# Shot Classification in Broadcast Soccer Video

Thesis by
Lionel Guimaraes

In Partial Fulfillment of the Requirements
for the Degree of
Master of Science in Computer Science

UNIVERSITY OF
KWAZULU-NATAL

University of KwaZulu-Natal
Durban, South Africa

2012

(Submitted 17th March 2013)

# Abstract

Event understanding systems, responsible for automatically generating human relatable event descriptions from video sequences, is an open problem in computer vision research that has many applications in the sports domain, such as indexing and retrieval systems for sports video. Background modelling and shot classification of broadcast video are important steps in event understanding in video sequences. Shot classification seeks to identify shots, i.e. the labelling of continuous frame sequences captured by a single camera action such as long shot, close-up and audience shot, while background modelling seeks to classify pixels in an image as foreground/background. Many features used for shot classification are built upon the background model therefore background modelling is an essential part of shot classification.

This dissertation reports on an investigation into techniques and procedures for background modelling and classification of shots in broadcast soccer videos. Broadcast video refers to video which would typically be viewed by a person at home on their television set and imposes constraints that are often not considered in many approaches to event detection. In this work we analyse the performances of two background modelling techniques appropriate for broadcast video, the colour distance model and Gaussian mixture model. The performance of the background models depends on correctly set parameters. Some techniques offer better updating schemes and thus adapt better to the changing conditions of a game, some are shown to be more robust to changes in broadcast technique and are therefore of greater value in shot classification. Our results show the colour distance model slightly outperformed the Gaussian mixture model with both techniques performing similar to those found in literature.

Many features useful for shot classification are proposed in the literature. This dissertation identifies these features and presents a detailed analysis and comparison of various features appropriate for shot classification in broadcast soccer video. Once a feature set is established, a classifier is required to determine a shot class based on the extracted features. We establish the best use of the feature set and decision tree parameters that result in the best performance and then use a combined feature set to train a neural network to classify shots. The combined feature set in conjunction with the neural network classifier proved effective in classifying shots and in some situations outperformed those techniques found in literature.

# Preface

The research discussed in this dissertation was done at the University of KwaZulu-Natal, Durban from April 2009 until November 2012 by Lionel Guimaraes under the supervision of Mr Anban Pillay and Professor Jules-Raymond Tapamo.

# Declaration – Supervisor

As the candidate’s supervisor I agree to the submission of this dissertation.

________________________________________

Mr Anban Pillay

# Declaration – Co-Supervisor

As the candidate’s co-supervisor I agree to the submission of this dissertation.

________________________________________

Prof Jules-Raymond Tapamo

# Declaration – Plagiarism

I, Lionel Guimaraes, declare that

i. The research reported in this dissertation/thesis, except where otherwise indicated, is my original work.

ii. This dissertation/thesis has not been submitted for any degree or examination at any other university.

iii. This dissertation/thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

iv. This dissertation/thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a) Their words have been re-written but the general information attributed to them has been referenced;

   b) Where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

v. Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

vi. This dissertation/thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation/thesis and in the References sections.

________________________________________

Lionel Guimaraes

To my family. Thank you for all the love and support you have shown me over the years and for helping me become the person I am today.

# Acknowledgments

Thank you to my advisors, Mr. Anban Pillay and Prof. Jules-Raymond Tapamo, for their guidance and understanding in the completion of this work. Thanks also to my fellow graduate students for creating an enjoyable environment for both learning and personal growth.

# Contents

# List of Figures

# List of Tables

# Nomenclature

| | |
|---|---|
| EM | Expectation Maximization |
| GMM | Gaussian Mixture Model |
| HSI | Hue, Saturation and Intensity |
| HSL/HLS | Hue, Saturation and Lightness |
| HSV | Hue, Saturation and Value |
| MLP-NN | Multi-layer Perceptron Neural Network |
| pdf | Probability density function |
| RGB | Red, Green and Blue |
| SVM | Support Vector Machine |

# Chapter 1

# Introduction

## 1.1 Background

Every year billions of hours of video are produced, created and recorded. As with other forms of media, users would benefit from the ability to sort, search and retrieve items relevant to them. The retrieval of images and video based on their content is a very active area of research with numerous applications ranging from medicine to entertainment. To enable content based retrieval requires the analysis of the video to produce content descriptions or annotations. The annotations facilitate content based sorting, searching and retrieval of the video sequences. This process can and often is performed manually but doing so is generally impractical due to the quantity of produced video and the subjectivity of the annotations, thus automated solutions need to be developed. There are numerous applications besides content based retrieval especially for sports video. These include automatic highlight generation, live commentary, post match tactics and performance analysis and content aware low bit-rate streaming. Automatic highlight generation is of particular interest. Being able to automatically generate a series of highlights from a game allows a user to watch most of the interesting activities from a game in a much shorter time. This also opens up possibilities that allow a user to receive a customised set of highlights based on user specified criteria.

Due to the massive variety of video available, developing general content analysis techniques that are able to process all types of video is infeasible at this stage, so this work focuses on a narrower context, that of broadcast soccer video. By limiting the focus, the goal of this work was to produce a more thorough exploration that could then be expanded beyond the initial scope. Sports video provides a rich context for analysis due to its constrained nature, strong structure and the adherence to specific rule-sets. This allows for a set of events to be defined which can be universally applied to all videos of a particular sport. Unlike general video content where different users may associate a different meaning to an event, a framework can be established for sporting events which can provide relatable information to all users. By creating a structured and systematic method for labelling events based on activities in the video sequences, the process of automatically producing event descriptions is greatly simplified. The focus on soccer is due to its worldwide popularity. General principles can be investigated that can then be expanded in future work to cover other sports.

Broadcast video refers specifically to video that has been produced for reception on television and is the most common format of sports video, which makes its analysis the sensible choice when considering general applicability. The nature of broadcast video poses certain challenges related to how the views change from camera to camera but it also brings certain advantages related to the specific nature of its production. Broadcast production strategies convey information and the identification of the techniques used can be used to infer semantics. This is opposed to situations where a camera setup may be purposely constructed to allow for video analysis, sometimes used for a single purpose such as the Hawk-Eye[2] visual tracking system. The output of these types of systems are generally not meant to be seen by the viewers and may have little relevance outside of their direct association with corresponding broadcast sequences.

## 1.2 Motivation

Significant amounts of research has been focused on content analysis for soccer video [7, 9, 10, 13, 14, 22, 25, 28, 32, 39, 41, 44, 46]. A broad range of solutions have been offered but many are incomplete or still in their infancy. An important goal of content analysis of sports video is to generate semantic event understanding which is to take the low level representation of a video sequence and translate it into a human relatable descriptions that make assertions about the content contained within the sequences. There are numerous events in soccer but a conservative set of events would include those directly related to game play such as goals, free kicks, penalties, throw-ins and corners. Additional details about the events can also be included such as who scored the goal or which team was awarded the free kick. Numerous techniques have been explored to solve the problem of event understanding and classification, for example, using multi-modal approaches that combine visual, textual and audio information [8, 7, 14, 16, 17, 43]. The visual and audio approaches are the most popular and are often used in combination with the textual information to enhance the classification made by the visual features.

When only the visual mode is used, the goal is to translate low-level visual features into semantic concepts. When constructing such a system for processing broadcast sports video, characterising views or shots is often a key process. A shot is a continuous frame sequence captured by a single camera action such as long shot, close-up and audience shot. Due to the way broadcast videos are produced, the information the producer wants to highlight for the audience influences the types of shots used and the order of their usage. Because the choice of shot is made with the specific intention of conveying information, identifying and classifying shots can provide a good basis to extract semantic information. Thus a common approach to event detection in broadcast soccer video has been shot pattern analysis, which detects events based solely on the order in which shots occur in a video sequence. In addition to production strategies, different shot classes also offer opportunities to extract different kinds of information useful for event detection. A shot which displays most of the play field area may be used to identify where the interesting actions are occurring or to determine the direction of play to suggest which team is attacking or defending, while a close up shot of a single player may give information about the player or team performing the actions. Even a shot of the audience may be an indicator that something interesting has occurred on the field and the producer wishes to show crowd reactions.

Thus solving the problem of shot classification is of significant importance to event classification and understanding, which in turn creates a basis for content analysis and retrieval.

This work focuses on methods and techniques that facilitate the classification of shot types in broadcast soccer video. To advance this goal three key problems have been identified: the modelling of the play-field area, the selection of shot features and the methods used to classify shots.

Shot are classified by examining a feature set extracted from a set of frames. These features are selected based on their ability to characterise certain shot types. This choice of features is crucial to differentiating among the various shot types. Importantly though, all the features used in this work rely on the quality of the play-field model. Thus the modelling of the play-field is a vital first step. Finally, with an appropriate feature set, the task of assigning a shot to a specific class is performed by a classifier. There are numerous classifiers with different complexities and other characteristics.

## 1.3 Objectives

The main aim of this work is the investigation of techniques and methods used in shot classification of broadcast soccer video and the establishment of each method's suitability for such a purpose.

The objectives of this work are to:

- Survey the literature to gain an understanding of the state of the art shot classification techniques.

- Investigate shot classification through the three key processes of background (play-field) modelling, feature selection and classifier selection.
  - Compare two background modelling techniques (viz. the colour distance model and the Gaussian mixture model) for their ability to model the field colour in broadcast soccer video. Key points of comparison include; the model's training requirements and how the requirements affect the model's usability, the accuracy with which the model represents the play-field, the ability of the model to adapt to changing illumination conditions within a sequence.
  - Compare three feature sets (viz. the field colour ratio, the vertical projection and the object size ratio) for shot classification and their efficacy for shot classification using the decision tree and neural network classifiers. Key points of comparison include; the features ability to work within a classifier to accurately classify shots, the potential of the feature to operate within a larger feature set and/or in conjunction with a more advanced classifier.

## 1.4 Contributions

The main contribution of this work is the establishment of a feature set and comparison of background modelling techniques for shot classification in broadcast soccer video.

A set of four shot classes have been proposed: the long, medium, close-up and out of field shot. Each provides an opportunity to extract certain kinds of information and are often used in the event classification process. Three sets of features; the field ratio, the vertical projection and the object size ratio, have been identified as having the potential to classify these shot types due to their ability to discriminate between the various shot types. These features have been individually explored in this work with particular attention given to each feature's classification potential.

The extraction of features remains largely dependant on the play-field model. Establishing effective techniques to do this is essential to the task of shot classification. Thus two techniques for field modelling in broadcast soccer video were identified as being both popular and accurate viz. the colour distance model and Gaussian mixture model. The ability of these techniques to perform the function of field modelling has been investigated with each technique being evaluated against similar techniques found in the literature and shown to perform at the same level.

