|\end{aligned} \quad \textbf{(A 5)}$$

As described, the third detection mechanism utilizes a mixture of the histogram and moment evidence in a combined approach, and stipulates that for a shot-cut to be declared, each dissimilarity measure must be at least moderately high in both mechamsms simultaneously

A.3 3 Thresholds & Performance Evaluation

For the purposes of evaluating ***Cut_detect***, its authors employed an eight-hour long test suite of broadcast TV video for analysis, which included a variety of programming, such as news, quizzes, dialogue, cookery, sport, gardening, and advertisements [79] Overall, the entire corpus consisted of 6159 shot transitions The locations of these were marked up as a ground truth for evaluation Of the transition types, 5380 (87%) were found to be shot-cuts

In its operation, each of the three detection mechanisms that comprise ***Cut-detect*** requires an optimum threshold be set On the basis of the thresholds, the dissimilarity measures of each mechamsm may be probed and peaks (shot transitions) detected In evaluating the histogram-based detection mechanism, the corresponding DACSM measures were generated for the test corpus, and these were compared with the ground truth It was subsequently determined that best retrieval statistics generated, i e **88%** precision and **85%** recall, were ascertained by using a DACSM threshold value of **0 040** [79] In evaluating the moments-based detection mechanism, again the corresponding frame-to-frame dissimilarity distance measures (**Δ**) were generated for the same corpus, which were then compared to the ground truth This time it was determined that best retrieval statistics generated, i e **87%** precision and **80%** recall, were ascertained by using a dissimilarity threshold value of **25 0** [79] By combining the two methods using such thresholds, it was determined that a unified approach yields

retrieval improvement towards **88%** precision and **87%** recall Finally, by adding the safety net mechanism, it was determined that for this corpus the retrieval was again improved towards **90%** precision and **92%** recall via safety net thresholds of **0 025** for **DACSM**, and **5 0** for **Δ** [79] These thresholds and their corresponding retrieval statistics for the test suite used are summarized in **Table A 1**

A 3 4 Field-Sports-Video Performance

While the abovementioned evaluation of ***Cut_detect*** as performed by its authors is comprehensive, it was required to appraise its performance explicitly on the video explicitly on the video content type specific to this thesis To this end, ***Cut_detect*** was executed on the content comprising the FSV training-corpus Precision and recall statistics were generated from this analysis and are presented in **Table A.2** As mentioned in *Section 4 5*, it was quantified that 95% of all training-corpus shot transitions were shot-cuts, hence the inclusion of shot-cut only retrieval statistics in this table From these results, and following a post-analysis investigation, it was noted that ***Cut_detect***, while less dependable in detecting transitions such as dissolves, provides for a very reliable performance in detecting the hard shot-cuts of this corpus

Table A 1 Thresholds and corresponding retrieval statistics for evaluation of Cut_detect

Mechanism	Threshold	Precision	Recall
DACSM	0 040	88%	85%
Δ	25 0	87%	80%
Combined (+ Safety Net)	0 040, 25 0 (0 025, 5 0)	88% (90%)	87% (92%)

Table A 2 Results generated by execution of Cut_detect on FSV training-corpus

Shot Transitions	Precision	Recall
All Transitions	98%	91%
Shot-Cuts	98%	97%

Appendix B

Tools For Signal-Level Feature Extraction

This appendix describes the procedures involved in the signal-level extraction of fundamental audiovisual evidence from MPEG-1 video bitstreams The features targeted are essential to the implementation of the extraction methodologies for the set of critical frame-level features described in *Section 5 1*, and they include Y-DCT coefficients, motion vectors, pixel luminance/hue, edge data, Hough line space data, and audio subband scalefactors However, prior to describing each feature extraction process, the MPEG-1 decompression software platform(s) upon which they are built are first introduced

B.1. MPEG Decompression Tools

Given an MPEG encoded video sequence, the first step involved in providing for any level of audiovisual feature extraction, requires that the bitstream be parsed and/or decoded appropriately Depending upon the nature of the video data to be extracted, it may be necessary that the content be fully decoded into its original uncompressed format, or alternatively, it may be sufficient to merely parse the bitstream down to an appropriate decoded level, such that the extraction of required compressed domain data may be facilitated There exists a variety of standard MPEG decompression software packages in the literature To accelerate the development of the required signal-level feature extraction software tools, it is proposed that several of the fundamental MPEG decoding components of these packages be recycled where appropriate To this end, the

following subsections provide an overview of the software packages specifically exploited, i e the Berkeley MPEG video decoder, the XIL image-processing library, and the Maplay MPEG audio decoder

B 1.1. Berkeley MPEG Decoder

The MPEG research branch of the Berkeley Multimedia Research Center at the University of California, Berkeley [131], have developed a variety of software packages for the encoding/decoding and analysis of MPEG-1 video bitstreams Many of these have been made publicly available, and they include an MPEG-1 video encoder, ***mpeg_encode*** [132], an MPEG-1 decoder, ***mpeg_play*** [133], and statistical analysers for MPEG bitstreams, i e ***mpeg_stat*** and ***mpeg_bits*** [134] Of primary interest for the work herein is the algorithm ***mpeg_play***, which was one of the earliest available software implementations of an MPEG-1 decoder The source code was written in C++, and while originally developed for UNIX, it has been designed so that it remains portable across multiple platforms The scheme is comprised of a library of software implementations, which perform the individual tasks necessary to fully decode a compressed MPEG-1 video bitstream to its image-pixel display level While the decoder was designed primarily for video playback applications, due to its modular design, many of the fundamental routines, such as those performing the bitstream parse, the GOP/frame/macroblock/block segmentation, etc, may be sequentially executed as standalone processes This allows for a gradual bitstream unraveling, down to an appropriately decoded level Hence this provides an already suitable platform for the development of compressed domain video feature extractors It should be noted that ***mpeg_play*** provides a software implementation for the decompression of MPEG-1 video streams only, i e the decoding of multiplexed audio is not supported Source code may be obtained via [135], and a complete description of the scheme including a performance analysis, may be located via [136]

B.1.2. XIL Image & Video Library

Developed by Sun Microsystems Inc [137] in 1992, the freely available image and video-processing library ***XIL*** [130], is a set of libraries written in C, which was designed initially for the Solaris operating system ***XIL*** supports a wide range of coding standards, e g JPEG, MPEG, H 261, CCITT faxG3/4, etc and the latest version of the software is ***XIL 1 3*** The ***XIL*** application programming interface (API) layer provides a wide range

of functionality that is fundamental to most image and video processing applications, especially those encompassing the issues of image and video (de)coding While the library is quite extensive, for the most part it does not provide functionality for the extraction of compressed domain MPEG data However, ***XIL*** does provide very efficient means for the decoding of MPEG video into the decompressed domain, i e it facilitates the task of pixel data extraction from compressed MPEG video Accordingly, the work herein is primarily concerned with only a small subset of its available functionality That is, it is proposed that the XIL routines that relate to the manipulation, decompression, and pixel attribute extraction from compressed MPEG-1 images may be exploited towards the development of pixel-level feature extraction tools A description of the functions that facilitate the implementation of these tasks, and the protocol concerning their use, may be found via [130]

B.1 3 Maplay

Developed by Tobias Bading of the University of Technology, Berlin, ***maplay*** (***M***PEG ***a***udio ***play***) is an MPEG-1 audio decoder designed primarily for the real-time playback of Layer-I/II MPEG audio streams, which was made publicly available in 1994 As a decoder, ***maplay*** can support all common bitrates (22 05kHz, 44 1kHz, 48kHz), and all standard audio types (mono, stereo, joint stereo and dual channel) The source code was written in C++ and was developed primarily for execution on UNIX based platforms Not unlike the Berkeley MPEG video decoder, ***maplay*** is comprised as a library of software implementations, which perform the individual tasks necessary to fully decode a compressed MPEG audio bitstream down to its audio sample level Again, while the decoder is primarily designed for playback applications, a benefit of its modular software design is that many of the fundamental routines, such as those performing the bitstream parse, the frame/granule/scalefactor segmentation processes, etc, may be executed sequentially as standalone processes This allows for a gradual bitstream unravel to an appropriately decoded level It is proposed that such routines may be thus recycled towards the generation of a suitable platform for the development of compressed domain audio feature extractors It should be noted that ***maplay*** provides a software implementation for the decompression of Layer-I/II MPEG audio streams only, i e the decoding of multiplexed video is not supported Source code may be obtained via [138]

B.1.4. Summary

It has been described how an MPEG audiovisual bitstream may be partially decoded for the extraction of compressed domain data, and/or fully decoded for the extraction of uncompressed data, by recycling the functionality offered by existing decompression tools Given this, the following sections provide a top-down description of the development of tools for the extraction of critical signal-level feature evidence from MPEG encoded video Where appropriate, it will be described how the proposed tools exploit and recycle some of the basic components of the abovementioned decompression schemes

B.2. DCT Coefficient Extraction

B.2.1. Y-DCT Coefficient Extraction

The XIL library API does not provide functionality for the extraction of DCT coefficient data from an encoded MPEG video bitstream Hence, it was required that an original software tool be designed to provide for the implementation of this task However, for efficiency, many of the standard software components of the Berkeley MPEG decoder were recycled in developing such Specifically, a tool called ***Y-DCT_extract*** was developed, which was implemented in the C programming language Given an MPEG encoded video bitstream, ***Y-DCT_extract*** uses some of the standard routines of ***mpeg_play*** to parse and decode the luminance (Y) component of the bitstream down as far as the block level At this point the tool extracts the Y-DCT coefficients for each block of each frame, in zigzag sequence Because it invokes only a partial bitstream decode, ***Y-DCT_extract*** provides a very rapid and efficient method for the extraction of such data from MPEG encoded video

B 2.2. Illustration

To illustrate the process of Y-DCT coefficient extraction and the associated knowledge extrapolation possible, consider the colour video image presented in **Fig B 1** Within this image, a particular region has been selected for illustration The selected region is of dimension (48x48) pixels and its corresponding YC_bC_r components are as shown[1] A

[1] The lower resolution of the colour difference components compared to the luminance component is due to the downsampling of the chrominance signals in the source compression

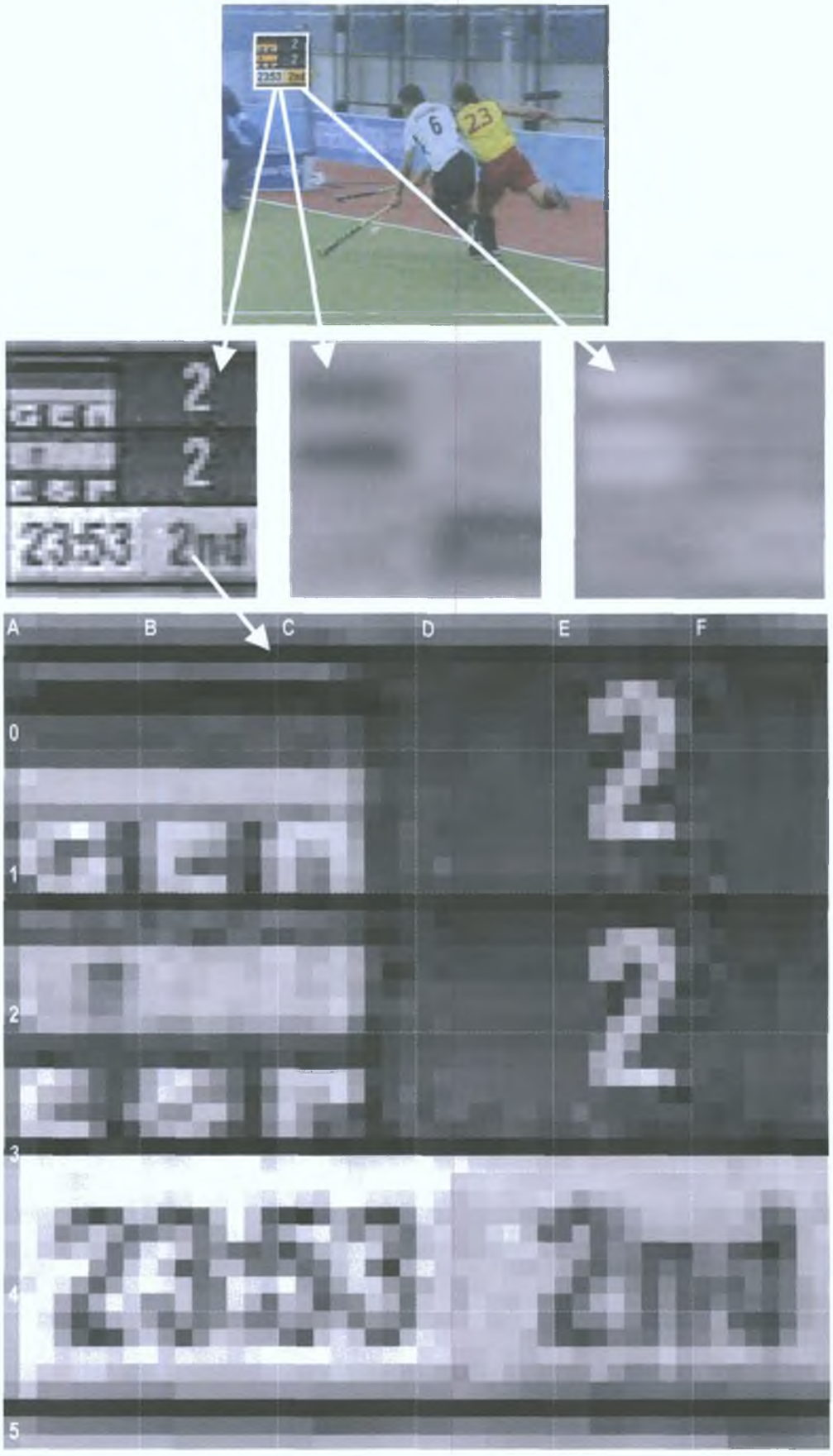


Fig. B.1. A colour video image; a selected region; YC_bC_r components of selected region; and an enlarged view of selected region luminance component.

further enlarged view of the Y component of the selected region is also presented Within this view the demarcations of its 36 (8x8) pixel blocks are outlined Using the tool ***Y-DCT_extract***, the DCT coefficients of the pixel blocks were extracted for this component **Fig B 2** presents the DC-DCT coefficients extracted, and additionally illustrated within this figure are the corresponding mean luminance intensities for each block

To specifically illustrate how the DC-DCT coefficients relate the to the low-level attribute of mean block intensity, consider those of block-A4 and block-F2 Block-A4 has a DC-DCT coefficient value of 1280, while that of block-F2 is 472 If the corresponding mean luminance intensities of these blocks are considered, it is evident that block-F2 is significantly darker than block-A4 It may be shown that this characteristic is consistent, i e for any given transformed pixel block, the higher the

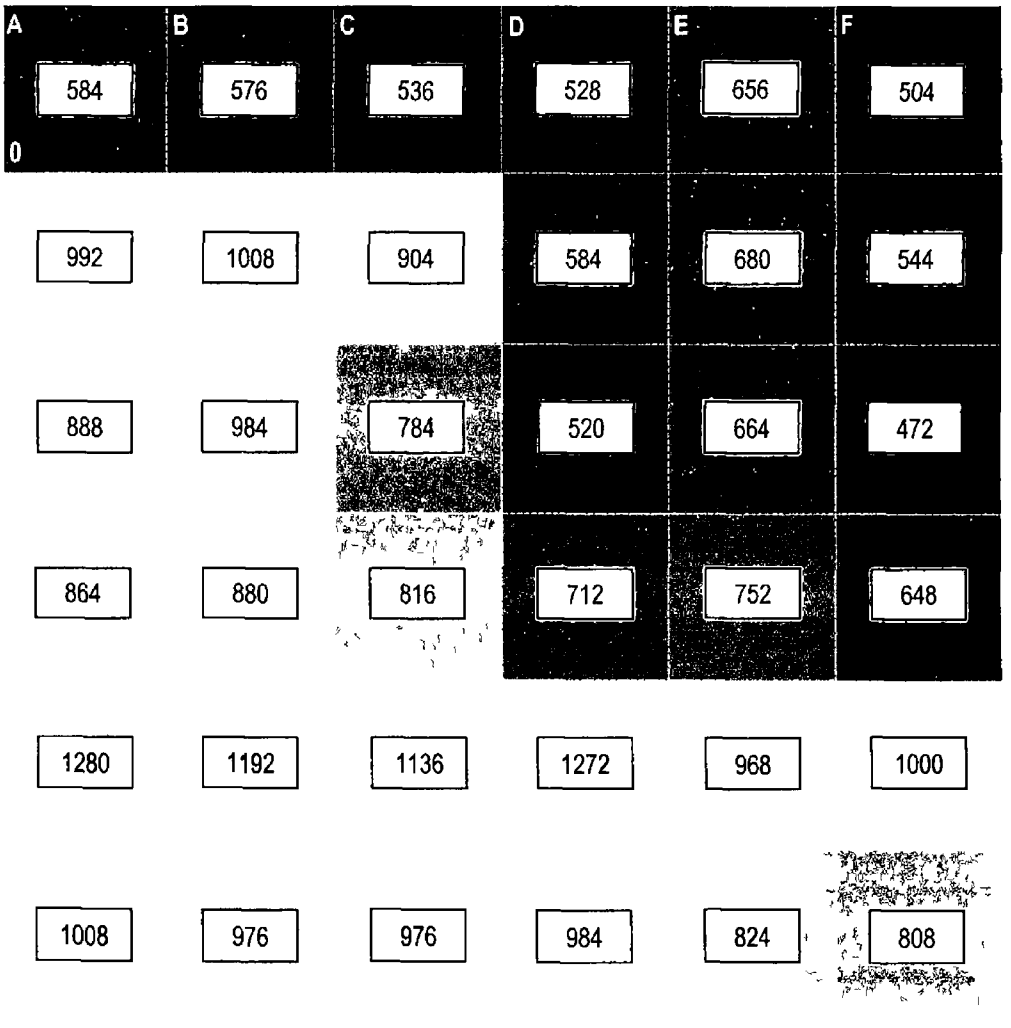


Fig B 2 DC-DCT coefficient values extracted by ***Y-DCT_extract*** for the 36 pixel blocks of the luminance component of the selected region of Fig B 1

representative DC-DCT coefficient value, the greater (brighter) the mean intensity level ***Y-DCT_extract*** also determined the AC-DCT coefficient count for each of the 36 blocks of the luminance component of the selected region **Table B 1** presents this output To illustrate how the AC-DCT coefficient count relates the to the low-level attribute of block intensity variance level, consider that of block-C4, and of block-D2 Block-C4 required 49 AC-DCT coefficients for its representation, whereas block-D2 required 11 If the corresponding intensity variance levels of these blocks are considered (see Fig B 1), it is evident that block-C4 is more intensely variant than block-D2 Again, this characteristic is consistent, i e for any given transformed pixel block, the higher the AC-DCT coefficient count, the greater the intensity variance level