To determine the effectiveness of the features a decision tree classifier was used for each individual feature. They were then combined into a single feature set and input into a neural network classifier. Although the decision tree is one of the most basic classifiers, it was selected for this reason as other more complex classifiers may have obfuscated the performance of the underlying features. Given the lack of well established features for shot classification it was important to individually assess the performance of features rather than just the classification system as a whole. As a result the individual features with decision tree classifier showed a lower performance than other techniques in the literature but this provided a means to assess the relationship and potential of these features. To offer levels of performance closer to those which can be expected from these features in a more advanced system, a set of features were combined and used in conjunction with a neural network classifier for the purpose of classifying shots. This provided a solid baseline to use when comparing them to similar works. The feature set gave favourable results, outperforming other techniques for certain shot classes.

## 1.5 Dissertation Structure

This dissertation is structured as follows: Chapter 2 provides a review of shot classification techniques in the literature, Chapter 3 details the various methods and techniques implemented for achieving shot classification, Chapter 4 presents and discusses the results of the field modelling and shot classification techniques and Chapter 5 concludes this dissertation with a discussion of the various techniques and results and offers avenues for further study.

# Chapter 2

# Literature Review

This chapter reviews relevant techniques for background modelling, shot feature selection and shot classification. As background modelling is a fundamental step in video processing systems this topic has been reviewed in Section 2.1. Three approaches of particular relevance for broadcast sport video have been reviewed viz. colour distance modelling, homogeneous regions detection and Gaussian Mixture Modelling and Histogram Matching and the Code Book approach. Automated shot classification is a very common step which precedes higher level processing of broadcast soccer video because certain shot types, e.g. slow motion or close up, are good indicators that important events have occurred, such as goals or free kicks. A review of various features and classifiers used in shot classification is given in Section 2.2.

## 2.1 Background Modelling

One of the fundamental processes in computer vision is separating areas of interest, the *foreground*, from the rest of the image, the *background*. Common techniques developed for general background modelling are not directly applicable to broadcast video because many of these techniques have been developed in situations where there is a single static camera while broadcast video typically involves multiple dynamic cameras which perform actions such as panning, tilting and zooming [34]. The core assumption of most background modelling techniques, a correspondence between pixel locations and object locations, thus does not hold for broadcast video. Much of the research in detecting events in sports video has attempted to address the problem of adapting or creating background modelling techniques suitable for use in broadcast video [8, 16, 18, 23, 24, 31, 36, 44, 48].

Numerous approaches to background modelling were inappropriate for broadcast soccer video, at least without alteration. Disregarding simple techniques such as frame differencing, many other techniques operate on a per pixel basis, where each pixel is modelled individually, but this is unsuitable for broadcast video because of the camera movement. Common techniques which operate on this basis include the single Gaussian, the mixture of Gaussians, kernel density estimate and the temporal median filter. Common techniques for broadcast video operate on the frame level, whereby the colour distribution is modelled for the frame as a whole rather than per pixel. While it is possible to adapt each for use on a frame level rather than pixel level, certain techniques lend themselves more to this process due to their ability to robustly model a wider array of inputs.

In this section we describe three techniques appropriate for background modelling of broadcast soccer video viz. colour distance modelling which models colour similarity by the distance in colour space; Homogeneous Regions with Gaussian Mixture Modelling that models the probability of model membership with a Mixture of Gaussian Distributions and Histogram Matching & The Code Book Approach which maintains a set of reference colours used to establish the model.

### 2.1.1 Colour Distance Modelling

Many sports are typically played on uniformly coloured playing surfaces such as a soccer fields or tennis courts. One popular and frequently used approach to background modelling in the broadcast sports video domain uses this property by modelling the dominant scene colour instead of modelling the background directly [8, 16, 18, 23, 24, 31, 36, 44, 48]. The reasonable assumption here is that the playing surface colour is the dominant colour of the scenes and that most interesting behaviour occurs on the playing field. Liu et al. [31] describe the play field as having an *"essential role in analysing many kinds of sports"* due to the play field colour dominating most shots in sports video. The importance of play field is further emphasised by [24] as they state that it plays a *"fundamental role in automatically analysing many sport programs"*.

In [44] the field is modelled as the dominant colour and pixels are determined to be field or non-field based on their distance from this dominant colour. The dominant colour is modelled by performing a density estimation using a histogram from appropriate training images or video sequences. Analysis is done in the HSI (Hue, Saturation, Intensity) colour space and a histogram is generated for each colour component. The distance from this dominant colour is determined by the *robust cylindrical metric*[1] and thresholds are set according to a selection schema. Dominant and non-dominant colour classification is determined by the distance from the model and separated by the given thresholds.

[1] *Euclidean distance in the cylindrical co-ordinate system*

A crucial problem which is not addressed by this technique is the proverbial ‘chicken and egg’ situation which affects all background modelling techniques, i.e. it is not known at the time of learning whether foreground is being learnt as background or visa versa [17, 52]. This technique, therefore, relies on the use of preselected sequences which are known to contain dominant field coloured scenes for learning the dominant colour. The use of manual thresholds also makes general applicability across different soccer videos difficult. This technique assumes a *uni-modal* field colour distribution, i.e a single dominant colour. If this is not strictly true, such as when the field contains large sections of shadow (see Figure 2.1), this assumption can introduce inaccuracies in these situations. Updating the background model using this technique is awkward and can decrease rather than improve accuracy. Without an effective updating scheme the performance over an entire video sequence may deteriorate as conditions begin to diverge from the training sequence due to global illumination changes caused by sun sets, cloud cover and stadium lighting.

Ekin et al. [18] improve upon the original approach of [44] in several ways. Two models are established, namely primary and control space, to avoid drift from the dominant colour due to adaptation. Additionally an algorithm is introduced to fuse segmentation results from two colour spaces. This is a similar approach used in [43] where both the HSV and Lab colour spaces are used. However, it is unclear what benefit the Lab colour space provides as it appears to perform significantly worse than the HSV colour space.

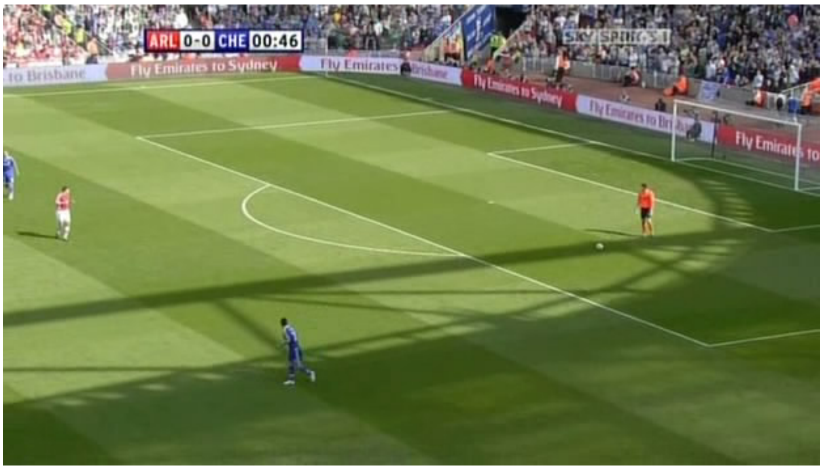


Figure 2.1: Areas of shadow on field

In [45] an approach is used that requires only the Hue and Saturation components of the HSV (Hue, Saturation, Value) colour space. No update scheme is provided and distance is measured using a polar coordinate system. Ignoring the value colour component can reduce the effect of illumination changes, such as shadows, on the performance. This is due to the Value component approximately representing the brightness[2] of a colour. However, because

[2] A notoriously difficult concept to formalise

this approximation is fairly rudimentary this component still contains colour information and thus not including Value in the colour model can reduce overall performance.

Certain approaches, such as the one presented in [50], use the RGB colour space to model the field colour. While this approach is feasible, it is less optimal than working with a transformed colour space such as HSV or HSI. This is due to how colours are distributed in the RGB colour space and how changes in illumination translate a colour from one point to another in the RGB space. Changing the brightness of a colour in the RGB colour space causes it to move along a vector passing through the origin. Thus, to detect an illumination change this vector needs to be calculated and used to determine if the colour shift was merely a result of changing illumination or if this is a different colour all together. The HSV and HSI colour spaces align this vector with their V/I axis which simplifies and accelerates this process. The other difficulty with the RGB colour space lies in how similar[3] colours are distributed. This colour similarity is referred to as the hue of a colour and is a useful quantity when trying to automatically match associated colours. The HSI/HSV colour spaces formally define hue and align it to an axis making comparisons simpler. Because these quantities are aligned to axes in a cylindrical coordinate space, the *euclidean distance* measure becomes a meaningful metric that could be used to measure colour similarity. To reiterate the HSV colour space is described in [29] as *natural and approximately perceptually uniform* while definition of perceptual uniformity is provided in [48] as that property of a colour space where t*wo colours of equal Cartesian distance in the colour space are also equally distant perceptually.*

Due to difficulties in the direct use of the RGB colour space, Yoon et al. [50] choose to manually define the range of required colours used to specify the field region (see Equation 2.1). Here strong assumptions are made about the colours of the playing fields which may not hold for all types of soccer playing fields, specifically that the field's colour is predominantly green. Initial threshold values are also set manually without any initial training and a threshold update algorithm is suggested but not provided. The lack of training or updating can result in decreasing performance during lengthy video sequences. The binarized field mask output image $O(x, y)$ is given by;

[3] Defined by human perception

$$O(x,y) = \begin{cases} 1, \quad if \begin{cases} I_G(x,y) > I_R(x,y) \\ I_G(x,y) > I_B(x,y) \\ |I_R(x,y) - R_{peak}| < R_t \\ |I_G(x,y) - G_{peak}| < G_t \\ |I_B(x,y) - B_{peak}| < B_t \\ GL(x,y) < GL_t \end{cases} \\ 0, \quad otherwise \end{cases} \tag{2.1}$$

> $O(x,y)$ is the binarized output image. $I_R$, $I_G$ and $I_B$ represent the red, green and blue pixel values with $R_t$, $G_t$ and $B_t$ representing the respective threshold values. $GL(x,y)$ refers to the grey level information and $GL_t$ is the respective threshold value. $R_{peak}$, $G_{peak}$ and $B_{peak}$ are the peak component values of the colour histogram. The threshold values of $R_t$, $G_t$ and $B_t$ are manually set to 10, 15 and 10, respectively and are manually adjusted based on colour variance.