Table B 1 AC-DCT coefficient count extracted by Y-DCT_extract for the 36 pixel blocks of the luminance component of the selected region of Fig B 1

Pixel Block	# Non-Zero AC-DCT Coefficients	Pixel Block	# Non-Zero AC-DCT Coefficients
A0	13	A3	44
B0	9	B3	42
C0	35	C3	37
D0	18	D3	25
E0	44	E3	42
F0	27	F3	23
A1	42	A4	44
B1	40	B4	41
C1	39	C4	49
D1	20	D4	45
E1	45	E4	41
F1	30	F4	34
A2	35	A5	38
B2	23	B5	32
C2	37	C5	46
D2	11	D5	33
E2	45	E5	32
F2	23	F5	27

B.3. Motion Vector Extraction

B.3.1. Motion Vector Extraction

Like the DCT coefficients, the XIL library API does not provide functionality for the

extraction of MV data from an encoded MPEG video bitstream Hence, it was again required that an original software tool be designed to provide for the implementation of this task Again, for efficiency reasons, many of the standard software components of the Berkeley MPEG decoder were recycled in the development of such Specifically, a tool called ***MV_extract*** was developed, which was implemented in the C programming language Given an MPEG encoded video bitstream, ***MV_extract*** reuses ***mpeg_play*** routines to parse the bitstream, isolate the P-frames, and decode the images to the macroblock level At this point the MVs for each macroblock of each P-frame are extracted Again, because it invokes only a partial bitstream decode, ***MV_extract*** provides a very rapid and efficient method for the extraction of MVs from MPEG encoded video

B.3 2 Illustration

To illustrate the process of MV extraction via ***MV_extract***, consider the two successive MPEG video images presented in **Fig B 3** The first image (reference frame) is an I-frame encoded image, while the second (predicted frame) is a P-frame encoded image The slight differences evident between the two images are due to the motion present within the temporal interval that separates them In encoding these two images, the ME algorithm was employed to estimate (in the luminance domain) the displacement of reference frame macroblocks in the predicted frame This estimation was then represented by a set of MVs, which were tagged onto the P-frame for reconstruction Using the tool ***MV_extract***, the P-frame bitstream was analysed and partially decoded such that its MVs were extracted To illustrate this output, a particular region has been selected within the reference frame The selected region is of size (96x96) pixels, and an enlarged view of the luminance component of this region is also presented Within this enlarged view, the demarcations of its 36 (16x16) macroblocks are illustrated Each one of these macroblocks has an associated type (I or P) and a corresponding set of MVs **Table B 2** presents this data as extracted for each macroblock of the selected region To illustrate how the MVs relate to reference frame macroblock displacement, consider those of macroblock-A4, macroblock-D2, and macroblock-F4 Macroblock-A4 is a predicted (p-) macroblock and has MVs of (48,-21) Therefore, in the predicted frame this reference frame macroblock is displaced by 48 pixels in the +x direction and 21 in the –y direction Macroblock-D2 is also a p-macroblock, however, it has MVs of (0,0) meaning its position has not been altered during the interval between the frames

Macroblock-F4 is an intra-encoded (i-) macroblock, which has no MVs. However, since this i-macroblock exhibits new data, it does not represent zero motion.

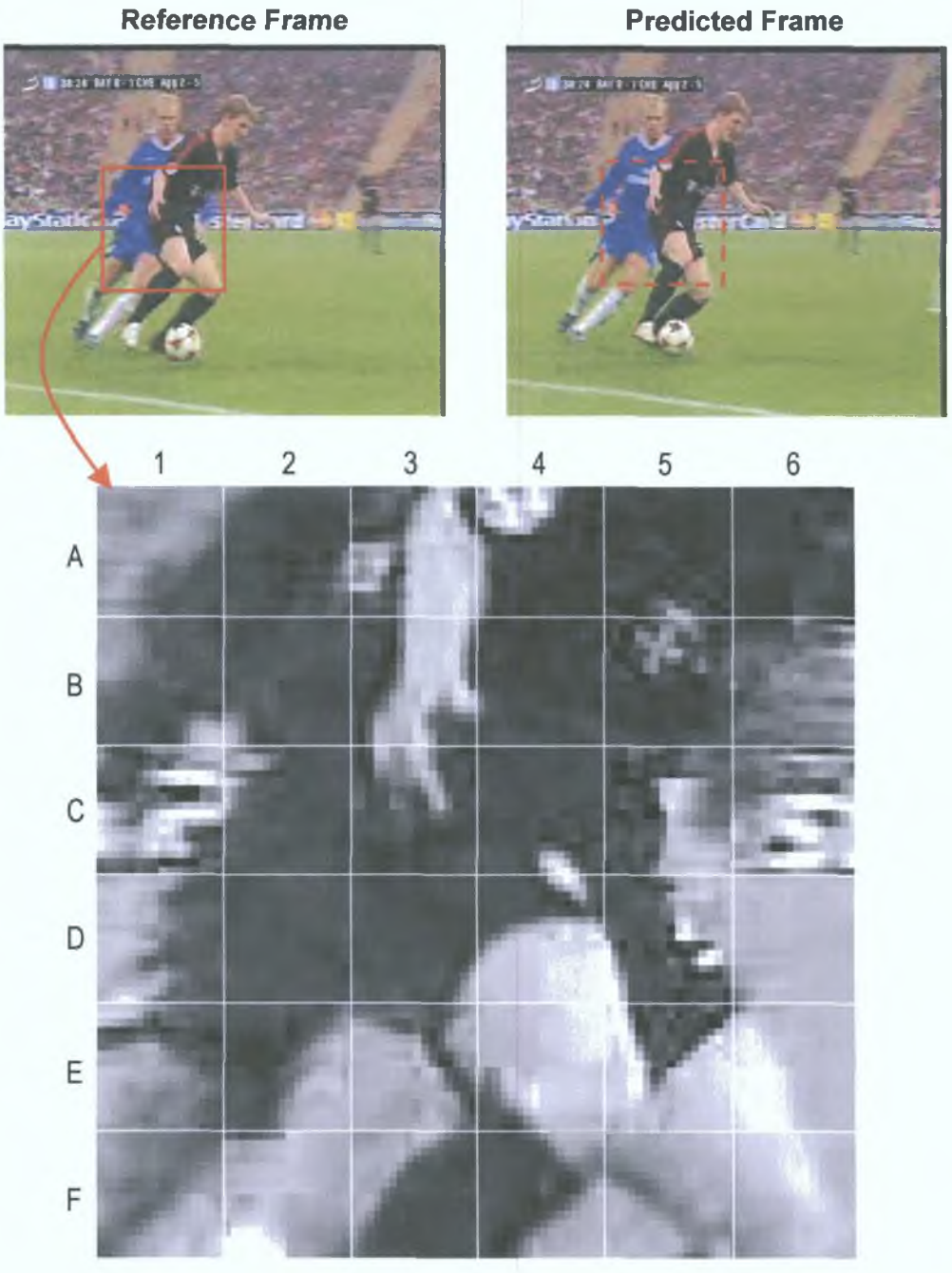


Fig. B.3. Two successive MPEG video images; a selected region; an enlarged view of luminance component of selected region.

Table B.2. MV_extract output for the 36 macroblocks of the selected region.

Macroblock	Type	MVs	Macroblock	Type	MVs
A1	I	(0,0)	D1	I	(0,0)
A2	P	(-49,-35)	D2	P	(0,0)
A3	P	(11,9)	D3	P	(3,5)
A4	P	(48,-21)	D4	P	(49,41)
A5	P	(11,9)	D5	P	(-51,29)
A6	P	(7,9)	D6	P	(32,8)
B1	I	(0,0)	E1	I	(0,0)
B2	I	(0,0)	E2	I	(0,0)
B3	P	(9,9)	E3	I	(0,0)
B4	I	(0,0)	E4	P	(1,13)
B5	I	(0,0)	E5	P	(1,3)
B6	I	(0,0)	E6	P	(24,0)
C1	P	(23,1)	F1	I	(0,0)
C2	I	(0,0)	F2	I	(0,0)
C3	P	(1,13)	F3	P	(-15,19)
C4	P	(1,9)	F4	I	(0,0)
C5	P	(-47,37)	F5	P	(17,1)
C6	P	(27,1)	F6	P	(19,-1)

B.4. Pixel Luminance Extraction

B.4.1. Luminance Extraction

The process of luminance data extraction uses the XIL library to provide for the required decompression of MPEG video images. Specifically, a tool called ***Y_extract*** was developed for the extraction of pixel luminance data, which was implemented in the C programming language. Given an MPEG encoded video frame, ***Y_extract*** decodes the image into the decompressed colour space (YC_bC_r), using XIL library API functionality. Following this, the pixel data is manipulated and the Y component of each pixel is extracted.

B.4.2. Illustration

To illustrate the process of pixel luminance extraction, consider the colour video image presented in **Fig B.4** (A). A zoomed-in region is also shown (B), and within this region a single pixel block (C) has been selected for illustration. Shown for this selected block are the demarcations of its individual pixels, and an image of its luminance component (D). Using the tool ***Y_extract***, the luminance values were extracted for each pixel of

this block, and the output is presented in **Table B.3**. To illustrate how the extracted data relates to actual luminance intensity, consider that of pixel-C4 and pixel-D7. Pixel-C4 has an extracted luminance intensity value of 184, while that of pixel-D7 is 16. If the corresponding luminance intensities of these pixels are considered (see Fig B.4), it is evident that pixel-D7 is significantly darker than pixel-C4. Again, this characteristic is consistent, i.e. for a given pixel, the higher the extracted luminance intensity value, the greater the pixel brightness.

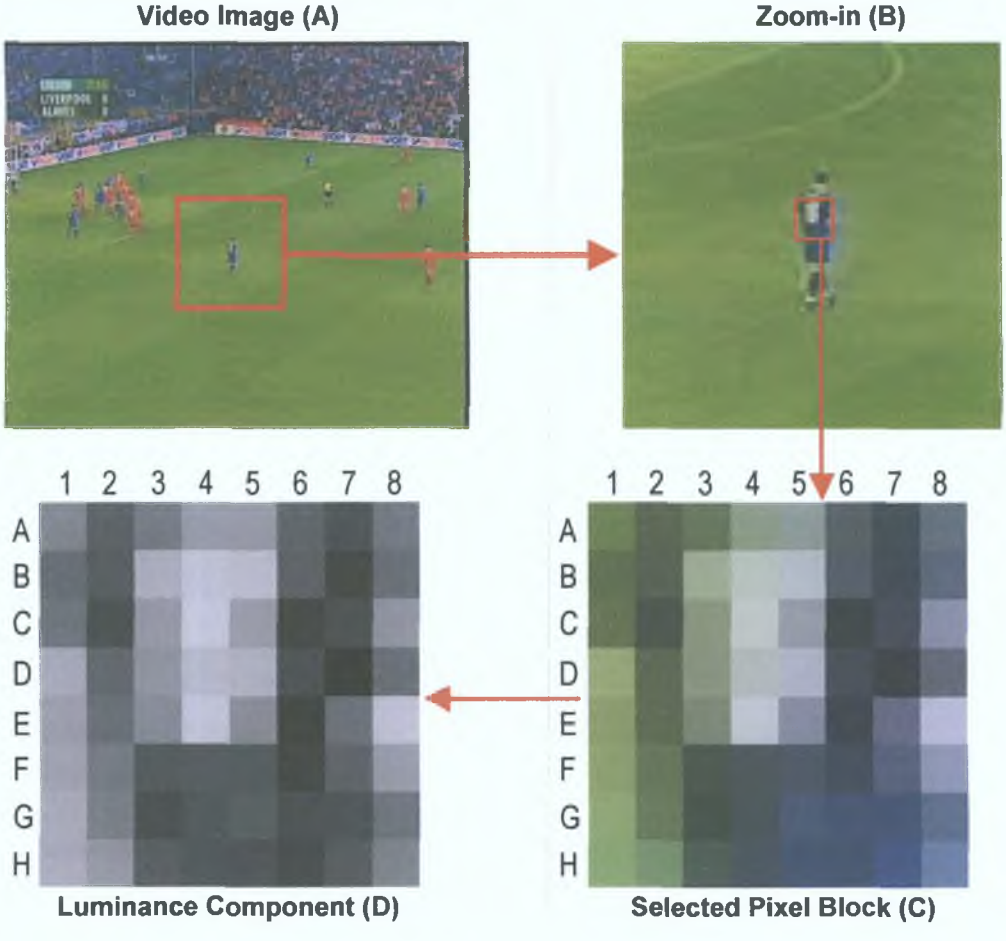


Fig. B.4. (A) A colour video image; (B) a zoomed-in view; (C) A selected pixel block; (D) the luminance component of selected block.

Table B 3 Y_Extract output for 64 pixels of selected block

Pixel	Luminance Intensity (Y)	Pixel	Luminance Intensity (Y)
A1	102	E1	130
A2	64	E2	81
A3	93	E3	109
A4	126	E4	181
A5	126	E5	114
A6	73	E6	25
A7	46	E7	86
A8	90	E8	126
B1	79	F1	134
B2	59	F2	94
B3	146	F3	54
B4	176	F4	61
B5	168	F5	58
B6	70	F6	29
B7	30	F7	70
B8	82	F8	126
C1	87	G1	143
C2	41	G2	99
C3	123	G3	40
C4	184	G4	53
C5	129	G5	66
C6	28	G6	45
C7	44	G7	47
C8	120	G8	88
D1	142	H1	141
D2	76	H2	122
D3	120	H3	62
D4	162	H4	51
D5	155	H5	49
D6	46	H6	37
D7	16	H7	60
D8	82	H8	104

B.5. Pixel Hue Extraction

B.5.1. Hue Extraction

To extract hue information from MPEG encoded video images it is required to first decompress the data into YC_bC_r pixel space A subsequent conversion into RGB components may be achieved via the inverse of the formulae given in **(3 2)** Then from these, equivalent HSV signals may be derived via the formulae of **(3 3)**

A software tool, ***H_extract***, was designed in the C programming language to implement these procedures. Specifically, given an MPEG encoded video image, ***H_extract*** utilizes appropriate XIL library functionality to decode a compressed image into YC_bC_r colour space. The procedures concerning the conversion of these signals into HSV space are then implemented, from which the pixel hue components are extracted.

B.5.2. Illustration

To illustrate pixel hue extraction consider the colour FSV video image presented in **Fig. B.5** (A). A zoomed-in region is also shown (B), and within this region a single pixel block (C) has been selected for illustration. For this block, the hue components of the image pixels were extracted using the tool ***H_extract***, and the output is presented in **Table B.4**. To illustrate how the extracted data relates to actual chrominance, consider that of pixel-H2 and pixel-H8. Pixel-H2 has an extracted hue position of 104°, while

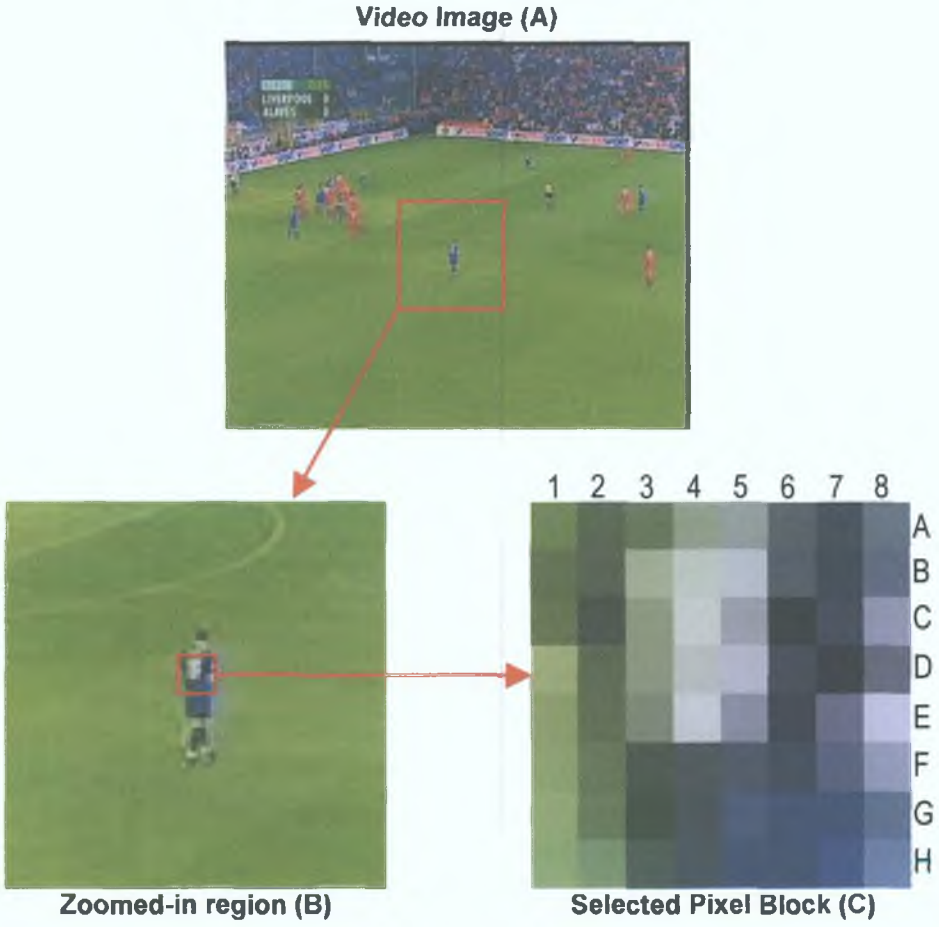


Fig. B.5. A colour video image; a zoomed-in view; a selected pixel block.

Table B 4 H_Extract output for 64 pixels of selected pixel block

Pixel	H	Pixel	H	Pixel	H	Pixel	H
A1	84°	C1	81°	E1	81°	G1	88°
A2	86°	C2	86°	E2	83°	G2	95°
A3	89°	C3	85°	E3	88°	G3	122°
A4	96°	C4	88°	E4	120°	G4	170°
A5	116°	C5	165°	E5	225°	G5	207°
A6	170°	C6	240°	E6	243°	G6	215°
A7	199°	C7	243°	E7	246°	G7	216°
A8	190°	C8	232°	E8	240°	G8	209°
B1	83°	D1	79°	F1	84°	H1	93°
B2	84°	D2	82°	F2	87°	H2	104°
B3	86°	D3	84°	F3	102°	H3	136°
B4	97°	D4	92°	F4	170°	H4	178°
B5	147°	D5	228°	F5	217°	H5	206°
B6	206°	D6	249°	F6	228°	H6	211°
B7	222°	D7	249°	F7	228°	H7	214°
B8	210°	D8	249°	F8	217°	H8	206°

that of pixel-H8 is 206° In considering these pixels in Fig B 5 (C), it is evident that pixel-H2 is exhibits a greenish tint, while pixel-H8 exhibits a bluish tint Considering the theoretical hue positions of their primary colours (Table 3 1), it is evident that the extracted hue data correlates well Furthermore, consider pixels-D6, -D7, and -D8 in Fig B 5 (C) These pixels exhibit a significant variance in colour shading for an ostensibly common tint However, their extracted hue values are equal, i e 249° This illustrates how the hue attribute reliably characterises chrominance, while transcending variances in intensity and saturation

B.6. Roberts Cross Edge Data Extraction

B 6.1 Roberts Edges

Given a 2-D binary image map, the Roberts Cross operator uses two (2x2) masks to determine the spatial gradient measurement in two distinct diagonal directions (i e the cross-differences) [76] In real world images, regions of intense spatial gradient typically correspond to object edges The two Roberts Cross masks are presented in **Fig B 6** They are designed to respond maximally to edges running at 45° to the pixel grid, i e one mask for each of the two perpendicular orientations

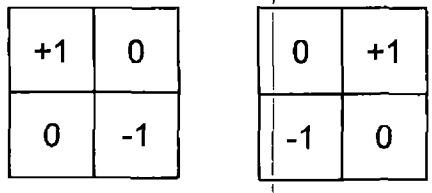


Fig B 6 Roberts cross operator masks

To extract edge information from an MPEG encoded video image it is required that it be first decompressed into its original YC_bC_r space Disregarding the sub-sampled chrominance components, the next step requires that the luminance component be thresholded, such that a binary image is produced, in which each pixel is represented either by a black (0) or white (1) level This may be achieved by applying an appropriate threshold ($\mathbf{T_{Bin}}$) to the extracted Y values For a given image, a typical choice for T_{Bin} is given in **(B 1)**, which effectively determines the median value between the brightest and darkest values of its luminance component

$$T_{Bin} = \frac{\max(Y) + \min(Y)}{2} \qquad \textbf{(B 1)}$$

Finally, the Roberts cross operators are applied to the pixels of the binary image map, yielding an output binary map, which exhibits the detected edges

To implement the procedures described above, a software utility called ***Edge_extract*** was developed in the C programming language Specifically, given an MPEG encoded video image, ***Edge_extract*** utilizes appropriate XIL library functionality to decode a compressed image into YC_bC_r colour space Following this it invokes both the luminance binarisation and Roberts Cross operations as outlined

B 6.2 Illustration

To illustrate the effectiveness of ***Edge_extract*** in the discernment of image edges, consider the colour video image presented in **Fig B 7** (A) Within this image a region has been selected for illustration (B), and ***Edge_extract*** was applied to this The Y component of this selected region was extracted and is presented in (C) Using T_{Bin} as defined in **(B 1)** the Y values were thresholded and the binary output yielded is shown in (D) By applying the Roberts cross operators to this binary image map, the pixels

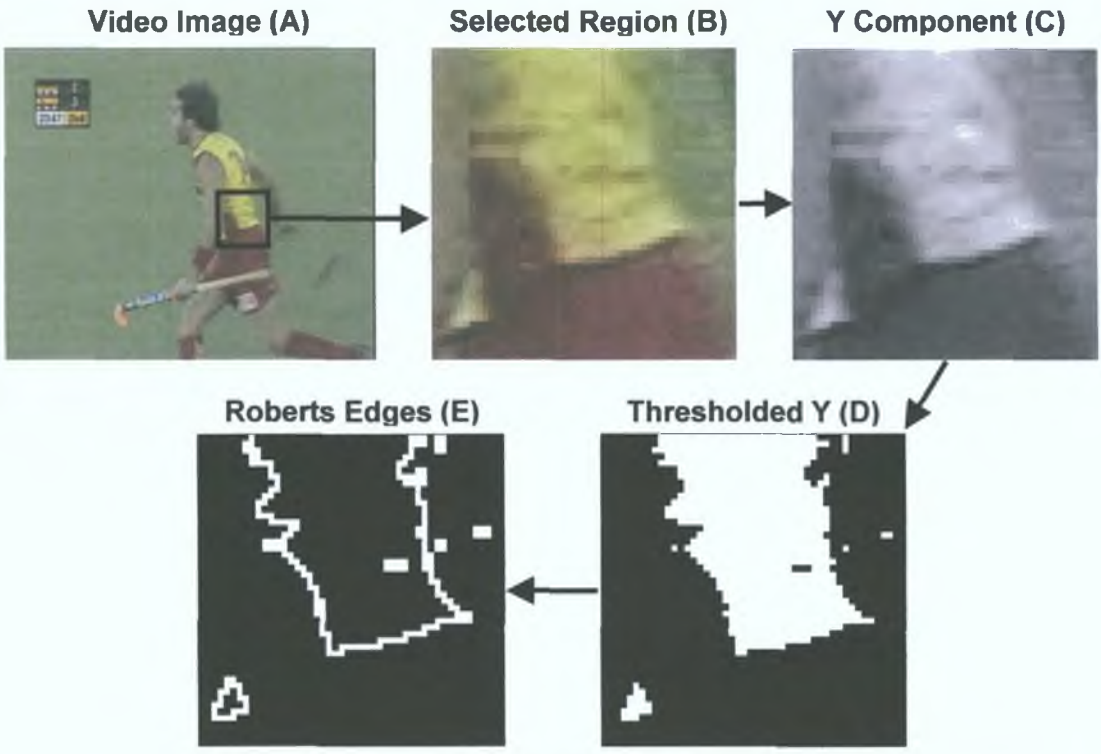


Fig. B.7. Colour video image (A); selected region (B); luminance component (C); thresholded luminance component (D); Roberts edges (E).

corresponding to intense spatial gradient were mapped to binary-1, while others were mapped to binary-0. That is, the edges of the binary map were discerned and isolated, as illustrated in (E).

B.7. Hough Line Space Data Extraction

B.7.1. Hough Line Space Data Extraction

The HLT assumes binary images as input. Furthermore, to eliminate large-scale line detection redundancy, it is also preferable to first apply edge detection to the binary images. The processing procedures involved in yielding such output from MPEG encoded video images was outlined in the previous section. By processing an image to this format, the HLT may be invoked in retrieving its linear content as follows.

In its parameterisation, the HLT utilizes the normal-form line representation, which has the format shown in **(B.2)**. This equation is the polar description of a line

passing through a point **(x,y)** that has a normal of length **d** from the origin, which itself makes an angle **θ** radians with the x-axis – see **Fig B 8**

$$d = xCos\theta + ySin\theta \qquad \textbf{(B 2)}$$

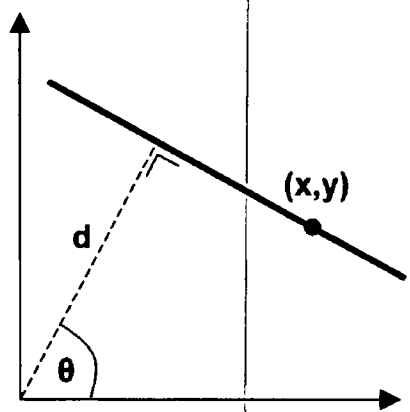


Fig B 8 Normal-form representation of a line

Hence, using this format to generate a parametric description of all the possible lines in a digital image space, the coordinates of the pixels serve as constants **(x,y)** in the above equation, which is then solved for variables **d** and **θ** Thus for an edge-detected binary image map, the HLT implements this process as follows

For each edge-pixel **(x,y)**, an equivalent value for **d** is calculated by iterating through a discrete set of possible line angles **θ**, i e **θ** ranges through a cycle of π radians for a chosen step-size - see **Fig B 9** The resultant values of **d** are then quantised using an *a priori* chosen quantisation (a process which is akin to setting the line thickness) From this each edge-pixel in Cartesian image space is mapped to its own **(d,θ)** relationship in Hough space - a relationship which turns out to be sinusoidal in nature Since the line angles are chosen discretely, and the corresponding value of **d** is quantised, the resulting **(d,θ)**

Hough space domain exhibits a latticed form, the resolution of which is determined by the chosen levels of discretisation and quantisation **Fig B 10** illustrates an edge-detected image and its corresponding HLT lattice for a high **(d,θ)** resolution

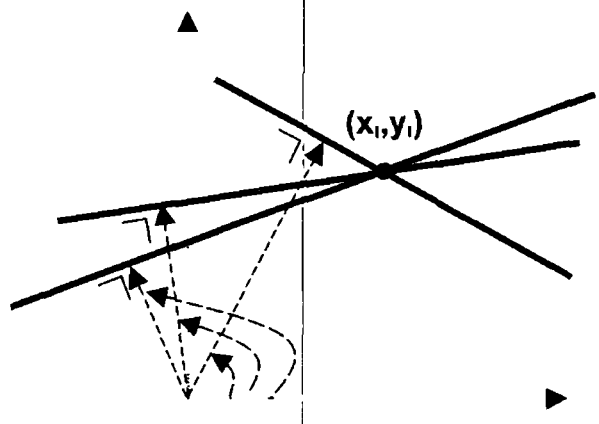


Fig B 9 Line angle iteration through a common point

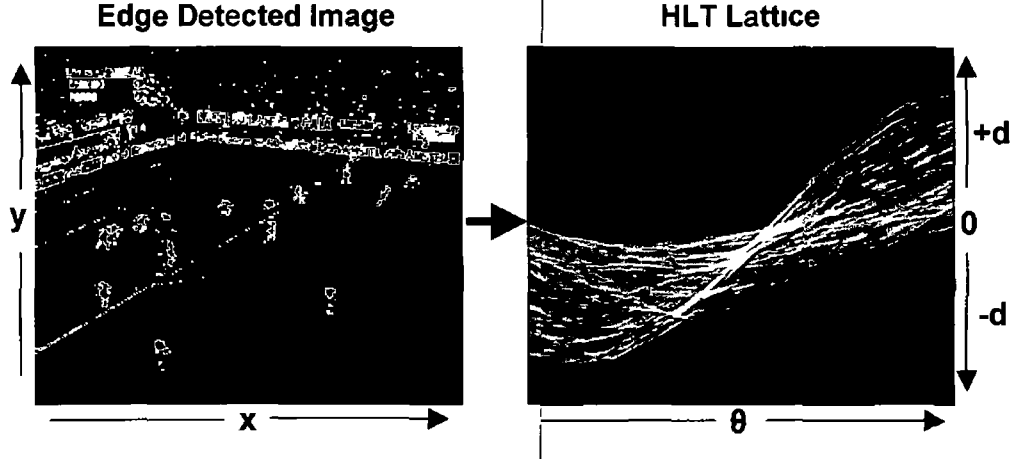


Fig B 10 Edge-detected image and its HLT lattice equivalent

Since collinear pixels exhibit common values of **d** and **θ**, following the iteration, edge-pixel points that are collinear in the Cartesian image space yield intersecting curves in the Hough Space Hence the position **d**, and orientation **θ**, of the most prominent lines in the image data may be discerned from the Hough space lattice, by simply locating the cells that exhibit the highest curve intersection tallies Thus for a given image, by implementing the abovementioned procedures, its Hough line space data may be retrieved, and hence knowledge of its linear content inferred

To implement the above procedures, a software-based utility called ***HLT_extract*** was implemented in the C programming language. Given an MPEG encoded frame, ***HLT_extract***, utilizes appropriate XIL library functionality to decode the image, and then the binary edge-detected equivalent is extracted as outlined in a previous section. Following this, the HLT is performed as described, facilitating the extraction of corresponding HLT lattice intersection tallies.

B.7.2. Illustration

To illustrate the extraction of Hough line space data using ***HLT_extract***, consider the colour video image presented in **Fig. B.11** (A). The edge-detected binary equivalent image is also presented (B). As described above, in applying the HLT to such an image, for each edge-detected pixel, θ (in radians) is iterated through an 180° cycle for a specific step-size, and in each case a corresponding value for **d** calculated. Hough space intersections are then tallied, indicating line occurrence probabilities. A standard step-size for θ is 1°, i.e. $\pi/180$ radians. However such processing is not practical for

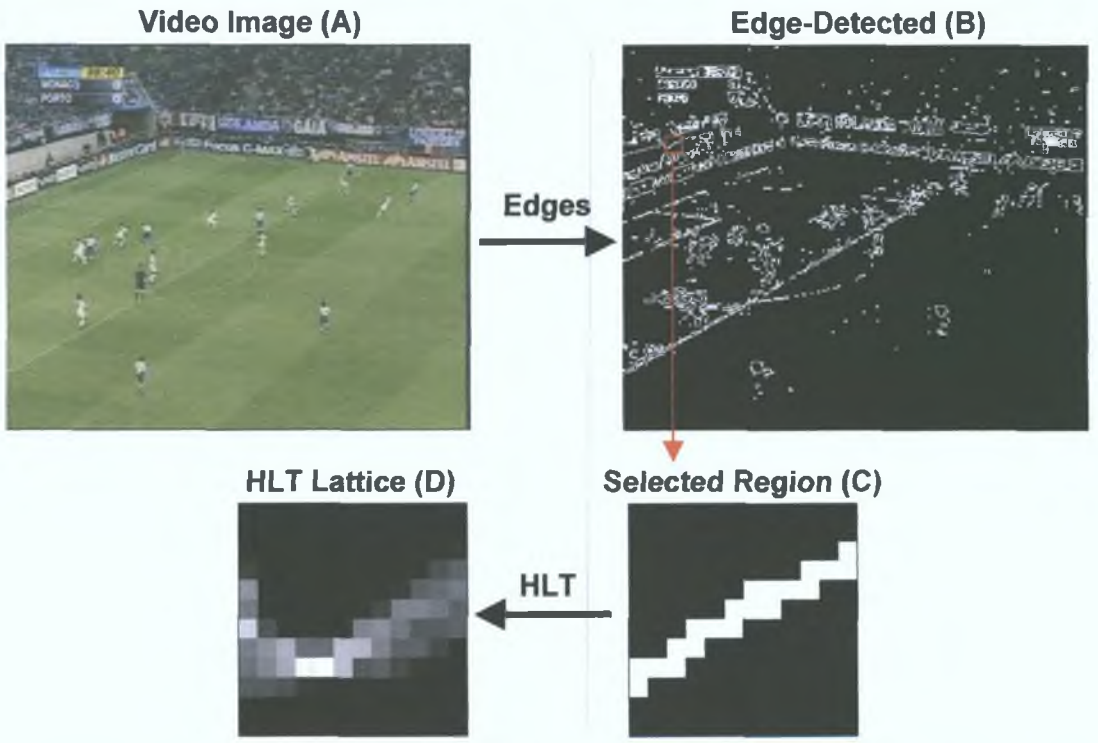


Fig. B.11. Video image (A); edge-detected equivalent (B); selected region (C); HLT lattice (D).

illustration, hence, for demonstration a more simplistic scenario is required To this end, within the edge-detected image a [12x12] pixel region has been selected, which exhibits a single line of well-defined orientation - see Fig B 11 (C) Using a step-size of 15° for the range [$-90° \leq \theta < 90°$], and 12 levels of **d**, the HLT was applied to this sample region The HLT accumulator lattice produced is illustrated in Fig B 11 (D) **Fig B 12** presents the actual tallies of the HLT lattice cells From this data it is evident that the highest tally (24) occurs at $\theta = 30°$ That is, for this sample region, the most prominent set of collinear points in this image correspond to an orientation of 30° Clearly this concurs well with that of the line displayed in the image

	90°	75°	60°	45°	30°	15°	0°	15°	30°	45°	60°	75°
+d	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	2
	3	0	0	0	0	0	0	0	0	1	5	7
	11	0	0	0	0	0	0	0	2	7	8	9
d=0	9	0	0	0	0	0	0	4	10	10	9	7
	18	5	0	0	0	0	10	11	9	6	6	7
	9	11	14	8	8	16	15	10	7	7	4	0
	8	10	13	22	24	16	6	7	4	1	0	0
	6	6	5	2	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	0	0
-d	0	0	0	0	0	0	0	0	0	0	0	0

Fig B 12 Intersection tallies of HLT lattice cells for selected region

B.8. Audio Subband Scalefactor Extraction

B 8.1 Scalefactor Extraction

To provide for the extraction of audio scalefactor data, it was again required that an original software tool be designed For efficiency, many of the standard software components of the MPEG audio decoder ***maplay*** were recycled in its development Specifically, a tool called ***Scf_extract*** was developed, which was implemented in the C programming language Given an MPEG encoded audio bitstream, ***Scf_extract*** uses some of the standard routines of ***maplay*** to parse and decode the bitstream down as far

as the subband level. At this point the scalefactors from each/any of the 32 subbands of each audio frame are extracted. Because it invokes only a partial bitstream decode, ***Scf_extract*** provides a very rapid and efficient method for the extraction of such from MPEG encoded audio.