The work in [26] builds upon that of [50] by introducing a better threshold selection and updating scheme. The threshold update scheme is modelled after the scheme used in [44] where threshold values are adjusted based on colour component histograms, peak values and standard deviation as;

$$A'_{peak} = \frac{\sum_{H(i) \geq \alpha H(A_{peak})} i \cdot H(i)}{\sum_{H(i) \geq \alpha H(A_{peak})} H(i)} \tag{2.2}$$

$$A_t = std(I_A(x,y)) \ \{(x,y) \in H_A(I_A(x,y)) \geq \alpha \cdot A_{peak}\} \tag{2.3}$$

$$GL_t = GL_{peak} + \beta \cdot std(GL(x,y)) \tag{2.4}$$

> $A = \{R, G, B\}$ represents the set of colour components, $A_{peak}$ refers to the peak value and $A'_{peak}$ is the colour mean value in the vicinity of the peak. $H(i)$ refers to the value of the colour histogram at index $i$. $\alpha$ and $\beta$ are constant coefficients.

This update scheme allows the model to adapt to changing conditions that alter the appearance of the playing field over the course of a video sequence. Because this approach is still based on the model introduced in [50] the same assumptions of colour apply which may be invalidated by certain field conditions.

### 2.1.2 Homogeneous Region Detection and Gaussian Mixture Modelling

One problem with using only colour information is that it can result in noise being included into the background model because all colours in the scene are included in the modelling process without any application of heuristics to determine more likely matches. However, it is possible to exploit further knowledge of this uniformly coloured field by assuming that it constitutes the largest homogeneously coloured region in the image. It is therefore possible to be more selective of candidate field regions by performing connected component analysis and filtering regions which are considered too small to be regarded as field regions. By removing areas which are unlikely to contain field pixels, fewer non-field colours are learnt into the model thereby increasing model accuracy. Variations of this approach have become popular and can be seen in literature (see [17, 24, 31, 39, 40, 48]).

Duan et al. [17] make use of motion analysis to further reduce possible sources of noise by eliminating shot classes which are unsuitable for field colour learning, such as those shots whose largest components may not be associated with the playing field. The background model is based on the mixture of Gaussians technique which is popular with many traditional background modelling approaches [33, 38, 47] and is commonly known as *Gaussian Mixture Model* (GMM). The average colours of candidate field regions are modelled by $K$ Gaussian distributions. Foreground classification occurs by determining a candidate's distance to the mean of the nearest Gaussian distribution and thresholding based on the standard deviation of that Gaussian. This approach does not require pre-selected training samples and the model is easily updated to reflect changes in the environment by altering or replacing the Gaussian distributions. However, certain techniques used, such as the *mean shift procedure* and connected component analysis, are computationally expensive and may not be appropriate for certain applications. The mean shift procedure [15] is used to cluster the colour information before being processed by a connected component algorithm. Unfortunately this procedure incurs a high computation cost when requiring the level of clustering needed to produce a useful output by a connected component algorithm. The cost of connected component analysis can be reduced significantly by using the contour tracing technique presented in [12].

Liu et al. [31] offer a simpler approach by applying the (4-)connected component analysis directly on a 2-Dimensional colour histogram to establish peaks generated in the CbCr colour space. Depending on quantisation levels, this significantly reduces the computational cost associated with both the clustering and connected component analysis used to establish the homogeneous regions. The dominant colours are then modelled using a GMM and updated using an incremental EM algorithm. Here, the YCbCr colour space is used instead of the HSI space as their results show improved performance when the 2D Colour histogram is constructed form the CbCr components over the H, I components. No filtering takes place to establish more suitable candidates but due to the use of a 2-dimensional histogram the correlation between these colour components are maintained, as opposed to the typical assumption of independence used for separate channel analysis. This naturally lends itself to reduced noise from other scene colours.

### 2.1.3 Histogram Matching and The Code Book Approach

Instead of modelling the background/field colour directly, Zhong and Chang [52] attempt to identify whether a scene displays the appropriate colour characteristics to be classified as a dominant play-field coloured scene. This approach is based on 3-dimensional colour histograms used to characterise a scene. $K$-means clustering, applied to manually selected training data, is used to generate $K$ representative feature vectors. Initially a large number of colour models are included and then narrowed down by adapting to the initial portion of the video. The adaptation process involves an initial filtering threshold and then an object-level verification process. Similarly Coldefy and Bouthemy [14] use colour histograms computed in the CIE Lab colour space along with an on-line $K$-means clustering algorithm to determine the presence of a green dominant colour scene. This approach works well within the context of their research but it does not provide adequate information with which to perform further feature extraction. Lin and Zhang [29] establish a shot correlation measure via the use of quantised 3-dimensional colour histograms in the HSV colour space and what they describe as colour objects i.e. significant clusters in the HSV space of a frame formed by a scene's dominant colours. Unfortunately due to the strong colour correlation between shots in sports video it becomes difficult to apply such a method. In [37] colour histogram differencing based on multiple timescales is used to detect scene changes, the details of which however, are not present due to its implementation and subsequent use of a third party annotation tool (IBM VideoAnnEx[4]). In a survey on automatic video classification [11] the use of histograms to detect shot or scene changes is fairly common.

In [41] a colour code book is created to define all colours which are considered field colours. A green colour table is manually constructed defining all permissible field colours. Then a training video sequence is processed and all colours in the upper half of each image in the sequence which match those in the green colour table are kept in another table referred to as the upper green table. A similar table is constructed for the lower half of each image in the sequence and the matches are stored in a lower green table. The target sequence is then processed and colours are compared to those stored in the lower and upper tables and determined to be either green or non-green. Unfortunately, the algorithm used to determine colour similarity is not given. The need to manually construct a table of

[4] The VideoAnnEx annotation tool assists authors in the task of annotating video sequences with MPEG-7 metadata.[3]

expected field colours may be impractical for many applications as the field conditions can vary dramatically from sequence to sequence. This approach also does not include any form of update scheme.

With the review of background modelling concluded, the discussion now turns to the features and classifiers used in the shot classification process.

## 2.2 Shot Classification

This section reviews the two components of the shot classification process, namely the shot features and the classification techniques. The features identify shot properties which are then used by classification techniques to identify type of shot.

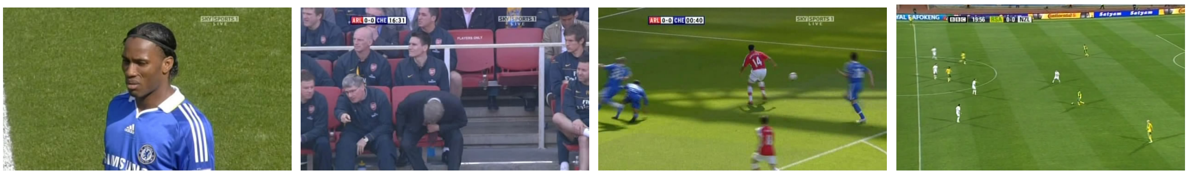
### 2.2.1 Features for Shot Classification

An important first step in classifying shots is to determine the shot classes that are semantically relevant and separable. Tekalp and Ekin [44] propose three shot classes; long, in-field medium and out-of-field/close-up shots (see Figure 2.2). Long shots contain the global field view and are useful for shot localisation. Medium shots provide a zoomed in view of a specific part of the field. Close-up or out-of-field shots describe a single person or audience shot. Classifying shots in this manner is useful due to the nature and context of their usage in a typical broadcast which facilitate the extraction of certain semantics. Close or audience shots are good indicators of important events such as goals or free kicks. Medium shots are often used to highlight interesting activity such as dribbling or tackling. Long shots are useful for activity localisation i.e. identifying the area of the field in which interesting activity is currently happening. This can be used to infer probabilities of the current state of events, e.g. which team is attacking or if a set piece is taking place. The field to non-field ratio is the primary feature used to determine shot classes and is sufficient to separate long shots and close/out-of-field shots. Medium shots, however are typically more difficult to classify or define, especially when compared to certain types of close-up shots.

Field colour ratio is widely used, in whole or in part, in many shot detection and classification algorithms [18, 23, 36, 37, 39, 48, 49]. The global field ratio and subsection ratios based on sections marked by the golden section spatial composition rule[5] are thresholded and passed to a Bayesian classifier. Figure 2.3 shows a frame segmented using golden section

[5] The image is divided in a ratio of 3:5:3 in both the vertical and horizontal directions

spatial composition, with the middle sections being used for processing. By processing only a specific subsection of each image certain issues are avoided such as interference from out-of-field sections in long/medium shots (see Figure 2.4a). Sub-sectioning experiences problems in instances with low angle medium shots or other situations such as a shot which is not framed by the playing field (see Figure 2.4b).



(a) Close-up (b) Out of field (c) Medium (d) Long

Figure 2.2: Shot classes proposed by [44]

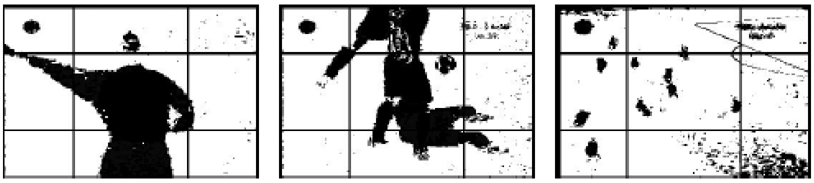
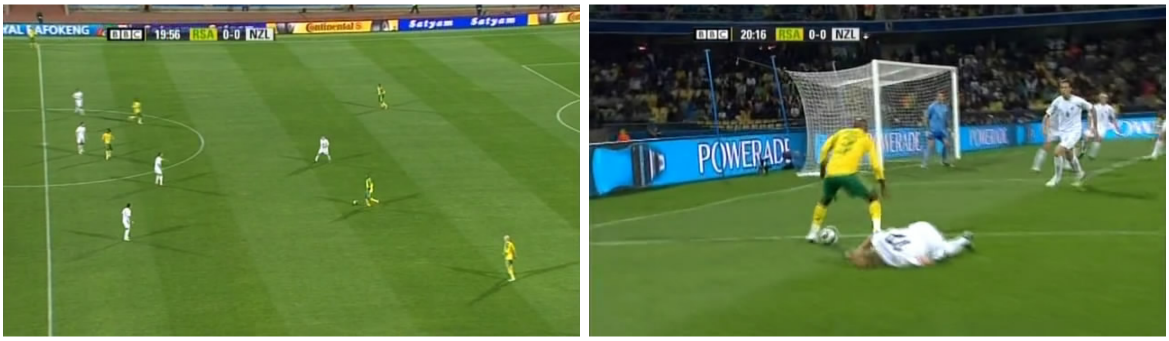


Figure 2.3: Images showing golden section spatial composition from [44]

Tong et al. [45] offer similar shot classes but use a broader set of classification features. Global field ratio separates shots containing a field background such as long and medium shots with shots such as out-of-field and close up shots. Using only field ratios does not produce sufficient accuracy when classifying medium and close-up shots therefore further features are needed. The existence of a head area along with texture features are used at the next level of separation to improve the accuracy with which medium and close-up shots can be classified. Here, the presence of a large central head area is a strong indicator for a zoomed in single person shot. However, potential for further misclassification may arise due to the large variance in human skin tones and thus the difficulty in separating skin tones from the vast array of other colours present in a broadcast video. A more robust facial detection algorithm such as those based on principle component analysis or Gabor filters would greatly improve detection accuracy compared to the simple skin tone comparison. This would unfortunately result in significantly increased computation complexity which, depending on the application, may not be tolerated. Finally, object size ratios help distinguish medium and long shots. Simple object detection is used to generate an estimate of object sizes in relation to the field area. By assuming object scales are associated with the zoom level of a shot an inference about a shot's type can be made. Shots with large object ratios have a high probability of being medium shots, similarly those with low ratios have a high probability of being long shots. Replay shots form a shot meta-class which are detected by identifying shot transitions where a logo is present. The detection of replay or slow-motion shots provides a reliable indicator for semantically important events in a video sequence. Detecting replays this way performs adequately in certain instances but there are many instances where replays are not bounded by logo transitions therefore invalidating this method for general application.