B.8.2. Illustration

To demonstrate the process of scalefactor extraction and the knowledge they impart, a short MPEG encoded sample audio clip of duration of 5s (approx.) was utilized. For illustration purposes the segment was decompressed and the resulting audio waveform is presented in **Fig B.13** (A). Operating on the MPEG encoded bitstream of this clip, the scalefactors from each of the 32 subbands were extracted using the tool ***Scf_extract***. For comparison purposes these are plotted in **Fig. B.13** (B). In considering

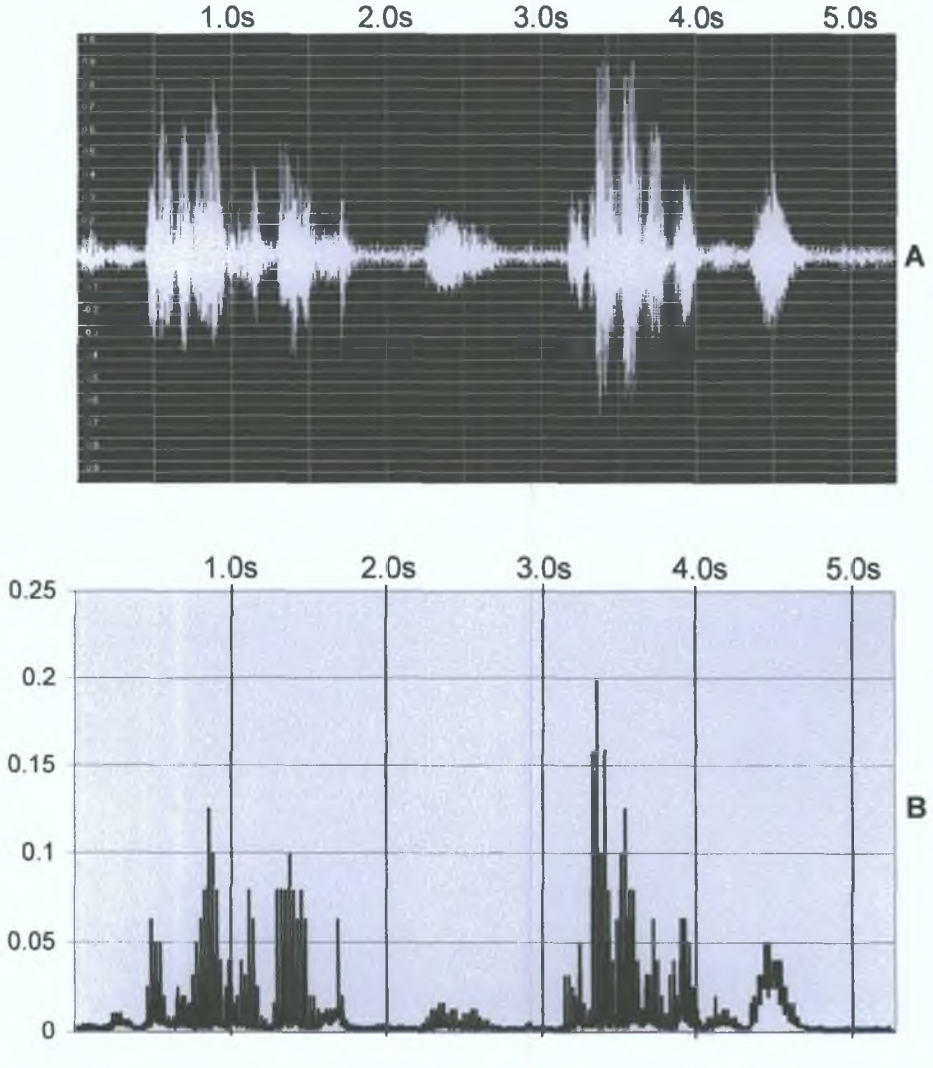


Fig. B.13. Audio clip waveform and a plot of its corresponding scalefactor data.

the two graphs, the correlation between waveform volume envelope and scalefactor intensity is evident For example, the waveform clip exhibits high energy between 3 and 4 seconds This is echoed by relatively high valued scalefactor intensity in the same interval Again, this characteristic is consistent, i e for any given audio segment i e , the more intense the audio energy level, the higher the representative scalefactor values for relative subbands

Appendix C

Pixel Erosion

This appendix illustrates the process of pixel erosion, which is employed in the filtering of the field pixel segmentation map as described in *Section 5 1 6 1*

Consider the sample binary pixel map shown in **Fig C 1**, where in the input pixel map, a mass grouping of positive pixels (binary-1) is adjacent to a small isolated cluster of such (shown in bold)

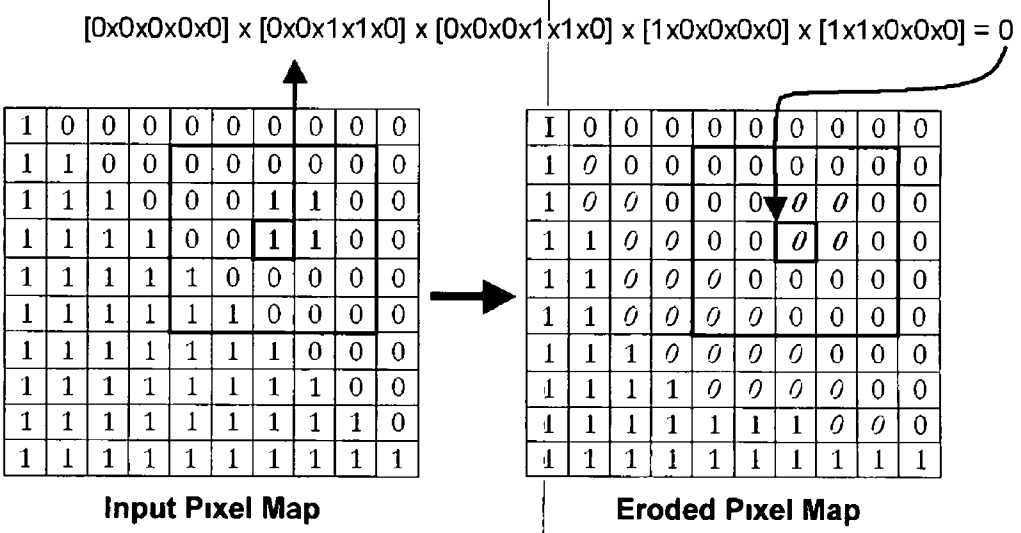


Fig C 1 Erosion filtering of a sample binary field-pixel candidate map

It is required that this input map be erosion filtered by the 2-D [5x5] pixel mapping defined in **(C 1)**, where b is the input pixel value (binary-0/1) and b' is the filtered

output pixel value, which corresponds to the combined product of its input value and that of all the other pixels contained within its surrounding [5x5] window

$$b'_{x\,y} = \prod_{i=x-2}^{x+2} \prod_{j=y-2}^{y+2} b_{i\,j} \qquad \textbf{(C 1)}$$

This operation has the effect of suppressing binary-1 pixels that are not wholly enclosed by binary-1 neighbours to the degree defined by the window size For instance, consider the input pixel highlighted in the figure In this case, since at least one of the windowed pixels is zero, their combined product, and hence the filtered equivalent of the current pixel equals zero In the figure, pixel bits that have changed from 1-to-0 are shown in italics in the eroded pixel map output It is evident that the erosion operation has the effect of shrinking the frontier of the mass group, while wholly obliterating the smaller isolated cluster In an object segmentation scenario, the idea is that, for a suitably sized window, the frontier shrinkage of the segmented objects should be negligible, while isolated falsely segmented pixels are suppressed

Appendix D

An Introduction To Support Vector Machines

Introduced by Boser *et al* in 1992 in [139], and based on Vapnik's earlier work on statistical learning theory [85], this appendix provides an overview introduction to Support Vector Machine (SVM) technology, which as explained in *Section 6.3.4*, represents the pattern classification approach employed in this work. Firstly, it is discussed how the field of generalization theory applies to the SVM solution. Following this, it is described how SVMs handle various scenarios, i.e. linear separable data, non-separable data, and ultimately the non-linear case. Finally, some of the issues critical to SVM implementation and performance are then outlined.

D.1. Generalisation Theory

Assume that each point in a given set of training data has the form $(\mathbf{x}_i,y_i)$, where $\mathbf{x} \in \mathbf{R}^n$, and $\mathbf{y}$ is the associated class. In binary classification $\mathbf{y}$ is either positive (1) or negative (-1)[1]. As explained earlier, the aim of the LM is to discern the target function, which is the relationship $\mathbf{x} \rightarrow \mathbf{y}$. However, ultimately the challenge is to select from the set of all possible hypotheses, the one that maximally reduces the risk of error in the classification of an unseen test point. Minimizing this risk of error will lead to a better generalization performance [96].

D 1 1. Bounding The Risk Of Error

Clearly the actual risk of error, Λ, cannot be determined since it requires knowledge of the unknown probability distribution from which the data are drawn However, it has been shown [95] that Λ is inherently bounded by an upper limit, which is both calculable and statistically reducible This process is called bounding the risk of error and is summarized as follows

As before, given a training set and a learned decision function, the empirical risk, λ, corresponds to the number of training set points that would be classified incorrectly by the decision function whence applied It may be shown [97] with certain probability that Λ is bounded by an upper limit, which corresponds to the empirical risk value, offset by a measure called the ***VC confidence*** That is, the inequality given in **(D 1)** holds with probability $1-\eta$, where **L** is the number of training points and **h** is a quantity known as the ***VC dimension*** of a set of functions [96]

$$\Lambda \leq \lambda + \sqrt{\frac{(h(\log(\frac{2L}{h})+1) - \log(\frac{\eta}{4}))}{L}} \qquad \textbf{(D 1)}$$

Hence, while Λ cannot be computed outright, the right-hand-side of this inequality may be Therefore by minimizing this, Λ is also minimized and hence the generalization performance should be enhanced

D 1.2. VC Confidence & VC Dimension

The value of the VC confidence term of the inequality given in **(D 1)** is dominated by the ratio **h/L**, i e the VC confidence varies almost as significantly as **h/L** varies Therefore, to ensure a lower overall risk of error limit, it is clearly desirable to maintain a low VC dimension value, **h** Given a set of functions, their VC dimension equates to the maximum number of training points that can be arbitrarily labeled (***shattered***) by that set of functions [96] For example the VC dimension of oriented lines in $\mathbf{R}^2$ equals 3 as illustrated in **Fig D 1** [96] Clearly a set of functions with infinite VC dimension can learn any set of training points correctly Therefore **h** may be viewed as an explicit quantification of LM capacity

[1] The value –1 is used to represent the false class rather than the value 0 such that later formulae are simplified [107]

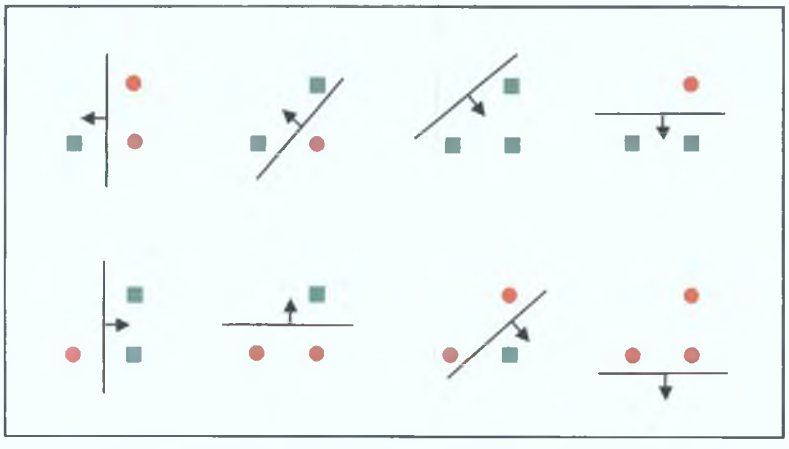


Fig. D.1. All 8 (2^3) possible binary labelings of 3 points in $\mathbf{R}^2$, with the orientated lines that correctly label them. [96].

D.1.3. VC Dimension & The Margin

It may be shown [96] that the above phenomenon may be generalized to Euclidean spaces in any dimension (note: in moving from R^2 to R^n the separating linear functions correspond to orientated ***hyperplanes***), i.e. the VC dimension of a set of orientated hyperplanes in $\mathbf{R}^n$ equals **n+1**. Such a hyperplane, **H**, may be described as shown in **(D.2)**, where $\boldsymbol{w}$ is the normal to **H**, and $\boldsymbol{x}$ is any vector lying on **H**.

$$w \cdot x + b = 0 \qquad \textbf{(D.2)}$$

Consider two further orientated hyperplanes, **H1** and **H2**, which are parallel to **H**, and lie on the decision boundary. These may be described as shown in **(D.3)**.

$$\begin{aligned} w \cdot x_i + b &\geq +1 \quad for\ y_i = 1 \\ w \cdot x_i + b &\leq -1 \quad for\ y_i = -1 \end{aligned} \qquad \textbf{(D.3)}$$

The inequalities in **(D.3)** may be amalgamated into one as shown in **(D.4)**.

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad \forall i \qquad \textbf{(D.4)}$$

Based on **(D.4)**, the equation given in **(D.5)** is true for any point that actually lies on either of the hyperplanes **H1** or **H2**, with the position of these hyperplanes with respect to **H** illustrated in **Fig. D.2** [83].

$$y_i(w \cdot x_i + b) - 1 = 0 \qquad \textbf{(D.5)}$$

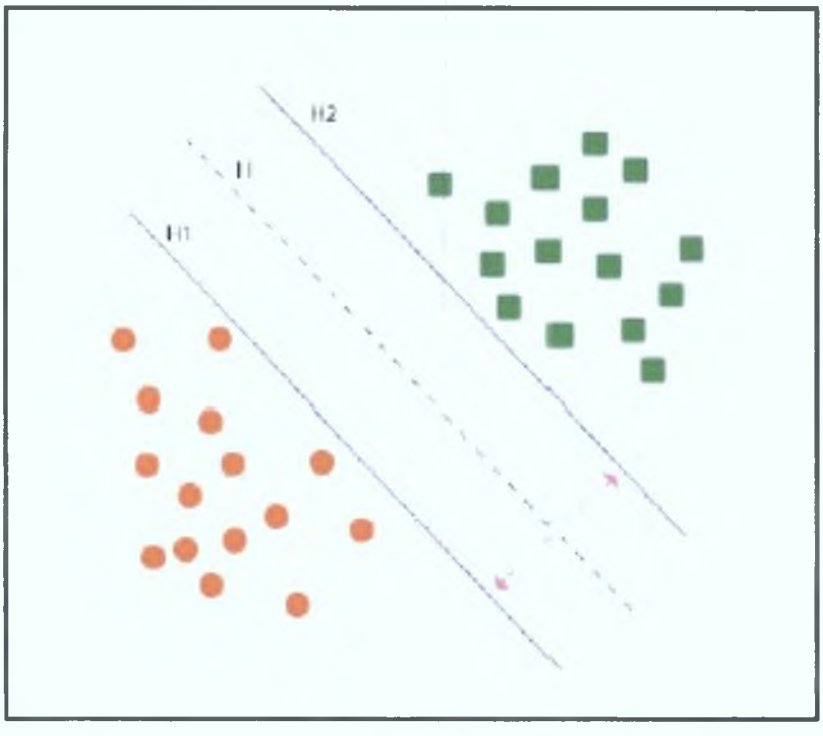


Fig. D.2. Orientated hyperplane H, with two further hyperplanes H1 & H2 lying on the decision boundary. Distance between H1 & H2 is called the margin [83].

It may be shown [96] that the distance between H1 and H2, i.e. the ***margin*** (**M**), may be calculated as in **(D.6)**, and that the hypothesis space, which represents the set of possible decision functions, is the set of functions given in **(D.7)**.

$$M = \frac{2}{\|w\|} \tag{D.6}$$

$$f(x) = \text{sgn}\,(w \cdot x + b) \tag{D.7}$$

Furthermore, in [140] it is shown that, assuming $\|\boldsymbol{w}\| \leq$ some value **A** and that the training points lie in an N-dimensional space completely within a sphere of radius **R**, then this set of functions has a VC dimension that satisfies the bound given in **(D.8)**.

$$h \leq \min(R^2 A^2, N) + 1 \tag{D.8}$$

However, given the margin as calculated in **(D.6)**, the term $\mathbf{R^2A^2}$ may be viewed as a function of the ratio between (i) the radius of a ball that contains all of the data, and (ii) the margin – see **Fig. D.3** [83]. That is, the bound on the VC dimension is proportional to $\mathbf{R^2A^2}$, where **R** is the radius of the smallest ball containing all of the data and $\|\boldsymbol{w}\| \leq$

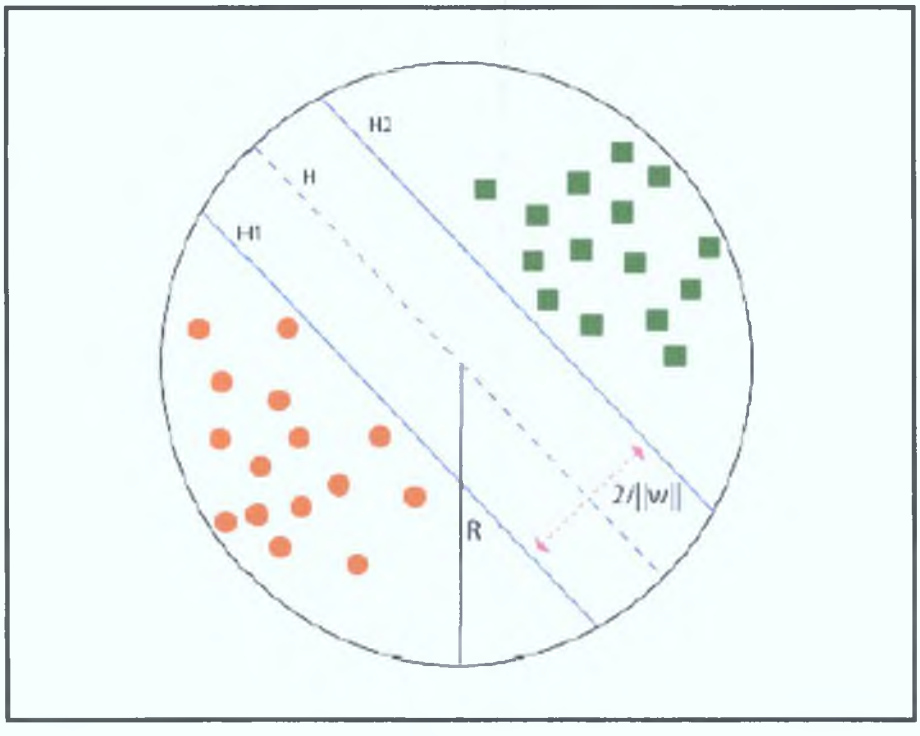


Fig. D.3. Margin between H1 & H2 is given by 2/||*w*||. R is the radius of the smallest ball containing all of the data [83].

A. Therefore, the larger the margin in relation to the radius, the tighter the bound will be. Hence, maximizing the margin will minimize the VC dimension. So, the separating hyperplane that gives the maximum distance between **H1** & **H2** will give a lower bound on the VC dimension, and as outlined above, this will yield improved generalization performance. This is the basis for the SVM approach, i.e. attempting to find the separating hyperplane that gives the maximal margin. A geometrical interpretation of why a wider margin will reduce the risk of error is given in **Fig. D.4** [83]. In this example, two separating hyperplanes correctly classify the same training set. However, scenario (b) uses a wider margin than scenario (a), and from this illustration it may be intuitively observed why this hyperplane would yield a lower risk of error for unseen data not within the training set.

D.1.4. The Structural Risk Minimization Approach

As mentioned in *Section 6.1.5.3*, the structural risk minimization (SRM) induction principle, proposed by Vapnik [85], is a methodology for controlling the capacity of a learning machine at the same time as minimizing the empirical risk. As described,

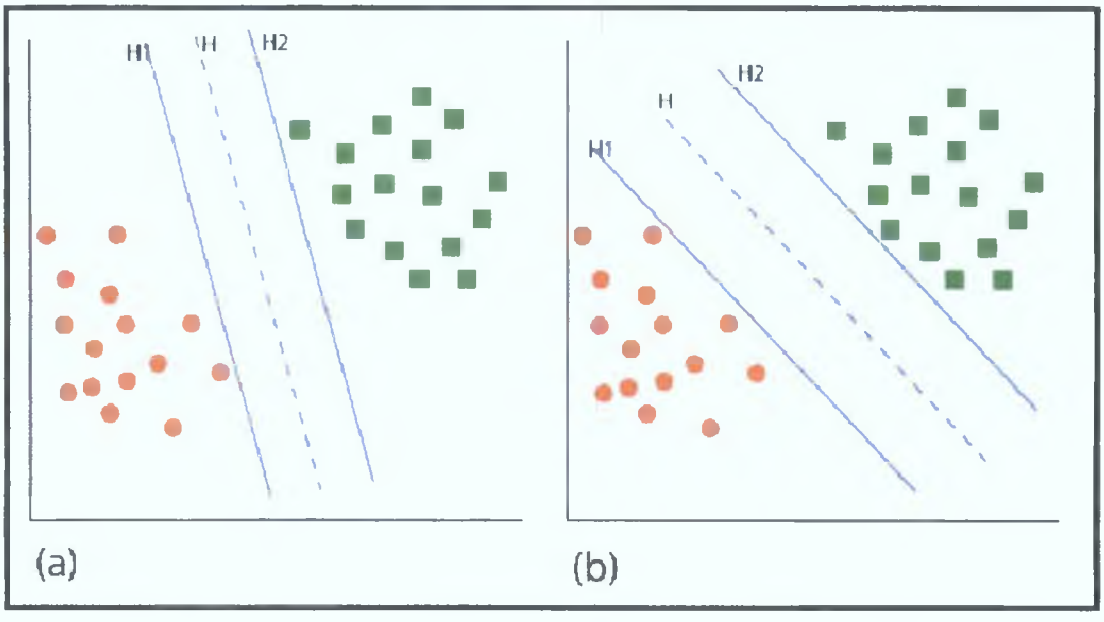


Fig. D.4. Two separating hyperplanes that correctly classify the same training set, but with varying degrees of risk of error [83].

controlling the capacity involves controlling the VC dimension (**h**), while minimizing the empirical risk involves minimizing the number of training errors. Since **h** is an integer, it cannot be varied smoothly, so a nested structure of hypothesis spaces is introduced, where each hypothesis space has a lower VC dimension than the previous one.

Although the VC dimension for a set of functions cannot always be calculated, it is possible to calculate a bound on the VC dimension. Since the empirical risk may be calculated based on the number of training errors, the structural risk minimization strategy involves searching through the structure of hypothesis spaces and choosing the one with a low capacity and that also has a low empirical risk. It will be shown in the following sections how SVMs implement this approach.

D.2. SVMs For Linear, Separable Data

D.2.1. Training A Support Vector Machine

It was described above how the separating hyperplane that yields the widest margin will reduce the bound on the risk of error in the test phase, i.e. providing for good generalization performance. Since the margin may be calculated as in **(D.6)**, a minimization of the value $\frac{1}{2}\|w\|^2$ is performed [96], which corresponds to a

maximization of its width In [97] it is shown how this can be posed as an optimisation problem to be solved using Lagrange Multipliers [141], where the objective function to be optimised, i e $\frac{1}{2}\|w\|^2$, is subject to the set of constraints given in **(D 4)** The details of this optimisation problem are outside the scope of this thesis However, an important point to note is that the problem may be formulated such that the input data does not appear directly in the expression, but rather as a dot product between training points It will be explained how this characteristic is exploited when moving to a non-linear scenario Furthermore, it may be shown [96] that once finalised, the solution is given in the form of, and is fully represented by, a typically minute subset of the training examples called ***support vectors*** The support vectors are the training points that are found to lie closest to the decision boundary, i e on either of the hyperplanes H1 and H2 In fact, all other training points end up having no further effect on the solution This is a beneficial characteristic of SVMs that is known as ***sparseness***, i e the final solution found is dictated by a subset of the training data In fact, if all other training points were removed, or were moved around so as not to cross H1 or H2, then the same separating hyperplane would be found This means that adding a larger number of really discriminating training points is unlikely to be of any benefit when training an SVM, although it is not always possible to know in advance which training points will lie far from the decision surface

D 2.2. Karush-Kuhn-Tucker Conditions

Certain conditions, known as the ***Karush-Kuhn-Tucker Conditions*** (**KKTC**), play a central role in both the theory and practice of any constrained optimization problem [96], and represent an extension of Lagrangian optimization theory, characterizing the solution to an optimization problem [97] In particular, it may be shown [96] that the KKTC are satisfied at the solution to any optimization problem in situations where the constraints are linear Furthermore, for optimisation problems involving a convex objective function, the KKTC being satisfied constitutes both necessary and sufficient proof that a given set of values is the correct solution [142] That is, for such a problem the KKTC are satisfied at the solution point, and the solution point only Hence, solving an optimization problem of this form involves finding a solution to the KKTC [97] As described, the constraints applied in SVMs are linear Furthermore, the objective function is always convex [97] Therefore, in SVMs the KKTC will always hold, and satisfying them is always sufficient proof that a proposed solution is correct Hence,

solving the SVM problem is equivalent to finding a solution to the KKTC Furthermore, it may be shown [96] that for a convex objective function with linear constraints any local minimum is also guaranteed to be a global minimum, meaning that there are no local minima in the SVM training problem like there can be with neural networks

D 2.3. Support Vector Machine Test Phase

The challenge of the test phase is to determine on which side of the decision boundary (the hyperplane lying half way between H1 and H2 and parallel to them) a given test sample lies The hypothesis space is the set of functions given in **(D 7)** It may be shown [97] that substituting-in the solution determined by satisfying the KKTC results in a decision function formulation, in which the decision surface appears as a dot product between data points - see **(D 9)** (ignoring for now α_i, which is the Lagrange Multiplier for $\mathbf{x}_i$)

$$f(x) = \text{sgn}\left(\sum_i \alpha_i y_i x_i \quad x + b\right) \qquad \textbf{(D 9)}$$

That is, the decision function involves calculating the dot product between a test point ($\mathbf{x}$) and each of the support vectors ($\mathbf{x}_i$) in turn, and then multiplying each time by $\mathbf{y}_i$ (i e 1 for positive examples and –1 for negative examples) Since the dot product can be understood as a similarity measure, it can be seen that the decision function essentially measures the similarity between the test point and each of the support vectors For instance, each positively labelled support vector will pull the result towards the positive direction depending on this similarity (and the weight α_i assigned to the particular support vector), and similarly each negatively labelled support vector will pull the result in a negative direction In this way, if the test point is more similar to the positive support vectors (taking the initial bias, $\mathbf{b}$, and the weights into consideration) then the point will be classified as positive Alternatively, if it is more similar to the negative support vectors it will be labelled as negative Furthermore, it will be later shown that the dot product formulation is crucial to allowing the procedure to be generalised to the non-linear case

D.3. SVMs For Non-Separable Data

The SVM system described so far is based on the assumption that the data is separable That is, it assumes that the separating hyperplane exists Sometimes however, due to noise in the data, no such separating hyperplane can be found In this case, the data is non-separable and no feasible solution will be found However, certain strategies for overcoming this problem have been developed as follows

D 3 1 Slack Variables & The Error Penalty

In [96] it is shown how the problem of non-separable data may be overcome if the set of constraints in **(D 4)** are relaxed when necessary That is, if some points are permitted to be classified incorrectly during training This is achieved by introducing slack variables, ξ_i, into the constraints of **(D 3)**, which then become those of **(D 10)** Again, the two inequalities can be combined as shown in **(D 11)**

$$\begin{aligned} w\ x\ +\ b\ &\geq\ +1-\xi_i \qquad \forall\ y_i = +1 \\ w\ x\ +\ b\ &\leq\ -1+\xi_i \qquad \forall\ y_i = -1 \\ \xi_i &\geq 0 \end{aligned} \tag{D 10}$$

$$y_i(w\ x\ +\ b)\ \geq\ 1-\xi_i \quad \forall i \tag{D 11}$$

The value ξ_i can be seen as a measure of how much a particular point violates the constraint From **(D 11)** it follows that any training points with a value for ξ_i greater than the value **1** will be misclassified, whereas points with a value between **0** and **1** will be classified correctly, but will fall inside the margin **Fig D 5** [83] illustrates two training points with slack variables greater than zero For point $\mathbf{x}_j$, the ξ_j value is greater than **1** Therefore, the point has crossed the separating hyperplane and will not be learned correctly, and is accepted as an outlier For point $\mathbf{x}_i$ the ξ_i value is between **0** and **1** In this case the point will be learned correctly by the hyperplane illustrated, but will still incur an error penalty because it lies inside the margin Because only points with $\xi >$ **1** are misclassified, $\Sigma\xi_i$, the total value of the error, can also be seen as an upper boundary on the total number of training points classified incorrectly That is, if all misclassified points had ξ_i=**1**, and all correctly classified points had ξ_i=**0**, then $\Sigma\xi_i$ would simply equate to the number of training points misclassified Thus, the total value for the error, $\Sigma\xi_i$, becomes another term in the objective function to be minimized [96] (which was simply $\frac{1}{2}\|w\|^2$ for the separable case) That is, the function to be minimized

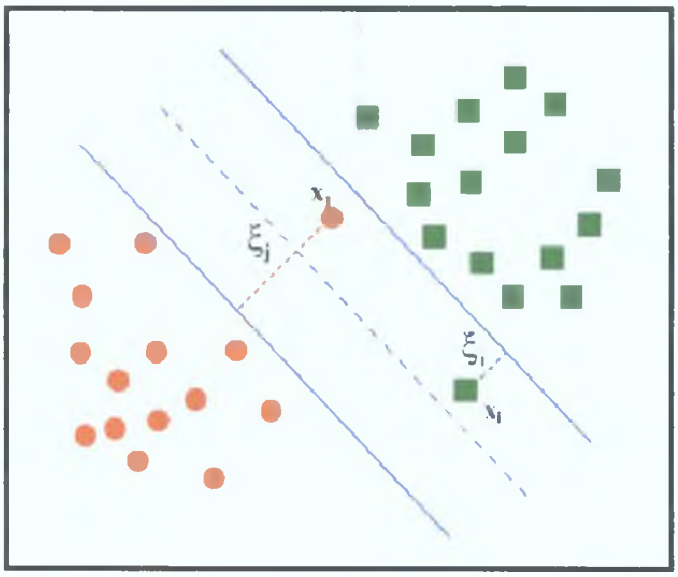


Fig. D.5. Two training points for which the slack variable ξ is greater than zero [83].

becomes that shown in **(D.12)** [2] [96], where **C** is a user chosen value known as the ***error penalty***.

$$\frac{1}{2}\|w\|^2 + C(\sum \xi_i)^k \qquad \textbf{(D.12)}$$

Note that the first term in this function acts as a bound on the VC dimension, and that the second term acts as a penalty for the number of training errors, or the empirical risk. This is how the SVM enforces the structural risk minimization induction principle, i.e. controlling both the VC dimension and the empirical risk simultaneously.