(a) Players framed by playing field

(b) Players not framed by playing field

Figure 2.4: Examples showing framed and non-framed shots

Sun et al. [41] introduce the idea of visual keywords which are claimed to have greater semantic linkage than shot classes. Shots are first segmented using motion analysis into active and static parts. Background modelling is then used to segment field and non-field regions. Each shot is segmented into four equally spaced rows. Shots are classified as one of four classes: green non-dominant, green dominant, green partially or field with player. Classification is based on the instances of green and non-green colours in each of the four rows.

Field of view is a strong feature which can be used to classify shots but this information is not readily available from the video sequences. [45] and [17] both offer techniques for field of view estimation. In [45] simple object detection is performed and the field to object height ratio is used to estimate the relative size of the objects and thus the field of view. Duan et al. [17] calculates the field ratio along each column of the image generating a 1-dimensional field-player interaction curve (see Figure 2.5) from which additional descriptors are extracted and used to estimate the depth of field by reasoning about the size and location of non-field elements in the image.

These features form a good basis for identifying scene properties but they can vary greatly from game to game requiring thresholds and classifiers to be trained per game. Alternatively, complex field properties can suffer significant decrease in accuracy with a poor performing background model. Therefore it becomes beneficial to combine colour and motion analysis, a popular approach seen in literature (see [8, 7, 16, 17, 52]). The inclusion of audio features as seen in [14, 28, 43] can also increase performance and accuracy, however, audio features are typically only introduced for mid to high-level analysis where scene and event detection

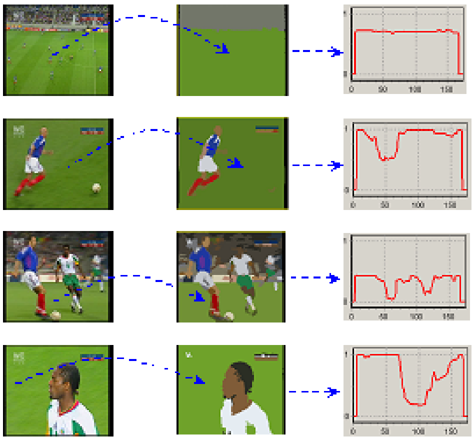


Figure 2.5: 1-Dimensional field player interaction curve from Duan et al. [17]

occur.

The goal of the shot features is to identify properties or characteristics of a shot which allows the identification of the shot's type or class. These features are processed by a classifier for the purpose of classifying a shot's type. The discussion of the classification process follows.

### 2.2.2 Classification

Finding the best features to use for shot classification is a difficult and as yet unsolved problem and the generation and extraction of these features is still the main focus of research. This often leaves the analysis and processing of these features as secondary considerations. Relatively few papers such as [16] provide a more detailed look at the classification process. Without a standardised, or at least 'best practise', feature set, papers like [46] which focus too heavily on higher level concerns can become too superficial for practical usage.

Due to this lack of maturity in the feature space, shot classification methodology is often basic. Techniques which rely on domain and a priori knowledge such as decision trees [17, 30, 44] are often used instead of adaptive techniques such as Bayesian classifiers [7, 18], support vector machines [16, 36] and neural networks [8]. There is generally little discourse around the selection of technique and its possible advantages or disadvantages over similar techniques. The variety of features used in literature makes it difficult to delineate between the strength of the feature set over the strength of the classification technique. Furthermore because of the often widely varying conditions between games, techniques which require significant training may require this training for each game and thus may become unsuitable for many applications.

In [18] and [44] a Bayesian classifier is used within the decision tree structure to separate a subset of shot classes, specifically the medium and long shot types, using as its feature set the three ratios generated by a golden section spacial decomposition of the shot. Due to the classifiers position and incorporation into the decision tree, its performance is heavily influenced by the decision tree itself.

Nan et al. [36] use a series of SVM classifiers to classify four different shot types. Long and medium shots are classified based on dominant colour features while close-up and out of field shots use an additional set of features based on texture and edge information. SVM classifiers also prove popular when analysing audio features, a topic not discussed here but nevertheless popular, especially in combination with visual features as seen in [16].

Assfalg et al. [8] use two neutral network classifiers in conjunction with numerous features to classify a broad range of shots, unfortunately the details of the classifiers are not elaborated.

## 2.3 Summary

Three types of background modelling approaches have been discussed: Colour Distance modelling, Homogeneous region detection with Gaussian mixture modelling and Histogram matching & the Code Book approach.

Colour Distance modelling uses colour spaces and distance measurement to construct a model of the dominant scene colour typically representative of the play field area. This is a largely autonomous approach as only the training set and a limited number of parameters need be manually selected. The approach is relatively computationally efficient and provides a modular platform which allows the various sections to be easily substituted and experimented with. For example the choice of colour space and distance measurement can be easily changed.

The homogeneous region detection and the Gaussian Mixture model is based on colour probability distributions created using a Mixture of Gaussian distributions. No training data need be supplied for the models generations which lessens the requirement for manual intervention. However, model parameters still require adjustment which is often performed manually. This approach however, is relatively computationally expensive compared with the others discussed and even more so when using multiple colour channels which increases the dimensionality of the distribution. This approach also requires computationally expensive pre-processing in order to generate the homogeneous regions for input into the model but a strong update schema provides the model with greater adaptability to changing conditions within the video sequence.

Histogram matching and the Code Book approach uses as its model a set of reference colours and/or colour histograms which is used to determine scene type. One problem with this approach is that no background mask is generated and thus it is not a feasible approach to use when further processing is required using the background mask as input. Furthermore, using these techniques to separate shots with the required fidelity becomes difficult due to soccer video sequences having a high colour correlation between shots.

The colour distance model and the Gaussian mixture model show the most potential for generating an accurate and robust background model for use in a shot classification system.

The shot classification process has been discussed in two parts; the features used for shot classification and then the classification techniques which use these features to identify and classify shots.

The various shot types typically defined in the literature, such as long, medium, close, audience and out of field shots were described. The relevance of each shot is discussed along with why the identification of the shot type is useful.

A basic yet popular shot feature is that of the field ratio, of which two variations are discussed, the global field ratio and the golden section field ratio. Field ratios are reliable for long and out of field shots but for medium and close shots, further features are required such as texture, motion or the identification of a person's head. A category of features attempt to identify the field of view of a shot which can be closely related to the the shot's type. These features include the object size ratio feature and the 1-D field/player interaction curve. The object size ratio identifies the objects in the shot and their size relative to the field size to estimate the field of view. The 1-D field/player interaction curve establishes a vertical projection of the field masks and extracts information from this projection for the purposes of identifying various shot properties.

Four shot classification techniques have been described: the decision tree, the Bayesian classifier, the SVM classifier and the neural network. The most widely used of the four is the decision tree which can operate in conjunction with SVM and Bayesian classifiers.

# Chapter 3

# Methodology

In this chapter various methods and tools for automatic shot classification in broadcast soccer video are motivated and described. Through the presentation of background modelling and its goal of modelling the play-field, a foundation is established from which to extract features necessary for shot classification. We discuss the features extracted from this field model and explain how and why the characteristics of these features are used for shot classification. Finally the classification process itself is detailed by examining how the various features are combined to produce a classification.

## 3.1 Field Colour Modelling

The modelling of a scene's background is a common procedure in image processing which is often used as a foundation for more complex processing and analysis. Most applications focus on a static camera arrangement. For broadcast video, however, it becomes necessary to adapt the approach to background modelling when operating in environments with dynamic camera movement. This is because moving cameras result in a moving background, something that is typically not tolerated in traditional background modelling techniques. A common approach used to address this issue, one typically associated with sports video processing, is to forgo attempts at modelling a 'generic' background and instead focus on modelling the play field area. This is the approach we have chosen to pursue. The field area is useful for numerous reasons, the most important of which is that it often contains the majority of the interesting activities one would wish to analyse in a sports video. By applying heuristics, the field model is significantly easier to generate than a traditional background model, especially when considering moving cameras and changing points of view. This is because there is no longer a reliance on pixel location to object location correlation, otherwise referred to as a spatio-temporal consistency. Hence it is possible to generate a reasonable model of the play field area given any number of different view points that may be present in a video sequence. Additionally [29] found colour features to be effective yet computationally inexpensive when establishing shot similarities making them highly appropriate for our application by ensuring appropriate levels of accuracy and performance.

Two colour modelling approaches have been explored in this work, the colour distance model and the Gaussian mixture model. Section 3.1.1 details the colour distance modelling method and covers the topics of colour models, density estimation, dominant colour estimation, threshold selection and finally connected component labelling and filtering. Section 3.1.2 details the Gaussian mixture modelling method and covers the topics of connected component analysis, clustering techniques and the Mixture of Gaussians technique.

### 3.1.1 Colour Distance Modelling

The colour distance approach to modelling the play field area leverages the idea that the similarity of two colours is determined by the distance between them. The core concept behind this colour distance, and its representation of similarity, is the idea of a colour space. A colour space is the subspace generated by the components of a colour representation for a particular colour model. Three is the typical number of components for most colour spaces although the dimensionality of the colour space does not alter the principles of the model. An individual colour becomes a point, or coordinate, in this subspace therefore making it possible to compare any two colours by measuring the distance between them. Subsequently a field model was created by establishing the coordinates of the primary field colour and then determining appropriate distance thresholds to separate colours from those belonging to the model and those which do not. If the colour space used were 3 dimensional and used a cylindrical coordinate system then one could visualise the model as a pie slice in this space. Figure 3.1 shows the generation procedure used for the colour distance model where training data is input and used to locate the dominant colour and establish a surrounding region where a colour is accepted as belonging to the model. Naturally, it becomes important to select a colour model which is able to maintain a high level of discriminatory power under the various shot and illumination changes typically present in broadcast soccer video.

The discussion begins with the motivation for the selection of the colour space model, which is used to generate the field model. This is followed by the discussion of the density estimation technique which was used to establish the most frequently occurring colours. This then leads into how the dominant colour is selected and what methods were used to select the thresholds for segmenting the dominant and non-dominant colours. The connected component filtering process is covered last.

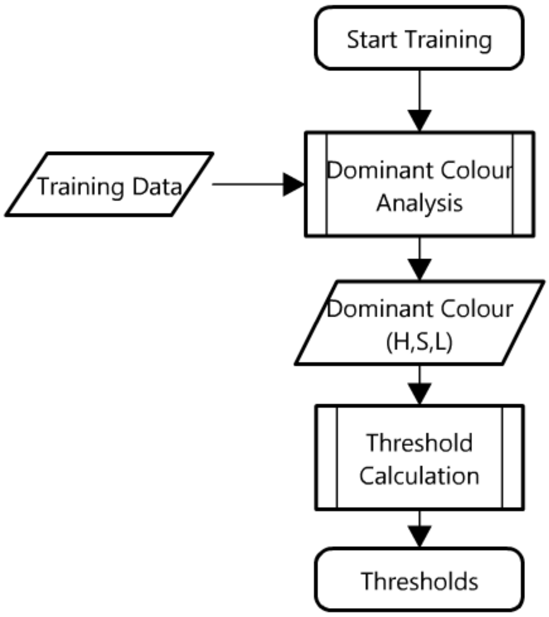


Figure 3.1: Colour Distance Model Generation

#### 3.1.1.1 Colour Models

A common method for representing colours for capturing and displaying digital images, used in devices such as digital cameras and display monitors, is the Red, Green and Blue (RGB) colour model. Its popularity stems from its similarity to the mechanical operation of the human eye and that of imaging sensors. The RGB colour model generates a subspace in the shape of a cube which can be visually represented as a colour solid (see Figure 3.2).

One of the problems with the RGB colour model is that it does not correspond well to how humans classify colours since we do not consider a colour to be merely a linear combination of primary colours [20]. Considering that humans are significantly superior at processing visual data than machines it seems prudent for the purposes of image processing to construct a model closer to that used by humans. The set of colour models which replicate this behaviour are known as perceptual colour models and include the Hue, Saturation and Value (HSV), the Hue, Saturation and Intensity (HSI) and the Hue, Saturation and Lightness (HSL/HLS) colour models. These models are not independent representations but are instead linear transformations of the RGB colour space that are unfortunately not standardised. Thus the specific transformation used needs to be specified and vary from one implementation to another. The advantage of HSV/HSI/HSL is that they attempt to approximate perceptual uniformity to ensure distances between two colours represents the perceptual difference between them [48]. This results in certain desirable characteristics such as those found by [21] that when images are contaminated by highlights, the hue colour component achieves the best performance. Even under consistent lighting conditions, hue is found to perform equal to normalised colour spaces while still being superior to the RGB colour space. This proves useful in analysing soccer video due to the frequent occurrence of shadows and highlights created by the various light sources at stadiums.

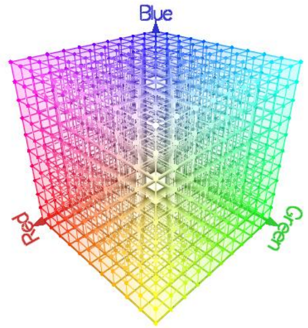


Figure 3.2: Visual representation of the RGB colour space

In this work, the colour distance modelling approach is based on the work done in [44] and uses the HSL colour space, with the specific RGB to HSL transformation defined[1] in equations 3.1, 3.2 and 3.3 (see Figure 3.4). The transformation from the red (R), green (G) and blue (B) colour space to the hue (H), saturation (S) and lightness (L) colour space is given as:

$$V_{max} = max(R, G, B)$$
$$V_{min} = min(R, G, B)$$

$$L = \frac{(V_{max} + V_{min})}{2} \tag{3.1}$$

$$S = \begin{cases} \frac{(V_{max} - V_{min})}{(V_{max} + V_{min})} & if\ L < 0.5 \\ \frac{(V_{max} - V_{min})}{(2 - (V_{max} + V_{min}))} & if\ L \geqq 0.5 \end{cases} \tag{3.2}$$

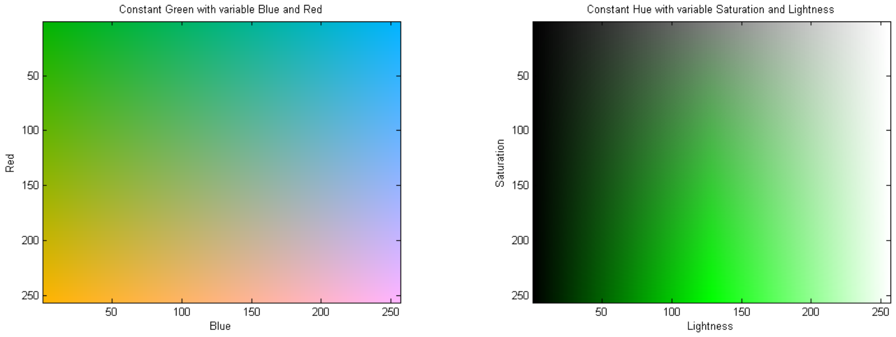
$$H = \begin{cases} \frac{(G-B)*60}{S} & if\ V_{max} = R \\ 180 + \frac{(B-R)*60}{S} & if\ V_{max} = G \\ 240 + \frac{(R-G)*60}{S} & if\ V_{max} = B \end{cases} \tag{3.3}$$

[1] Based on the OpenCV implementation [5, 6]

**Hue** describes the colour according to its similarity to the perceived colours of red, green and blue or a combination of two.

**Saturation** describes the pureness of a colour or level of chromacity. Achromatic colours being unsaturated, while pastel to primary colours are fully saturated.

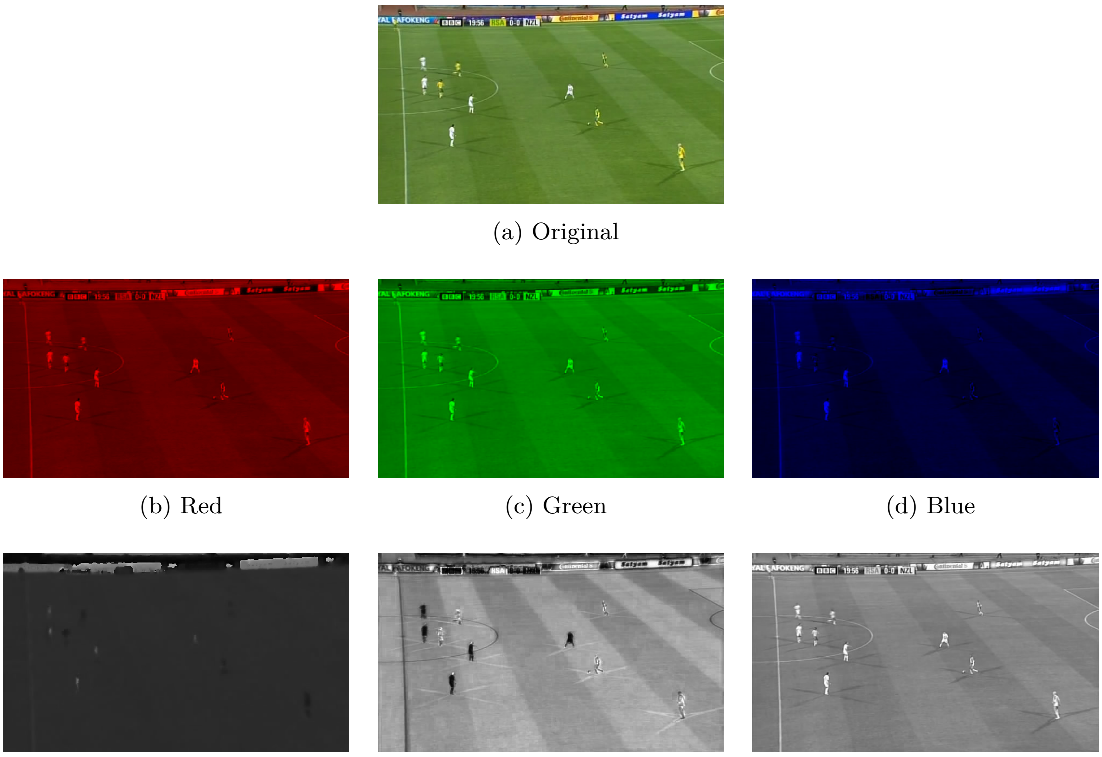
**Lightness** is the measure of light intensity.