The error penalty, **C**, is a user chosen parameter which determines the relative significance of training errors compared to the size of the margin in the objective function to be optimized. That is, as **C** varies through a range of values, the normal $\|w\|$ varies smoothly through a corresponding range [97]. Hence, for a particular problem, choosing a particular value for **C** corresponds to choosing a value for $\|w\|$ and then minimizing ξ for that value. Since there is a value of **C** corresponding to the optimal choice of $\|w\|$, that value of **C** will therefore give the optimal bound [97]. Note that if $C=\infty$ then this solution is identical to that for separable data. That is, if the error penalty is infinite in magnitude, then clearly the tolerance for errors is zero and no training errors will be allowed. As before, a comprehensive description of the Lagrangian

[2] For reasons outside the scope of this thesis, it is normally preferred to set k=1 [105].

optimization and the protocol for satisfying the KKTC for the non-separable case SVM may be found in [97]

D.4. SVMs for Non-Linear Data

The above techniques for SVM data classification perform adequately in cases where the target function can be expressed as a linear function of the data Yet most of the time this is not so, i e in most real-world applications the target functions are non-linear However, if the data can be mapped into a higher dimensional feature space, where it can be separated by a linear decision function, then the same linear techniques outlined above may be applied, such that the data is separated in the feature space as opposed to in the input space [96] For example, a set of input vectors **(x,y)** could be mapped to a higher dimensional space as shown in **(D 13)**, where **Φ** represents the mapping

$$\Phi(x, y) = (x, x^2, xy, y^2) \qquad \textbf{(D 13)}$$

This mapping, from an input space $\mathbf{R}^2$ to a feature space $\mathbf{R}^4$, is based on the features of the input vector, and would make it easier to separate the data with a linear decision function if the target function was a quadratic polynomial [83] Clearly, more complex mappings to very high dimensional feature spaces could be created to suit situations where the target function is more complex – see [97]

However, working in high dimensional feature spaces is often unfeasible from a computational perspective This is one side of a problem known as the ***curse of dimensionality*** [143] The other side of this problem is overfitting, although, it has already been shown how the SVM approach overcomes this by maximising the margin between the separating hyperplanes In overcoming the dimensionality problem, SVMs use a special type of function, known as a ***kernel function***, to implicitly map the data to a high dimensional feature space without having to explicitly create it Effectively, this means that SVMs gain all of the advantages of working in a high dimensional space (i e the ability to learn any training set correctly) without inheriting their disadvantages (i e the problems of overfitting and the computational difficulties of performing explicit calculations in high dimensional spaces)

D 4 1. Implicit Mapping Using Kernel Functions

It has already been shown how, in both the training and test phases, the data appears in the form of a dot product between points Specifically, in the case of the latter, the decision function is as given in **(D 9)** Working in a feature space defined by the mapping **Φ**, the decision function would thus be given as that in **(D 14)**

$$f(x) = \text{sgn}\left(\sum_i \alpha_i y_i \Phi(x_i)\ \Phi(x) + b\right) \qquad \textbf{(D 14)}$$

Therefore, if a kernel function $\mathbf{K(x_i, x_j)}$ may be found as defined in **(D 15)** then the RHS of such may be replaced by the LHS everywhere in the training and test algorithms, and therefore the mapping would not have to be explicitly calculated

$$K(x_i, x_j) = \Phi(x_i)\ \Phi(x_j) \qquad \textbf{(D 15)}$$

That is, the feature space would be implied, without the computational overhead of dimensionality, and then crucially, a linear separation of the data could be performed in this feature space using the techniques outlined for the linear case Hence, the only extra overhead is that of computing the kernel function Some common, well-studied kernel functions are given in **(D 16)**, **(D 17)**, and **(D 18)** [96]

$$K(x, y) = s(x\ \ y + c)^p \qquad \textbf{(D 16)}$$

$$K(x, y) = \exp(\gamma \|x - y\|^2) \qquad \textbf{(D 17)}$$

$$K(x, y) = \tanh(kx\ \ y - \delta) \qquad \textbf{(D 18)}$$

The kernel of **(D 16)** defines a decision surface that is a polynomial of degree **p** in the data, that of **(D 17)** gives a Gaussian Radial Basis Function classifier, and that of **(D 18)** gives a particular kind of two-layer sigmoidal neural network [96]

D.5. Implementation and Performance

D 5.1. Training Phase Performance

The solution to the SVM training problem is found via the process of constrained Lagrangian optimization subject to satisfaction of the KKTC While this analysis is

outside of the scope of this thesis, a comprehensive discourse on such may be found in [97] Furthermore, the efficiency aspect of how the optimization problem may be most economically implemented is also addressed therein For example, typical strategies aim to breakdown large sized training quantities into more manageable, but still representative parts, i e schemes known as ***Chunking***, ***Decomposition***, and ***Sequential Minimal Optimisation*** [97]

D 5 2 Test Phase Performance

As described above, the formulation for the SVM decision function is given as shown in **(D 19)**, where i iterates through the number of support vectors

$$f(x) = \operatorname{sgn}\left(\sum_i \alpha_i y_i K(x_i, x_j) + b\right) \qquad \textbf{(D 19)}$$

On this basis, it is evident that the main factors influencing the running time in the test phase are the number of support vectors and the complexity of the kernel function In [96] it is proposed that the running time of the kernel function will typically be $\mathbf{O(D_n)}$, where $\mathbf{D_n}$ is the dimensionality of $\mathbf{R^n}$ This can be explained by the fact that the kernel function will need to iterate through the features of the input vector The time taken to test a single point will therefore be $\mathbf{O(D_n(N_s))}$, where $\mathbf{N_s}$ is the number of support vectors [96]

It should be clear from the above that the execution time in the test phase tends to suffer in situations where there are a large number of support vectors In fact, SVM solutions may display very slow performance in the test phase for this reason [144] In [145] a solution to this problem is proposed, which aims to reduce the number of support vectors required to describe a given decision surface The technique starts with a trained SVM the number of support vectors required in the new solution is decided *a priori* The technique then approximately recreates the decision surface given by the input trained SVM using fewer support vectors The support vectors created by this technique are not part of the training set and may not lie on the decision boundary, rather they are created artificially to approximate the decision surface input to the algorithm This technique has been shown to speed up performance by a factor of ten, making the performance comparable with that of neural networks, without having any significant impact on generalization performance [145]

Appendix E

SVM Implementation

This appendix provides a brief introduction to the specific Support Vector Machine (SVM) implementation used in this work

The SVM implementation employed for these experiments is that known as **SVM**^{light} [146], which since development has been made freely available for scientific use Written in the C programming language, **SVM**^{light} is an implementation of Vapnik's Support Vector Machine [95], and complete descriptions of the optimization algorithms used may be found via [147] and [148] As described, **SVM**^{light} has minimal memory requirements [147], and in addition, it has been shown to handle problems with many thousands of support vectors efficiently [149] Furthermore, the implementation exploits the fact that many tasks have the property of sparse instance vectors, leading to very compact and efficient representations [147] In practice, **SVM**^{light} has been used on a large range of problems, including text classification, image recognition tasks, bioinformatics and medical applications

Appendix F

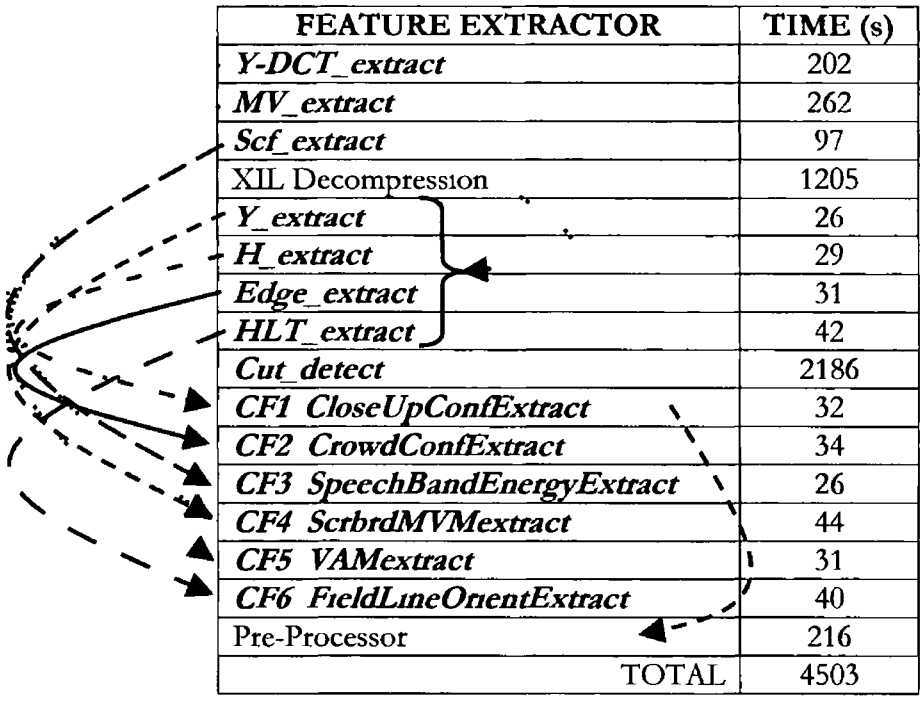
Speed Performance

This appendix provides an evaluation of the system speed performance of the developed scheme Taken as a whole, the description of the system may be broadly divided into two stages, i e that concerning the feature extraction (including pre-processing) process, and the pattern classification phase Hence, an assessment of the time taken to execute the underlying processes of these two stages was performed For a point of reference, it should be noted that (i) the tests were conducted on a 2GHz Intel Pentium-4 powered PC platform (512MB of RAM) running Red Hat Linux 7 2, (ii) the video images were captured at CIF resolution at 25fps, with audio data captured in 128kbps stereo at a sampling frequency of 44100 samples per second per channel, and (iii) the SVM implementation used was SVM^{light} version 6 01

F.1. Feature Extraction Speed Performance

Based on a one-hour video sample extracted from the test-corpus, **Table F 1** presents the processing time estimations for the components of the feature extraction stage as described in *Chapter 5* and *Appendix B*, where for clarity, the relationships between these are explicitly illustrated Note that in reflecting the actual scheme implementation, only one account of XIL-based decompression is accounted for, since even though four separate signal-level feature extractors were described as employing this process, i e ***Y_extract***, ***H_extract***, ***Edge_extract***, and ***HLT_extract*** (see *Appendix B*), it is clearly required to invoke this procedure only once Overall, the total time required to complete the feature extraction process was estimated to be 4503s, which corresponds to approximately 75-minutes, i e 1 25 times real-time for a one-hour video

Table F 1 Processing time estimations for system feature extractors and preprocessor based on the analysis of one hour of MPEG-1 video



FEATURE EXTRACTOR	TIME (s)
Y-DCT_extract	202
MV_extract	262
Scf_extract	97
XIL Decompression	1205
Y_extract	26
H_extract	29
Edge_extract	31
HLT_extract	42
Cut_detect	2186
CF1 CloseUpConfExtract	32
CF2 CrowdConfExtract	34
CF3 SpeechBandEnergyExtract	26
CF4 ScrbrdMVMextract	44
CF5 VAMextract	31
CF6 FieldLineOrientExtract	40
Pre-Processor	216
TOTAL	4503

From the data recorded in the table it is evident that the most time-consuming processes correspond to the shot boundary detection algorithm (***Cut_detect*** [79]), and the XIL-based image decompression process That is, the combined processing time required to execute these two procedures amounts to 3391s, which corresponds to in excess of 75% of the total time required to complete the overall feature extraction Moreover, of these two identified procedures, it is clearly that of the shot boundary detection that is by far the most time-consuming, i e the ***Cut_detect*** algorithm required 2186s to complete its task, which represents over 49% of the total time required

F.2. Pattern Classification Speed Performance

Recall that during the pattern classification phase of the experiments the SVM error penalty value (C) was varied throughout a critical set of values such that the range of possible performances of the scheme may be observed It was therefore considered desirable to gauge the effect this parameter variance had (if any) on the subsequent speed performance of the training and testing tasks To this end, **Figs F 1, F 2,** and **F 3** illustrate the fluctuations in SVM training time, the number of support vectors rendered

in each case, and in the time taken for SVM classification

From Fig F 1 it is evident that as the value of C was increased (i e from 0 02 to 2), the time taken to train the SVM on the training-corpus was found to progressively increase, i e from 490s up to approximately 1200s Hence the variance in the training times exhibited was quite substantial, i e a 700s variance across the spectrum that C traverses No explicit reason for this observed occurrence is immediately apparent, except to conclude that the value of the error penalty tends to have some bearing on the time required for the SVM to converge on an optimal solution It was described earlier how the training speed performance of an SVM suffers when the training set is large Given the sizeable training times recorded, this phenomenon is clearly apparent for the scenario herein

In contrast, it was found that as the value of C was increased, the number of support vectors rendered in each case decreased from 1474 to 1193, as illustrated in Fig F 2 Likewise, it was observed that for each corresponding trained SVM classifier, the time taken to classify the test-corpus content decreased from 39s to 31s for the increase in C, as illustrated in Fig F 3 It is described in *Section D 5 2* how the classification speed performance of an SVM is exclusively dictated by the number of support-vectors required to represent the solution This explains the close relationship between the number of support-vectors rendered for the variance in C, and the observed SVM classification times Recall that the training and test corpuses are essentially equal in size, and therefore consist of approximately the same number of training/test points However, bearing in mind the times taken for SVM training (in the order of hundreds of seconds), the times taken for SVM classification (in the order of tens of seconds) may be considered negligible in comparison The substantial difference between the two cases is due to the SVM attribute of sparseness, which is the fact that the final solution found is typically defined by a much smaller subset of the training data – see *Section D 2 1 (Appendix D)* Furthermore, the variance in training times observed in training as C traverses its prescribed range (approximately 700s) vastly exceeds that observed in classification (approximately 10s)

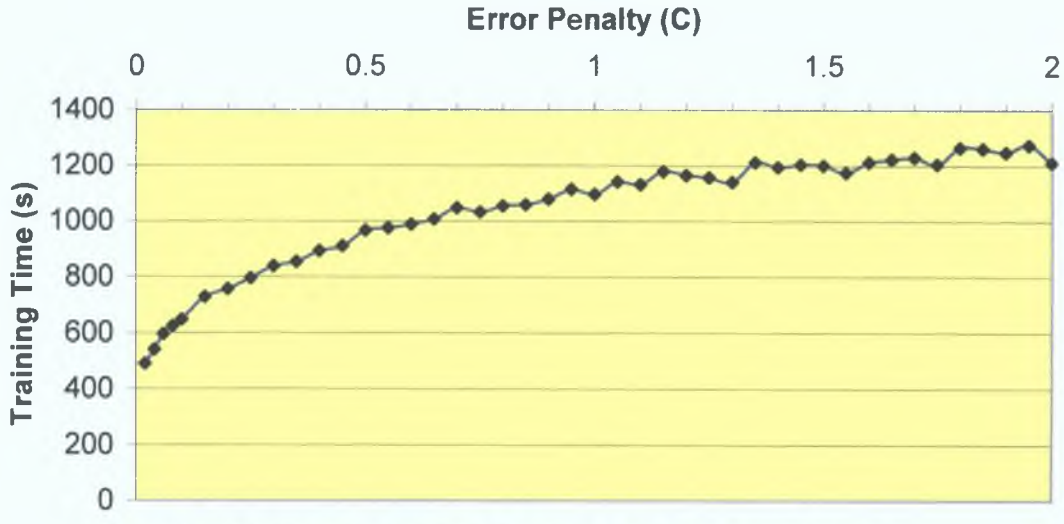


Fig. F.1. Variance of SVM training time with error penalty.

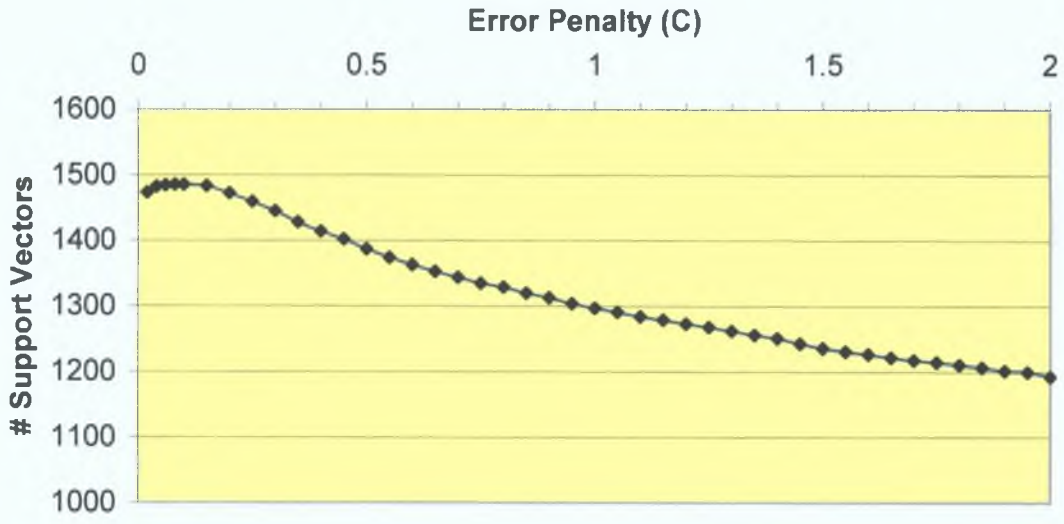


Fig. F.2. Variance in number of support vectors with error penalty.

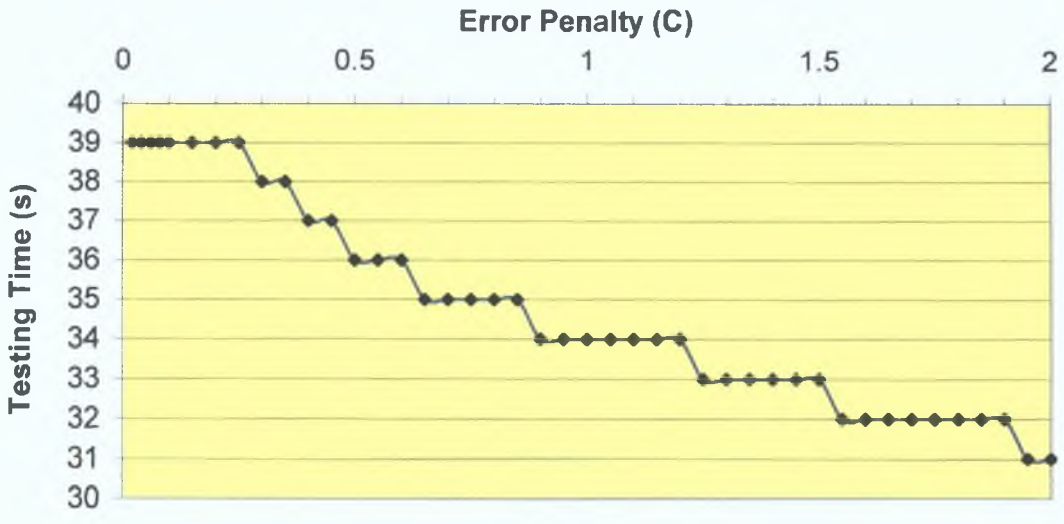


Fig. F.3. Variance in testing time with error penalty.