(a) Constant Green

(b) Constant Hue

Figure 3.3: RGB and HSL colour spaces projected on to the GB and SL planes respectively



(a) Original

(b) Red

(c) Green

(d) Blue

(e) Hue

(f) Saturation

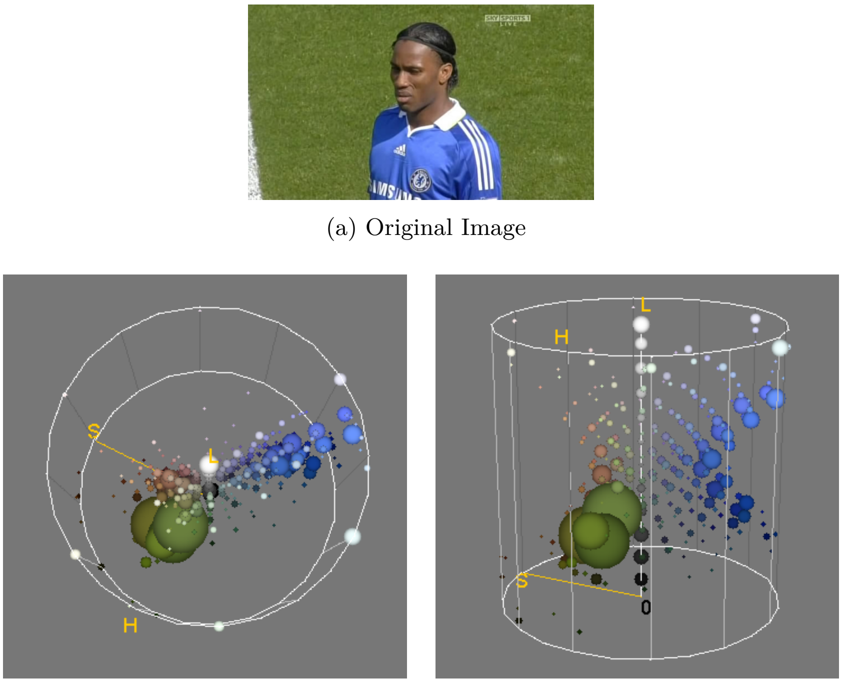
(g) Lightness

Figure 3.4: HSL colour space component separation

#### 3.1.1.2 Density Estimation

After establishing a model with which to represent colours the next step involved constructing a method to establish the dominant colour of any given frame. The dominant scene colours corresponds to the colours with the highest probabilities of occurrence. Thus to identify these colours, knowledge of the probability density function (pdf) of the scene colours is required. The pdf however, is unknown at the time of processing and therefore a density estimation is required to be generated from the observed data. For this purpose we employed a popular density estimator, the histogram. The histogram is used to estimate the underlying pdf of the frame colours from the observed data (pixel values). Thus the histogram can be used to determine the dominant scene colours.

Because individual colours are represented as a triple of strongly correlated colour components, a full colour histogram results in a 4-dimensional construction. The complexity of such a construct is illustrated in Figure 3.5. This would result in 16777216 ($256^3$) histogram bins resulting in a prohibitively costly computation per frame. However, if the assumption is made that the components are independent, and thus there is no correlation between them, we can generate density estimates for each component individually (see Figure 3.6) thus reducing the total bins to 768 ($256 \times 3$). Of course this assumption is inaccurate as the components are not independent but it provides a useful approximation for speeding up calculations. If further performance gains are required simple clustering can be performed to reduce the number of component bins and thus the total number of bins. This did not appear necessary for this work thus a full set of three 'independent' 1-D component histograms to represent the colour density were used. It is possible to create a compromise and use a 2-D histogram as used by [31] however this still adds substantial computational complexity when establishing peaks for dominant colour selection and was therefore not considered.

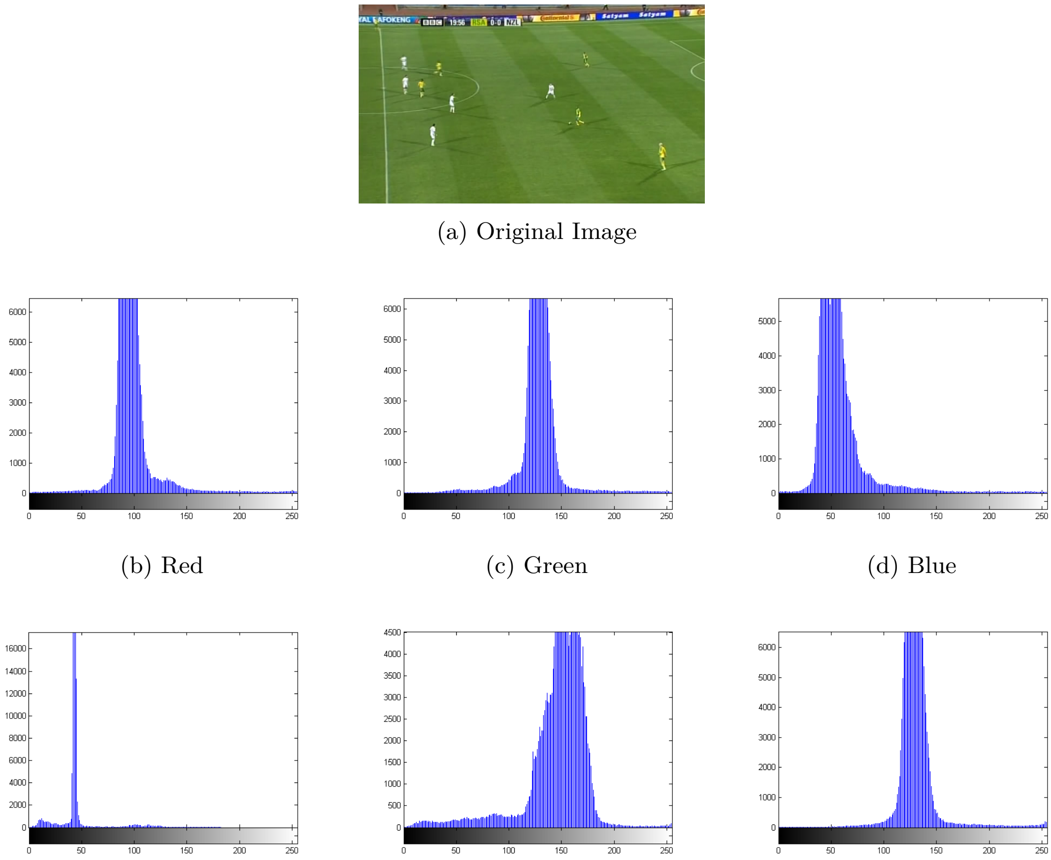


(a) Original Image

(b) Top down view of HLS colour histogram

(c) Side view of HLS colour histogram

Figure 3.5: 3-D colour histogram visualisation generated by 3D Color Inspector/Color Histogram [1]



(a) Original Image

(b) Red

(c) Green

(d) Blue

(e) Hue

(f) Saturation

(g) Lightness

Figure 3.6: Component histograms

#### 3.1.1.3 Dominant Colour Estimation and Threshold Selection

By using the histogram for density estimation the probability distribution for a given frame was established, but to generate a more accurate and reliable estimation of the dominant scene colour of a video sequence, multiple frames and their respective histograms were analysed.

Algorithm 3.1 [18, 44] was used to determine the dominant scene colour by using the HSL colour space and histogram density estimation. The mean values of the highest represented colours ($H_{mean}$, $S_{mean}$, $L_{mean}$) were established for each colour component, and thus represented the dominant scene colour. This was done by creating a combined histogram representing a number of frames and establishing the set of bins to represent the dominant colour. The set was established by first considering the peak bin for each colour component and including all adjacent bins which fit a criterion. The criterion used for a bin to be included as part of the dominant colour was that it should be no less than $K\%$ of the peak bin value, where $K$ is the predetermined threshold. This technique assumes a uni-modal distribution for each colour component which does not hold true for every situation, but, for this work inaccuracies resulting from this assumption proved an acceptable compromise to maintain high performance.

Subsequently considering the set of means ($H_{mean}$, $S_{mean}$, $L_{mean}$) as a co-ordinate in the cylindrical co-ordinate system, the euclidean distance was used to measure the colour similarity of a pixel's colour from the mean by leveraging the perceptual uniformity of the HSL colour space. Algorithm 3.2 [18, 44] processed a training set of images to find appropriate threshold values to use for classification. The dominant/non-dominant classification was made by referencing these threshold values. Due to the varied conditions of soccer matches and the extent of illumination changes present in different videos, a leniency factor $\alpha$ was required to scale threshold values so as to ensure an acceptable level of accuracy while maintaining high recall rates.

**Algorithm 3.1** Find Dominant Colour [44]

1. Convert input from RGB (default) to HSL using Equations 3.1, 3.2 and 3.3.
2. Generate histograms ($H$) for each colour channel separately.
3. Establish parameters for each colour channel assuming a uni-modal distribution.
    (a) Find peak histogram bin ($i_{peak}$).
    (b) Find smallest ($i_{min}$) and largest ($i_{max}$) bin equal to $K$% of peak.
        i. Check all bins with indices equal to or lower than the peak ($i_{min} \leq i_{peak}$) and find bin that satisfies equation 3.4. This establishes the lower bin index, $i_{min}$.
        ii. Check all bins with indices equal to or higher than the peak ($i_{max} \geq i_{peak}$) and find the bin that satisfies equation 3.5. This establishes the upper bin index, $i_{max}$.

$$H_{i_{min}} \geq K \times H_{i_{peak}} \; and \; H_{i_{min}-1} < K \times H_{i_{peak}} \tag{3.4}$$

$$H_{i_{max}} \geq K \times H_{i_{peak}} \; and \; H_{i_{max}+1} < K \times H_{i_{peak}} \tag{3.5}$$

$$i_{min} \leq i_{peak} \; and \; i_{max} \geq i_{peak} \tag{3.6}$$

    (c) Calculate mean.

$$mean(H) = \frac{\sum_{i=i_{min}}^{i_{max}} i \times H_i}{\sum_{i=i_{min}}^{i_{max}} i} \tag{3.7}$$

**Algorithm 3.2** Calculate Distance Thresholds[44]

1. Convert input frame, $x$, from RGB to HSL

2. For each pixel, $i$, in frame, $x$, with colour components $H_i$, $S_i$, $L_i$, calculate the euclidean distance from dominant colour $H_{mean}$, $S_{mean}$, $L_{mean}$.

$$\theta = \begin{cases} |H_{mean} - H_i| & if \quad |H_{mean} - H_i| < 180^\circ \\ 360^\circ - |H_{mean} - H_i| & if \quad |H_{mean} - H_i| > 180^\circ \end{cases} \tag{3.8}$$

$$d_L(i) = |L_{mean} - L_i| \tag{3.9}$$

$$d_{polar}(i) = \sqrt{S_{mean}^2 + S_i^2 - 2S_iS_{mean}\cos(\theta)} \tag{3.10}$$

$$distance(i) = \sqrt{d_L^2(i) + d_{polar}^2(i)} \tag{3.11}$$

3. Calculate mean distance.

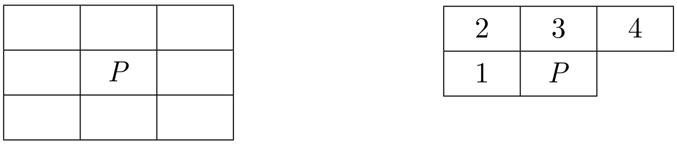
$$meanDistance(x) = \frac{\sum_{i=0}^{i_{max}} distance(i)}{i_{max}} \tag{3.12}$$

4. Set threshold as $\alpha\%$ of mean where $\alpha$ is a leniency factor and has been determined from domain knowledge.

$$threshold = \alpha \times meanDistance(x) \tag{3.13}$$

#### 3.1.1.4 Connected Component Labelling

Once the basic background model was generated by the previously described techniques the model was further refined by performing a connected component analysis. Using heuristics, components were filtered based on size and shape to eliminate unsuitable regions. Components which were too small were filtered as they are likely to be as a result of field colours present in objects other than the field. Initially a basic single pass depth-first connected component algorithm was implemented but the performance was unacceptable for this work. A more efficient two pass algorithm, modified from the multi-pass algorithms described in [42], was used.



(a) 8-connected neighbourhood

(b) Scan mask

Figure 3.7: Pixel neighbourhood

**First Pass.** The binary image is scanned sequentially from top left to bottom right. For each non-background pixel $p$, the neighbouring pixels are scanned according to the scan mask ($M = [1, 2, 3, 4]$); in Figure 3.7b to determine connected regions. Only provisional labels are assigned during this pass with the assignment procedure as follows;

$$L[p] = \begin{cases} 0, & if\ I[p] = 0 \\ l, (l \leftarrow l + 1), & if\ I[i] = 0,\ for\ all\ i \epsilon M \\ T[\{ \min\limits_{i|I[i]=1} (L[i]) | i \epsilon M \}], & otherwise \end{cases} \tag{3.14}$$

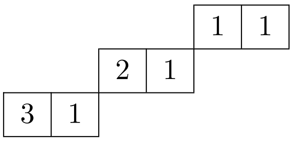
For further explanation Equation 3.14 is translated into Algorithm 3.3.

**Algorithm 3.3** First pass connected component labelling

```
1. Begin iterating through all pixels $p$ in image $I$ using scan mask $i$ to assign labels $L$,
   initializing label counter $l$ as 1.
   (a) Case 1: $I[p] = 0$ (Pixel represents background)
       i. Assign label of zero
          $L[p] = 0$
   (b) Case 2: $I[i] = 0$, $\forall i$ (No foreground neighbours in scan mask)
       i. Assign new label
          $L[p] = l$
       ii. Increment label counter $l$
          $l \leftarrow l + 1$
   (c) Case 3: (A foreground neighbour is present in scan mask)Assign label as smallest
       equivalent label determined by equivalence table $T$
       $L[p] = T[\min_{i|I[i]=1} (L[i])]$
```

The first case for labelling a pixel is when the pixel represents the background, $I[p] = 0$, in this case a zero or null value will always be assigned. The second case represents a scenario where all pixels in the scan mask are background, $I[i] = 0$, therefore a new label is assigned to the current pixel and the label counter is incremented, $l, (l \leftarrow l + 1)$. Otherwise if the pixel is not a background pixel and the scan mask contains labelled neighbours then the pixel label must be assigned the equivalence label of the minimum neighbouring label. The equivalence table $T$ tracks labels found to be equivalent thus $T[i]$ represents the smallest label equal to $i$.

**Analysis.** After the first pass is complete all pixels have been assigned a provisional label and an equivalence table has been constructed. At this point the equivalence table is still incomplete as only 'local' relationships have been established, meaning a label's equivalence may not have been fully propagated [42]. Figure 3.8 shows the state of both the labels and the equivalence table after the first pass has been completed. In this instance the labels have fully propagated as both labels 2 and 3 have been found equal to label 1.



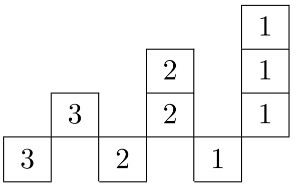
(a) Labels

| i | 1 | 2 | 3 |
|---|---|---|---|
| T[i] | 1 | 1 | 1 |

(b) Equivalence table

Figure 3.8: Connected component labelling procedure, first pass completed and labels fully propagated

This is not typically the case however and often labels are not fully propagated, as demonstrated in Figure 3.9. To correct this issue and ensure full propagation of labels the equivalence table is processed and labels are back propagated such that the resulting table reflects the lowest equivalent label. This procedure is shown in Algorithm 3.4.



(a) Labels

| i | 1 | 2 | 3 |
|---|---|---|---|
| T[i] | 1 | 1 | 2 |

(b) Equivalence table

Figure 3.9: Connected component labelling procedure, first pass completed and labels not fully propagated

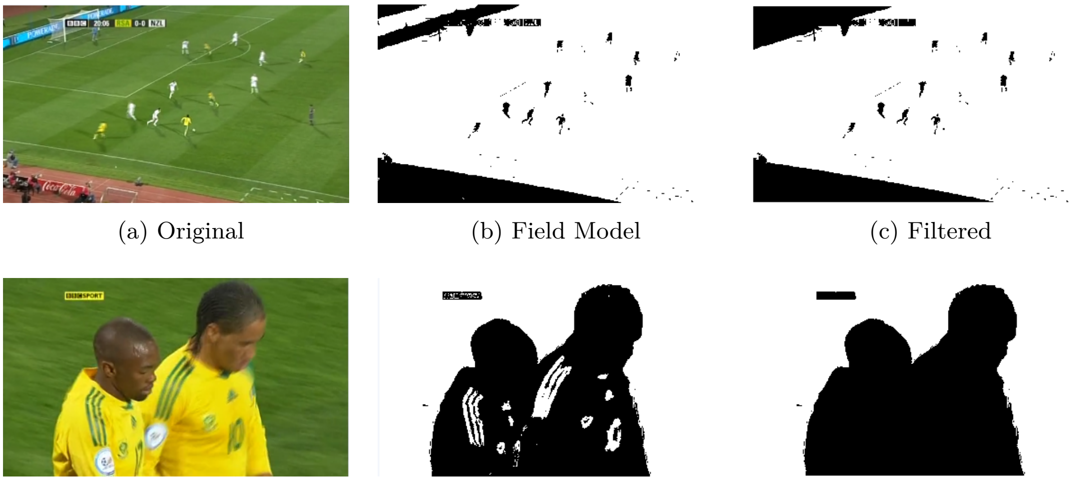
**Second Pass.** The second pass is simple: iterate through the labels and change each label to its lowest equivalent. However, using only the above procedures the final set of labels

**Algorithm 3.4** Equivalence Table Label Propagation

```
for each $label$
    if $T[label] \neq label$
        $T[label] \leftarrow T[T[label]]$
    end
end
```

are often non-sequential which can be inconvenient. A few extra steps can be added if sequential labels or additional label data such as label counts, centroids or bounding boxes are desired.

After the connected component labelling procedure has completed, the components are filtered based on size and shape (aspect ratio only) and components found outside acceptable parameters are removed from the background model. The results of this procedure can be seen in Figure 3.10 from which we can see that small non-field regions which were initially labelled as background in Figures 3.10b and 3.10e have been correctly filtered and relabelled as foreground in Figures 3.10c and 3.10f.


(a) Original (b) Field Model (c) Filtered

(d) Original (e) Field Model (f) Filtered

Figure 3.10: Field colour model and connected component filtering

### 3.1.2 Gaussian Mixture Modelling

The Gaussian mixture model is a popular technique often used for background modelling. Traditionally this model operates by modelling each pixel individually as a mixture of Gaussian distributions. This forms a probability density function which is used to determine the probability of a pixel value belonging to the background. Due to the lack of a fixed camera view it is not possible to use this technique on a per pixel level for this work, thus instead of modelling individual pixels the Gaussian mixture models the field colour.

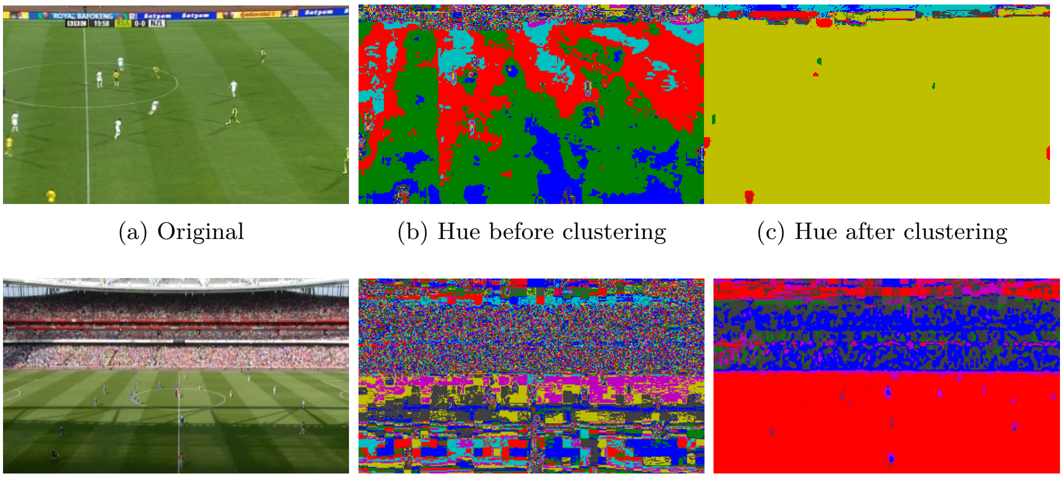
The discussion begins with the pre-processing techniques used to simplify the modelling process. This is followed by detailing the on-line Gaussian Mixture learning process used to generate the Gaussian Mixture model.

#### 3.1.2.1 Connected Component Analysis and Clustering Techniques

Before constructing the model using a mixtures of Gaussians, connected component analysis was performed to isolate probable field candidate regions. This improves model accuracy by preventing most non-field regions from being introduced into the model by using heuristics to rank regions based on their probability of containing field pixels. The ranking was based on the size and shape of the regions, with regions falling outside of the expected aspect ratios of a field region being filtered out and the remaining regions ordered by size with the largest three regions being included into the model.

The connected component analysis was however performed before the field/non-field mask had been calculated, thus the analysis was not performed on a binary image but rather a grayscale Hue channel image. This meant that some form of clustering was required prior to determining connected components. To simplify the process of establishing connected components, similarly valued pixels are clustered together to form homogeneous regions. This reduces the total number of components and provides a better segmentation of components. Some approaches have used the mean shift procedure on the hue colour channel to cluster pixels. To attain a sufficient level of clustering however, the mean shift procedure becomes computationally expensive.

Instead we observe that hue distributions for field regions commonly occupy a narrow band of values ($10° - 15°$). By using a simple binning technique for clustering, sufficient accuracy was still possible but at a low computational cost. The technique used was a basic 10x range reduction, linearly reducing the maximum range of hue values from 180° to 18° effectively discarding the least significant digit. Combining this technique with the expected range of field hues it becomes evident that field regions will typically occupy only one or two bins which reduces the complexity of establishing connected regions. Figure 3.11 shows the results of applying this binning technique to the hue colour channel, observing how the many hue values of the field region have been assigned to a single bin.