Appendix G

Improving Speed Performance

As part of an overall future work proposal, this appendix describes potential avenues for improving the speed performance of the developed scheme, the actual evaluation of which is described in *Appendix F*

G.1. The Probing Domain

As described in *Section 5 3*, the pre-processing filter is a bilateral mechanism, which de-lists shots from SFV pattern analysis probing based on whether or not they are (i) immediately followed by a close-up image and/or (ii) deemed to constitute advertisements Hence a potential future work task concerns somehow further improving the content rejection capability of the pre-processor stage That is, it is proposed that if the probing domain of the pattern classification stage may be made more selective, a noticeable improvement in the computation efficiency should be apparent for the scheme

G.2 Training & Classification

From Fig F 1 (*Appendix F*), it is estimated that at the global optimum performance point of the system (i e C=0 5), the time required to train the SVM using the prescribed training corpus was 966s, which may be considered quite large However, the SVM implementation utilized in the developed scheme, i e **SVM**^{light} (see *Appendix E*), incorporates many of the training time optimization algorithms mentioned in *Section D 5 1* (*Appendix D*) Therefore, it is considered reasonable to conclude that the long

training times observed have more to do with the magnitude of the training dataset used, rather than any deficiencies in the chosen SVM implementation However, it must be noted that while the SVM training phase has, in general, been shown to be an all-round time-consuming process, in terms of the scheme implementation it is a one-off procedure That is, once it is performed, the corresponding classifiers are generated and no further training is required

In contrast, the SVM classification process is a procedure that needs to be performed each time a given video is to be processed However, given the processing times rendered for the test-corpus classification (Fig F 3), it is concluded that once the learned model is to hand, the time required for the SVM to perform the classification task should be negligible compared to the duration of the input video Given this, it is argued that the processing times for the classification stage need not be considered in terms of the proposed future work task of system acceleration

G.3. Feature Extraction

The task of feature extraction is another process that must be performed each time a video is to be processed It has been shown that, relative to the duration of an input video, the combined processing times of the processes currently underpinning this stage are large Specifically, it was estimated in *Section F 1* that for a 1-hour sample video the time taken to complete the feature extraction stage equates to approximately 1 25 times real-time Given that the training phase is done off-line, and that the time demands for the classification phase are negligible, the main bottleneck in terms of system implementation corresponds primarily to the feature extraction stage Hence, a description on how compressed domain processing may be applicable in order to alleviate this now follows

As explained in *Section 5 1*, the implementation of the frame-level critical feature extraction methodologies are rooted in the processing of extracted signal-level feature evidence, and as described in *Appendix B*, most of these signal-level features are extracted from the decompressed audiovisual signals of the videos (e g pixel luminance/hue, edge data, etc) However, three signal-level components are extracted directly from the compressed domain video bitstreams (i e the DCT coefficients, the motion vectors, and the scalefactor data) Of the frame-level critical features, two (i e CF3 and CF5) are derived exclusively from this compressed domain data alone From

the feature extraction processing time estimations given in Table F 1, it is evident that, as expected, the extraction of these two features was significantly more efficient in terms of processing speed/time compared to those requiring a full (XIL-based) decompression On this basis, it is proposed that a potential future work task involves undertaking the redevelopment of the frame-level feature extractors currently based on exploiting decompressed signal-level data, such that they may be derived from compressed domain equivalents

For example, one possibility might be to implement CF2 (the crowd image detection algorithm) based on extracted DCT coefficient evidence alone Recall that within the current scheme, crowd image detection is facilitated by exploiting the fact that such views represent inherently uniform high-frequency textured images On this premise, crowd image confidence values are then generated based on an uncompressed domain edge-proliferation attribute (see *Section 4 4 2 2*) However, it is also recognized that discrimination between high-frequency and low-frequency image texture may be made at the pixel-block level by examining the encoded profusion of non-zero AC-DCT coefficients (see *Section B 2 2*) It is proposed that this suggests a hypothesis upon which a methodology for the extraction of crowd image confidence values exclusively on the basis of compressed domain signal data may be developed

Finally, from Table F 1 it is evident that of the individual processes underpinning the feature extraction, it is the task of shot-boundary detection, implemented herein by [79], that is by far the most time consuming procedure Recall that [79] performs this task by generating frame-to-frame dissimilarity measures based on a comparison of colour histograms/moments That is, it requires access to colour information from decompressed video images Therefore, another future work task concerns either sourcing or developing an alternative, less time-consuming algorithm, such that when plugged into the system a significant improvement in overall processing time might be observable (e g the compressed domain shot boundary detection algorithms described in *Section A 2*)

References

[1] R Lienhart, S Pfeiffer, and W Effelsberg, "Video Abstracting," in Communications of the ACM, Vol 40, pp 55-62, 1997

[2] S Nepal, U Srinivasan, G Reynolds, "Automatic Detection of 'Goal' Segments in Basketball Videos," proc 9th ACM international conference on Multimedia (ACM MM'01), pp 261-269, Ottawa, Canada, 2001

[3] A V Ratnaike, B Srinivasan, S Nepal, "Making Sense Of Video Content," proc 11th ACM international conference on Multimedia (MM'03), pp 650-651, Berkeley, CA, USA, 2003

[4] A Kinane, V Muresan, N O'Connor, N Murphy, S Marlow, "Energy-Efficient Hardware Architecture for Variable N-point 1D DCT," proc IEEE International Workshop on Power And Timing Modeling, Optimization and Simulation (PATMOS 2004), Santorini, Greece, 15-17 September 2004

[5] N O'Connor, V Muresan, A Kinane, D Larkin, S Marlow, N Murphy, "Hardware Acceleration Architectures for MPEG-Based Mobile Video Platforms A Brief Overview," proc 4th Workshop on Image Analysis for Multimedia Interactive Service (WIAMIS 2003), London, U K , 9-11 April 2003

[6] A Jain, S Chaudhuri, "A Spatio-Temporal Approach To Video Summarization And Retrieval " Technical report, School of Computer and Information Science, Indiana University-Purdue University Indianapolis, Indiana, USA, 2004

[7] H Sundaram, L Xie, S-F Chang, "A Utility Framework For The Automatic Generation Of Audio-Visual Skims," proc 10th ACM International Conference on Multimedia (MM'02), pp 189-198, Juan-les-Pins, France, 2002

[8] A Hanjalic, "Multimodal Approach to Measuring Excitement in Video," proc IEEE International Conference on Multimedia and Expo (ICME 2003), 2003

[9] N O'Connor, C Czirjek, S Deasy, S Marlow, N Murphy, A Smeaton, "News Story Segmentation in the Fischlar Video Indexing System," proc International Conference on Image Processing (ICIP'01), Thessaloniki, Greece, 10-12 October 2001

[10] J-Y Pan, H-J Yang, C Faloutsos, "MMSS Multi-modal Story-oriented Video Summarization," proc 4th IEEE Conference on Data Mining (ICDM'04), Brighton, UK, November 1-3, 2004

[11] B Lehane, N O'Connor, N Murphy, "Dialogue Sequence Detection in Movies," proc International Conference on Image and Video Retrieval (CIVR'05), W-K Leow et al (Eds), LNCS 3569, pp 286-296, Singapore, 20-22 July 2005

[12] B Lehane, N O'Connor, N Murphy, "Action Sequence Detection in Motion Pictures," proc European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology, London, U K, 25-26 November 2004

[13] K D Yow, B-L Yeo, M Yeung, and B Liu, "Analysis and Presentation of Soccer Highlights from Digital Video," proc 2nd Asian Conference on Computer Vision, Vol 2, pp 499-503, December 1995

[14] S Choi, Y Seo, H Kim, K-S Hong, "Where Are the Ball and Players? Soccer Game Analysis with Color Based Tracking and Image Mosaick," proc 9th International Conference on Image Analysis and Processing (ICIAP '97), Vol 2, pp 196-203, 1997

[15] D D Saur, Y-P Tan, S R Kulkarni and P J Ramadge, "Automated Analysis and Annotation of Basketball Video," in Symp Electronic Imaging Science and Technology Storage and Retrieval for Image and Video Databases, Vol 3022, pp 176-187, Jan 1997

[16] T Kawashima, K Tateyama, T Iijima, and Y Aoki, "Indexing of Baseball Telecast for Content-Based Video Retrieval" proc IEEE International Conference on Image Processing (ICIP'98), pp 871-875, 1998

[17] F Sudhir, J C M Lee, A K Jain, "Automatic Classification of Tennis Video for High-Level Content-Based Retrieval," proc International Workshop on Content-based Access of Image and Video Databases (CAIVD'98), pp 81-90, 1998

[18] W Zhou, A Vellaikal, C-C J Kuo, "Rule-Based Video Classification System for Basketball Video Indexing," proc ACM Multimedia 2000, pp 213-216, Los Angeles, USA, November 2000

[19] H Kim, K S Hong "Soccer Video Mosaicing Using Self-Calibration and Line Tracking," proc 15th International Conference on Pattern Recognition (ICPR'00), Vol 1, pp 1592-1595, 2000

[20] P Xu, L Xie, S-F Chang, A Divakaran, A Vetro, H Sun "Algorithms and System for Segmentation and Structure Analysis in Soccer Video," proc 2001

IEEE International Conference on Multimedia and Expo (ICME'01), pp 184-187, 2001

[21] V Tovinkere, R.J Qian, "Detecting Semantic Events in Soccer Games Towards A Complete Solution," proc International Conference on Multimedia and Expo (ICME'01), pp 1040-1043, Tokyo, Japan, Aug 22-25, 2001

[22] O Utsumi, K Miura, I Ide, S Sakai, and H Tanaka, "An Object Detection Method for Describing Soccer Games from Video," proc IEEE International Conference on Multimedia and Expo (ICME'02), pp 45-48, Lausanne, Switzerland, 2002

[23] J Assfalg, M Bertini, A Del Bimbo, W Nunziati, and P Pala, "Soccer Highlights Detection and Recognition using HMMs," proc IEEE International Conference on Multimedia and Expo (ICME'02), pp 825-828, Lausanne, Switzerland, 2002

[24] L Xie, S-F Chang, A Divakaran, and H Sun, "Structure Analysis of Soccer video with Hidden Markov Models," proc IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'02), pp 4096-4099, Orlando, Florida, USA, 2002

[25] P Chang, M Han, and Y Gong, "Extract Highlights from Baseball Game Video with Hidden Markov Models," proc International Conference on Image Processing (ICIP'02), Vol 1, pp 609-612, 2002

[26] M Lazarescu, S Venkatesh and G West, "On the Automatic Indexing of Cricket using Camera Motion Parameters," proc International Conference on Multimedia and Expo (ICME'02), pp 809-813, Lausanne, Switzerland, 2002

[27] A Ekin, A M Tekalp, and R Mehrotra, "Automatic Soccer Video Analysis and Summarization," in IEEE Transactions on Image Processing, Vol 12(7), pp 796-807, June 2003

[28] E Kijak, L Oisel, and P Gros, "Temporal Structure Analysis of Broadcast Tennis Video using Hidden Markov Models," in Symp Electronic Imaging Science and Technology Storage and Retrieval for Media Databases, Vol 5021, pp 277-288, Jan 2003

[29] J Assfalg, M Bertini, C Colombo, A Del Bimbo, W Nunziati, "Automatic Interpretation of Soccer Video for Highlights Extraction and Annotation," proc ACM Symposium on Applied Computing (SAC'03), Melbourne, FL (USA), pp 769-773, 2003

[30] Y Rui, A Gupta, and A Acero, "Automatically Extracting Highlights for TV Baseball Programs," proc ACM Multimedia 2000, pp 105-115, Los Angeles, USA, 2000

[31] D Zhang and D Ellis, "Detecting Sound Events in Basketball Video Archive" Technical report, Dept of Electronic Engineering, Columbia University, 2001

[32] R Cabasson and A Divakaran, "Automatic Extraction of Soccer Video Highlights using a Combination of Motion and Audio Features," in Symp Electronic Imaging Science and Technology Storage and Retrieval for Media Databases, Jan 2002, vol 5021, pp 272-276

[33] M Petkovic, V Mihajlovic, M Jonker, and S Djordjevic-Kajan, "Multi-Modal Extraction of Highlights from TV Formula 1 Programs," proc IEEE International Conference on Multimedia and Expo (ICME'02), pp 817-820, Lausanne, Switzerland, 2002

[34] B Li and M I Sezan, "Event Detection and Summarization in American Football Broadcast Video," in Symp Electronic Imaging Science and Technology Storage and Retrieval for Media Databases, Vol 4676, pp 202-213, Jan 2002

[35] R Dayhot, A Kokaram, and N Rea, "Joint Audio-Visual Retrieval for Tennis Broadcasts," proc IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03), pp 561-564, April 2003

[36] S-C Chen, M-L Shyu, C Zhang, L Luo, M Chen, "Detection of Soccer Goal Shots Using Joint Multimedia Features and Classification Rules," proc 4th International Workshop on Multimedia Data Mining (MDM/KDD2003) in conjunction with the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp 36-44, Washington DC, USA, August 2003

[37] H Pan, P Van Beek, M Sezan, "Detection of Slow-Motion Replay Segments in Sports Video for Highlights Generation," proc IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP'01), Salt Lake City, Utah, USA, 2001

[38] D Zhong and S-F Chang, "Structure Analysis of Sports Video using Domain Models," proc IEEE International Conference on Multimedia and Expo (ICME'01), pp 920-923, Tokyo, Japan, 2001

[39] C Wu, Y-F Ma, H-J Zhang, and Y-Z Zhong, "Events Recognition by Semantic Inference for Sports Video," proc IEEE International Conference on

Multimedia and Expo (ICME'02), Vol 1, pp 805-808, Lausanne, Switzerland, 2002

[40] J Assfalg, M Bertini, C Colombo and A Del Bimbo, "Semantic Annotation of Sports Videos," in IEEE Multimedia, Vol 9, No 2, pp 52-60, 2002

[41] H Pan, B Li, and M I Sezan, "Automatic Detection of Replay Segments in Broadcast Sports Programs by Detection of Logos in Scene Transitions," proc IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'02), vol 4, pp 3385-3388, Orlando, Florida, USA, May 2002

[42] Z Xiong, R Radharkrishnan, A Divakaran, T S Huang, "Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework," proc IEEE International Conference on Multimedia and Expo (ICME '03) Vol 3, July 2003

[43] K A Peker, R Cabasson, and A Divakaran, "Rapid Generation of Sports Video High-Lights using the MPEG-7 Motion Activity Descriptor," in Symp Electronic Imaging Science and Technology Storage and Retrieval for Media Databases, Jan 2002, vol 4676, pp 318-323

[44] N Babaguchi, Y Kawai, and T Kitashi, "Event Based Indexing of Broadcasted Sports Video by Intermodal Collaboration," in IEEE Transactions on Multimedia, Vol 4, pp 68-75, March 2002

[45] L-Y Duan, M Xu, X-D Yu, Q Tian, "A Unified Framework for Semantic Shot Classification in Sports Videos," proc 10th ACM International Conference on Multimedia (MM'02), pp 419-420, Juan-les-Pins, France, 2002

[46] B Li, J Errico, H Pan, and M I Sezan, "Bridging the Semantic Gap in Sports," in Symp Electronic Imaging Science and Technology Storage and Retrieval for Media Databases, Vol 5021, pp 314-326, January 2003

[47] Z Xiong, R Radhakrishnan, A Divakaran, "Generation of Sports Highlights using Motion Activity in Combination with a Common Audio Feature Extraction Framework," proc IEEE International Conference on Image Processing (ICIP'03), Vol 1, pp 5-8, September 2003

[48] A Hanjalic, "Generic Approach to Highlights Extraction from a Sport Video," proc IEEE International Conference on Image Processing (ICIP'03), Vol 1, pp 1-4, Barcelona, 2003

[49] C Jianyun, L Yunhao, L Songyang, W Lingda, "A Unified Framework for Semantic Content Analysis in Sports Video," proc International Conference on

Information Technology and Applications (ICITA'04), pp 149-153, Harbin, China, 2004

[50] S M Goldwasser, "TV and Monitor CRT (Picture Tube) Information Version 1 77 " Available http //arcadecontrols com/files/Miscellaneous/crtfaq htm

[51] K R Rao, J J Hwang, "Techniques & Standards for Image, Video and Audio Coding" Prentice Hall, 1996 ISBN 0-13-309907-5

[52] A Majumder, R Stevens, "Perceptual Photometric Seamlessness in Tiled Projection-Based Displays," proc ACM Transactions on Graphics, Vol 24, No 1, pp 118-139, January 2005

[53] A H Munsell (1946), "A Color Notation," Munsell Color Co , Baltimore, MD, USA

[54] P Browne, A Smeaton, N Murphy, N O'Connor, S Marlow, C Berrut, "Evaluating and Combining Digital Video Shot Boundary Detection Algorithms," proc IMVIP 2000, Belfast, Northern Ireland, 31st August – 2nd September 2000

[55] A Chandrashekhara, H M Feng, T-S Chua "Temporal Multi-Resolution Framework for Shot Boundary Detection and Keyframe Extraction " TREC (Text REtrieval Conference), Gettysburg, November, 2003

[56] R M Ford, "A Quantitative Comparison of Shot Boundary Detection Metrics," proc SPIE, Vol 3656, Storage and Retrieval for Video Databases VII, pp 666-676, December 1998

[57] R Lienhart, "Comparison of Automatic Shot Boundary Detection Algorithms " In Image and Video Processing VII 1999, proc SPIE 3656-29, Jan 1999

[58] W J Heng, K N Ngan, and M H Lee, "Comparison of MPEG Domain Elements for Low-Level Shot Boundary Detection," in Journal of Real-Time Imaging, Special Issue on Real-Time Digital Video over Multimedia Networks, Vol 8, No 5, pp 341-358, October 1999

[59] X-S Hua, D Zhang, M Li, H-J Zhang, "Performance Evaluation Protocol for Video Scene Detection Algorithms," proc 4th International Workshop on Multimedia Information Retrieval (MIR'02), Juan-les-Pins, France, 2002

[60] A B Watson, "Image Compression using the Discrete Cosine Transform " In The Mathematica Journal, Volume 4, Issue 1, 1994

[61] F Hoffman, "An Introduction to Fourier Theory," Available http //lanoswww epfl ch/studinfo/courses/cours_nonlinear_de/extras/Hoffman(1997)_An_Introduction_to_Fourier_Theory pdf

[62] R L Lux, "Principal Components Analysis An Old but Powerful Tool for ECG Analysis" In the International Journal of Bioelectricmagnetism, Vol 5, No 1, 2003, pp 342-345

[63] N O'Connor, N Murphy, S Marlow, "Module EE554 Image & Video Compression," course material corresponding to module, School of Electronic Engineering, Dublin City University

[64] T Sikora, "Digital Video Coding Standards and Their Role in Video Communications" In "Signal Processing for Multimedia" published by IOS Press, 1999, J S Byrnes (Ed) ISBN 90 5199 460 5

[65] D Marshall, "Multimedia," Module CM0340, Dept of Computer Science, Cardiff University
Available http //www cs cf ac uk/Dave/Multimedia/index html

[66] The official MPEG homepage http //www chiariglione org/mpeg

[67] The official JPEG homepage http //www jpeg org/

[68] D Pan, "A Tutorial on MPEG Audio Compression," proc IEEE Multimedia Vol 2, No 2, pp 60-74, 1995

[69] N Babaguchi, Y Kawai, Y Yasugi, T Kitahashi, "Linking Live and Replay Scenes in Broadcasted Sports Video, proc 2000 ACM workshops on Multimedia (MULTIMEDIA'00), pp 205-208, Los Angeles, California, United States, 2000

[70] L Wang, B Zeng, S Lin, G Xu, H-Y Shum, "Automatic Extraction of Semantic Colours in Sports Video," proc IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), Montreal, Canada, 2004

[71] J-C Terrillon and S Akamatsu "Comparative Performance of Different Chrominance Spaces for Colour Segmentation and Detection of Human Faces in Complex Scene Images," proc 12th Conf on Vision Interface, Vol 2, pp 180-187, May 1999

[72] T Ojala, M Pietikainen, "Texture classification," in CVonline - Compendium of Computer Vision, R B Fisher (ed), 2001 Available http //homepages inf ed ac uk/rbf/CVonline/LOCAL_COPIES/OJALA1/texclas htm

[73] X Wang, "On Cortical Coding of Vocal Communication Sounds in Primates," proc The National Academy of Sciences, USA 97 11843-11849 (2000)

[74] R C Gonzalez and R E Woods, "Digital Image Processing", Addison-Wesley, New York, 1992 ISBN 0201180758

[75] J Canny "Computational Approach to Edge Detection", IEEE Trans Pattern Analysis and Machine Intelligence, vol 8, no 6, pp 679-698, Nov 1986

[76] L Roberts, "Machine Perception of 3-D Solids," Optical and Electro-optical Information Processing, MIT Press, 1965

[77] T Risse, "Hough Transform for Line Recognition," Proc Computer Vision and Image Processing, 1989, 46, 327-345, 1989

[78] IBM Corporation homepage http //www ibm com

[79] C O'Toole, A Smeaton, N Murphy, S Marlow, "Evaluation of Shot Boundary Detection on a Large Video Test Suite," proc Challenges in Image Retrieval, Newcastle (UK), February 1999

[80] D A Sadlier, S Marlow, N O'Connor, N Murphy, "Automatic TV Advertisement Detection from MPEG Bitstream," in Pattern Recognition, Special Issue On Pattern Recognition in Information Systems, Vol 35, No 12, December 2002 ISSN 0031-3203

[81] C V Schwab, S Shouse, L Miller "Recognise Limitations to Avoid Injury", Fact Sheet Pm-1563j, The Safe Farm Program, Iowa State University Extension, USA, December 1994

[82] X Sun, B S Manjunath and A Divakaran, "Representation of Motion Activity in Hierarchical Levels for Video Indexing and Filtering," proc IEEE International Conference on Image Processing (ICIP'02), New York, USA, 2002

[83] N K O'Hare, "Using Support Vector Machines to Segment Digital Video," MSc, School of Computing, Dublin City University, 2003, ref M0098900DC

[84] T M Mitchell, "Machine Learning," McGraw-Hill, 1997, ISBN 0-07-042807-7

[85] V Vapnik, "Estimation of Dependences Based on Empirical Data," Nauka, Moscow, 1979 (English translation published by Springer-Verlag, New York, 1982)

[86] V Roth "Probabilistic Discriminative Kernel Classifiers for Multi-Class Problems," in Pattern Recognition--DAGM'01, pp 246-253, Springer, LNCS 2191, 2001

[87] P Joseph, "HMM Based Classifiers," technical paper, Department of Computer Science, North Dakota State University, USA Available at http //www cs ndsu nodak edu/~pjoseph/CSCI765Paper pdf

[88] K T Abou-Moustafa, M Cheriet, C Y Suen, "Classification of Time-Series Data Using a Generative/Discriminative Hybrid," proc Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR'04), pp 51-56, 2004

[89] ChengXiang Zhai "Text Categorization," course material corresponding to module on "Introduction to Text Information Systems", ref CS397/CS498-CXZ, University of Illinois at Urbana-Champaign, USA Available at http //sifaka cs uiuc edu/course/498cxz04f/ loc/textcat ppt

[90] S N Srihari "CSE 574 Machine Learning," course material for module, University at Buffalo, State University of New York, USA Available http //www cedar buffalo edu/~srihari/CSE574/

[91] A Y Ng and M Jordan, "On Discriminative Vs Generative Classifiers A Comparison of Logistic Regression and Naive Bayes," proc Neural Information Processing Systems, 2002

[92] R Nallapati, "Discriminative Models for Information Retrieval," proc ACM SIGIR'04, Sheffield, UK, 2004

[93] T M Bae, C S Kim, S H Jin, K H Kim, and Y M Ro, "Semantic Event Detection in Structured Video Using Hybrid HMM/SVM," proc CIVR'05, Singapore, 2005

[94] C M Bishop, "Neural Networks for Pattern Recognition," Oxford University Press, 1995, ISBN 0-19-853564]

[95] V Vapnik "The Nature of Statistical Learning Theory," Springer-Verlag, New York, 1995 ISBN 0-387-94559-8

[96] C Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," in Data Mining and Knowledge Discovery, 2(2) 121-167, 1998

[97] N Cristianini, J Shawe-Taylor, "Support Vector Machines and Other Kernel-Based Learning Methods," Cambridge University Press, 2000, ISBN 0-521-78019-5

[98] L Lu, S Z Li, H-J Zhang, "Content-Based Audio Segmentation Using Support Vector Machines," proc ACM Multimedia Systems Journal, Vol 8(6), pp 482-492, March 2003

[99] J Wang, C-S Xu, E Chng, X Yu, Q Tian, "Event Detection Based on Non-Broadcast Sports Video," Proc IEEE ICIP'04, Singapore, 2004

[100] C Ma, M A Randolph, J Drish, "A Support Vector Machines-Based Rejection Technique for Speech Recognition," Proc IEEE ICASSP'01, Salt Lake City, Utah, USA, 2001

[101] S R Gunn, M Brown, K M Bossely, "Network Performance Assessment for Neuro-Fuzzy Data Modeling," in Intelligent Data Analysis, Vol 1208 of Lecture Notes in Computer Science, pp 313-323, 1997

[102] S R Gunn "Support Vector Machines for Classification and Regression," a technical report for the Faculty of Engineering, Science and Mathematics, School of Electronics and Computer Science, University of Southampton, 1998 Available http //www ecs soton ac uk/~srg/publications/pdf/SVM pdf

[103] H V Khuu, H-K Lee, J-L Tsai, "Machine Learning with Neural Networks and Support Vector Machines," Technical report, Dept of Electrical and Computer Engineering, University of Wisconsin, USA, 2003 Available at http //www cs wisc edu/~hiep/Sources/Articles/svms_nns_research pdf

[104] P Browne, C Czirjek, G Gaughan, *et al* "Dublin City University Video Track Experiments for TREC 2003" in TRECVID 2003 – Text REtrieval Conference TRECVID Workshop, Gaithersburg, Maryland, 17-18 November 2003

[105] E Lee Kai Chuan, "Website Exclusivity Learner/Classifier (WELCome)," honours year project report 2003/2004, Dept of Computer Science, School of Computing, National University of Singapore Available at http //wing comp nus edu sg/publications/theses/edwinLeeThesis pdf

[106] Y Sun, M Robinson, R Adams, et al "Integrating Binding Site Predictions Using Meta Classification Methods," proc International Conference on Adaptive and Natural Computing Algorithms (ICANNGA'05), Coimbra, Portugal, 21-23 March 2005

[107] A Leykin, M Tuceryan, "Automatic Determination of Text Readability Over Textured Backgrounds for Augmented Reality Systems" proc IEEE International Symposium on Mixed and Augmented Reality (ISMAR'04), November, 2004

[108] G Wu, E Chang, "Class-Boundary Alignment for Imbalanced Dataset Learning," in ICML Workshop on Learning from Imbalanced Data Sets II, Washington DC, 2003

[109] R Akbani, S Kwek, N Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets," proc 15th European Conference on Machine Learning (ECML'04), Pisa, Italy, September 20-24 2004

[110] K K Chin, "Support Vector Machines Applied to Speech Pattern Classification" 1998 masters thesis, Dept of Engineering, University of Cambridge, U K
Available http //mi eng cam ac uk/~kkc21/thesis_main/thesis_main html

[111] H V Khuu, H-K Lee, J-L Tsai, "Machine Learning with Neural Networks and Support Vector Machines," Technical report, Dept of Electrical and Computer Engineering, University of Wisconsin, USA, 2003 Available at http //www cs wisc edu/~hiep/Sources/Articles/svms_nns_research pdf

[112] Y Fu, R Sun, Q, Yang, *et al* "A Block-Based Support Vector Machine Approach to the Protein Homology Prediction Task in KDD Cup 2004," in ACM SIGKDD Explorations Newsletter, Vol 6, issue 2, pp 120-124, December 2004

[113] TREC Video Retrieval Evaluation (TRECVID 2005) Available http //www itl nist gov/iaui/894 02/projects/tvpubs/tv pubs org html

[114] R Lienhart, "Reliable Dissolve Detection," in Storage and Retrieval for Media Databases 2001, proc SPIE 4315, pp 219-230, Jan 2001

[115] U Naci, A Hanjalic, "Spatiotemporal Block Based Analysis of Video for Fast and Effective Shot Transition Detection and Identification," proc Advanced School for Computing and Imaging Conference (ASCI'05), Het Heijderbos, Heijen, The Netherlands, June 8-10, 2005

[116] S Marchand, "An Efficient Pitch-Tracking Algorithm Using A Combination Of Fourier Transforms," proc The COST G-6 Conference on Digital Audio Effects (DAFX-01), Limerick, Ireland, December 6-8, 2001

[117] G S Ying, L H Jamieson, and C D Mitchell, "A Probabilistic Approach to AMDF Pitch Detection," proc International Conference on Spoken Language Processing, Philadelphia, PA, Oct 1996, pp 1201-1204

[118] Text Retrieval Conference (TREC) URL http //trec nist gov

[119] TRECVID URL http //www-nlpir nist gov/projects/trecvid/

[120] X Yu and D Farin "Current and Emerging Topics in Sports Video Processing", proc International Conference on Multimedia and Expo (ICME 2005), Amsterdam, July 2005

[121] J R Wang, N Parameswaran, "Survey of Sports Video Analysis Research Issues and Applications," proc CRPIT '36 Proceedings of the Pan-Sydney area Workshop on Visual Information Processing, 2004, pp 87-90, (publishers Australian Computer Society, Inc , Darlinghurst, Australia)

[122] K Wan, C Xu, Q Tian, and M Leong, "Automatic Sports Content Analysis – State-Of-Art And Recent Results," article from broadcastpapers com Available http //www broadcastpapers com/BcastAsia04/BAsiaIIRAutoSports pdf

[123] Hawkeye Innovations URL www hawkeyeinnovations co uk

[124] H J Zhang, A Kankanhalli and S W Smoliar, "Automatic Partitioning of Full-Motion Video", in Multimedia Systems, Vol 1, pages 10-28, 1993

[125] R Zabih, J Miller, and K Mai, "A Feature-Based Algorithm for Detecting and Classifying Scene Breaks," proc ACM Multimedia (MM'95), pp 189-200, San Francisco, California, USA, November 1995

[126] A Nagasaka and Y Tanaka, "Automatic Video Indexing and Full-Video Search for Object Appearances," in Visual Database Systems II, Elsevier Science Publishers, pages 113-117, 1992

[127] X U Cabedo and S K Bhattacharjee "Shot Detection Tools in Digital Video," in Proceedings Non-linear Model Based Image Analysis 1998, Springer Verlag, pp 121-126, Glasgow, July 1998

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

[129] J Boreczky and L A Rowe, "Comparison of Video Shot Boundary Detection Techniques," in IS&T/SPIE proceedings Storage and Retrieval for Images and Video Databases IV, Vol 2670, pp 170-179, February 1996

[130] Solaris XIL 1 3 Imaging Library Programmer's Guide Sun Microsystems Inc , November 1993 Available http //docs sun com/app/docs/doc/802-5863

[131] The Berkeley Multimedia Research Center homepage http //bmrc berkeley edu/

[132] K L Gong and L A Rowe, "Parallel MPEG-1 Video Encoding," proc Picture Coding Symposium, Sacramento, CA, September 1994

[133] K Mayer-Patel, B C Smith, and L A Rowe, "The Berkeley Software MPEG-1 Video Decoder," proc ACM Transactions on Multimedia Computing,

Communications, and Applications, (TOMCCAP'05), Vol 1, No 1, pp 110-125, ACM Press, New York, USA, 2005

[134] D Banks & L A Rowe, "Analysis Tools for MPEG-1 Video Streams," Internal publication, Berkeley Multimedia Research Centre, University of California Berkeley, CA 94720-1776, June, 1995

[135] URL http //bmrc berkeley edu/frame/research/mpeg/mpeg_play html

[136] K Patel, B C Smith, L A Rowe, "Performance of a Software MPEG Video Decoder," proc ACM MM'93, pp 75-82, Anaheim CA 1993

[137] URL http //www sun com

[138] URL http //web mit edu/afs/athena/contrib/graphics/src/mpeg_system-1 1/maplay2/

[139] B Boser, M Guyon, V Vapnik, "A Training Algorithm for Optimal Margin Classifiers," proc 5th Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, ACM Press, 1992

[140] E Osuna, R Freund, F Girosi, "Support Vector Machines Training and Applications," A I Memo 1602, Massachusetts Institute of Technology, Artificial Intelligence Library, 1997 Available at ftp //publications ai mit edu/ai-publications/1500-1999/AIM-1602 ps

[141] M Minoux, "Mathematical Programming Theory and Algorithms", John Wiley & Sons, 1986

[142] R Fletcher, "Practical Methods of Optimisation," John Wiley & Sons Inc, 2nd edition, 1987

[143] C J Van Rijsbergen, "Information Retrieval," Butterworths Press, 1979 (2nd edition)

[144] T Downs, K E Gates, A Masters, "Exact Simplification of Support Vector Solutions," in Journal of Machine Learning Research, MIT Press, Vol 2, No 2, pp 293-297, May 2002

[145] C Burges, "Simplified Support Vector Decision Rules," proc The Thirteenth International Conference on Machine Learning, pp 71-77, Bari, Italy, 1996

[146] SVM^{light} URL - http //svmlight joachims org/

[147] T Joachims, "Making large-Scale SVM Learning Practical," Advances in Kernel Methods - Support Vector Learning, B Scholkopf, C Burges and A Smola (eds), MIT-Press, 1999

[148] T Joachims, "Learning to Classify Text Using Support Vector Machines" Dissertation, Kluwer, 2002

[149] K Morik, P Brockhausen, & T Joachims, "Combining Statistical Learning with a Knowledge-Based Approach – A Case Study in Intensive Care Monitoring," proc 16th International Conference on Machine Learning (ICML'99), Bled, Slovenia, 1999