(a) Original (b) Hue before clustering (c) Hue after clustering

(d) Original (e) Hue before clustering (f) Hue after clustering

Figure 3.11: Hue channel clustering (colour mapped for visualisation purposes)

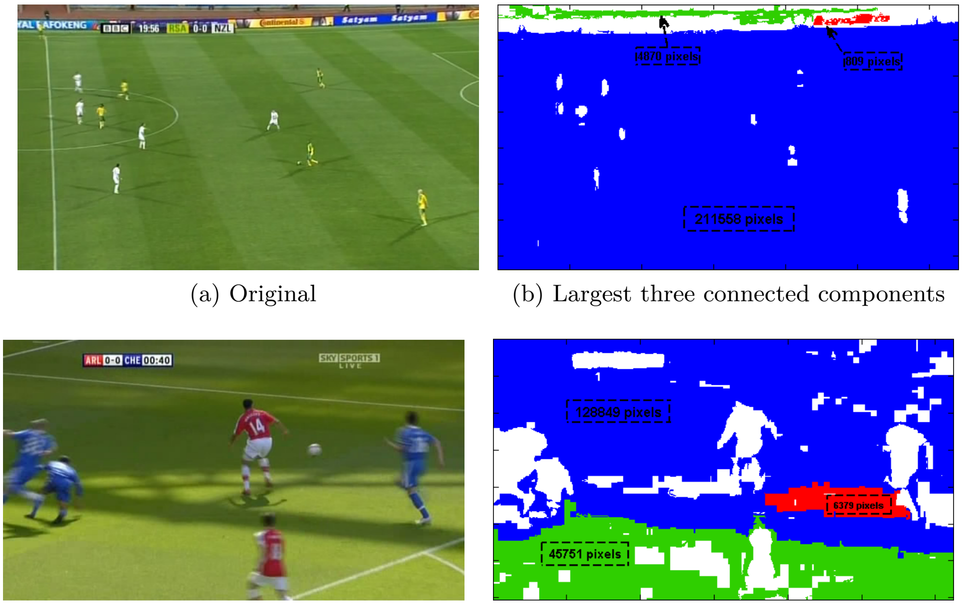
The connected component labelling algorithm used is the same as presented in section 3.1.1.4 with only a few alterations needed to operate on non-binary images. Because of the clustering already performed on the hue channel image, a connected component in this space was defined as a connected set of equal value pixels. As a consequence it could no longer be assumed that neighbouring labels are connected without first verifying the value of the original pixel. The new provisional labelling procedure became:

$$L[p] = \begin{cases} l, (l \leftarrow l+1), & I[i] \neq I[p], \forall i \\ T[\min\limits_{i|I[i]=I[p]}(L[i])], & otherwise \end{cases} \tag{3.15}$$

Figure 3.12 provides two examples of connected component labelling performed on the hue channel of the respective images after clustering has been performed. These components were candidates for inclusion into the field model, their eligibility determined by their size relative to the size of the frame, thus only sufficiently large components were included in the model.

#### 3.1.2.2 Mixture of Gaussians

The mixture of Gaussians technique generates a set of Gaussian distributions used to model the background and, more specifically for our application, to model the field colour. Due



(a) Original

(b) Largest three connected components

(c) Original

(d) Largest three connected components

Figure 3.12: Connected component labelling on hue channel after clustering

to the typically large amount of data present in soccer video sequences and the computationally expensive cost of processing, a batch processing model is not feasible. An on-line or incremental Gaussian mixture model technique was instead used based on expectation maximisation techniques found in [27] (see Algorithm 3.5).

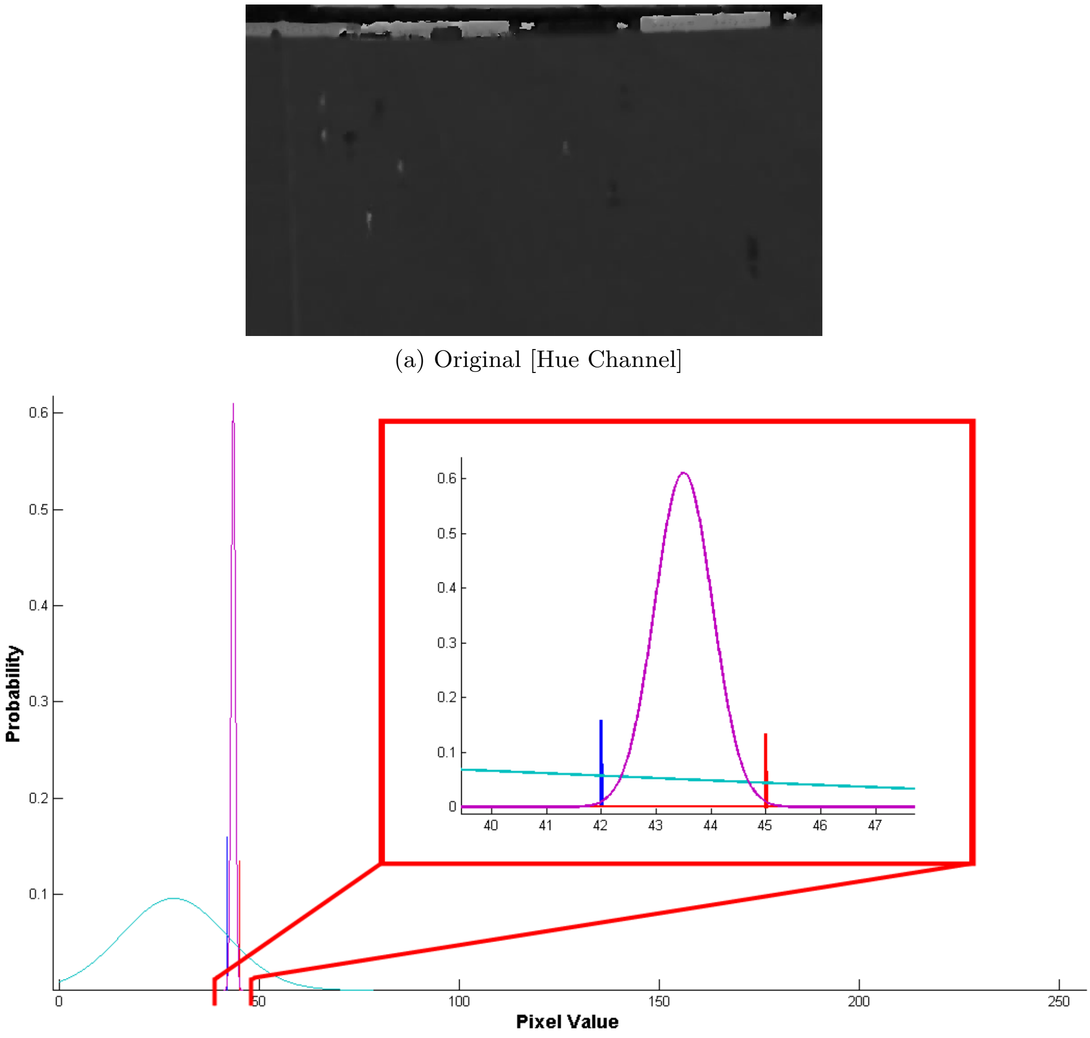
Here the model evolves based on the incoming data. If new data is found to be close to an existing distribution then that distribution is altered and increases in weight relative to other distributions. If new data is not sufficiently close to any existing distributions then the lowest weighted Gaussian is replaced with a new distribution with the mean located at the new data point. This allows dominant distributions to grow while smaller distributions continue to get replaced and remain small.

Focusing on the hue channel, Figure 3.13 shows the resulting set of Gaussian distributions modelled using a single frame. As can be seen in Figure 3.13b a heavily weighted Gaussian forms to encompass a narrow band of hue values which we have assumed to represent the field colour. Here the entire frame is used to model the field without any filtering and due to the specific frame composition the results remain respectable with the highest weight Gaussian achieving a maximum certainty of 62%. However, a different frame composition from a different shot type could result in a significant reduction in accuracy if no filtering was performed prior to model generation.

By restricting the model to include only good field candidates, accuracy was maintained throughout the video sequence. Figure 3.14 shows the Gaussian model of the example frame

**Algorithm 3.5** On-line Gaussian Mixture Learning[27]

```
Control Variables:
$K$        //the number of Gaussian distributions
$V_0$      //the starting standard deviation of a new Gaussian
$\alpha$   //the learning rate constant
$T_\sigma$ //the threshold for the number of standard deviations away from an existing Gaussian

Intialization:
$\forall_{j=1..K}$
    $w_j = 0$,          //set initial weight of all Gaussians to 0
    $\mu_j = \infty$,   //set initial mean of all Gaussians to inf
    $\sigma_j = V_0$,   //set initial std dev of all Gaussians to $V_0$
    $c_j = 0$           //set counter for number of effective observations of all Gaussians to 0

While new data x(t)

    //calculate probability $p$ for Gaussian distribution $g$ given parameters $w$, $\mu$ and $\sigma$
    //at observation $x$ for all $K$ distrobutions
    $\forall_{j=1...K} \quad p_j = \begin{cases} w_j \cdot g_j(x, \mu_j, \sigma_j) & if\ \frac{|x-u_j|}{\sigma_j} < T_\sigma \\ 0 & otherwise \end{cases}$

    If $\sum_{j=1}^{K} p_j > 0$ Then //at least one match is found

        For ($k = 1$, $k < K$, $k++$)

            $q_k = \frac{p_k}{\sum_{j=1}^{K} p_j}$

            If Winner-Take-All Then
                $q_k = \begin{cases} 1 & if\ k = argmax_j\ \{p_j\} \\ 0 & otherwise \end{cases}$
            End If

            $w_k(t) = (1-\alpha) \cdot w_k(t-1) + \alpha \cdot q_k$

            If $q_k > 0$ Then //for matched Gaussians
                $c_k = c_k + q_k$      $\eta_k = q_k \cdot (\frac{1-\alpha}{c_k} + \alpha)$
                $\mu_k = (1-\eta_k) \cdot \mu_k(t-1) + \eta_k \cdot x$
                $\sigma_k^2(t) = (1-\eta_k) \cdot \sigma_k^2(t-1) + \eta_k \cdot (x - \mu_k(t-1))^2$
            End If
        End For
    Else                      //no match found
        $\forall_{j=1...K}$      $w_j(t) = (1-\alpha) \cdot w_j(t-1)$
        $k = argmin_j\ \{w_j\}$      $w_k = \alpha$      $\mu_k = x$      $\sigma_k = V_0$      $c_k = 1$
    End If
    $normalize\ w$
End While
```



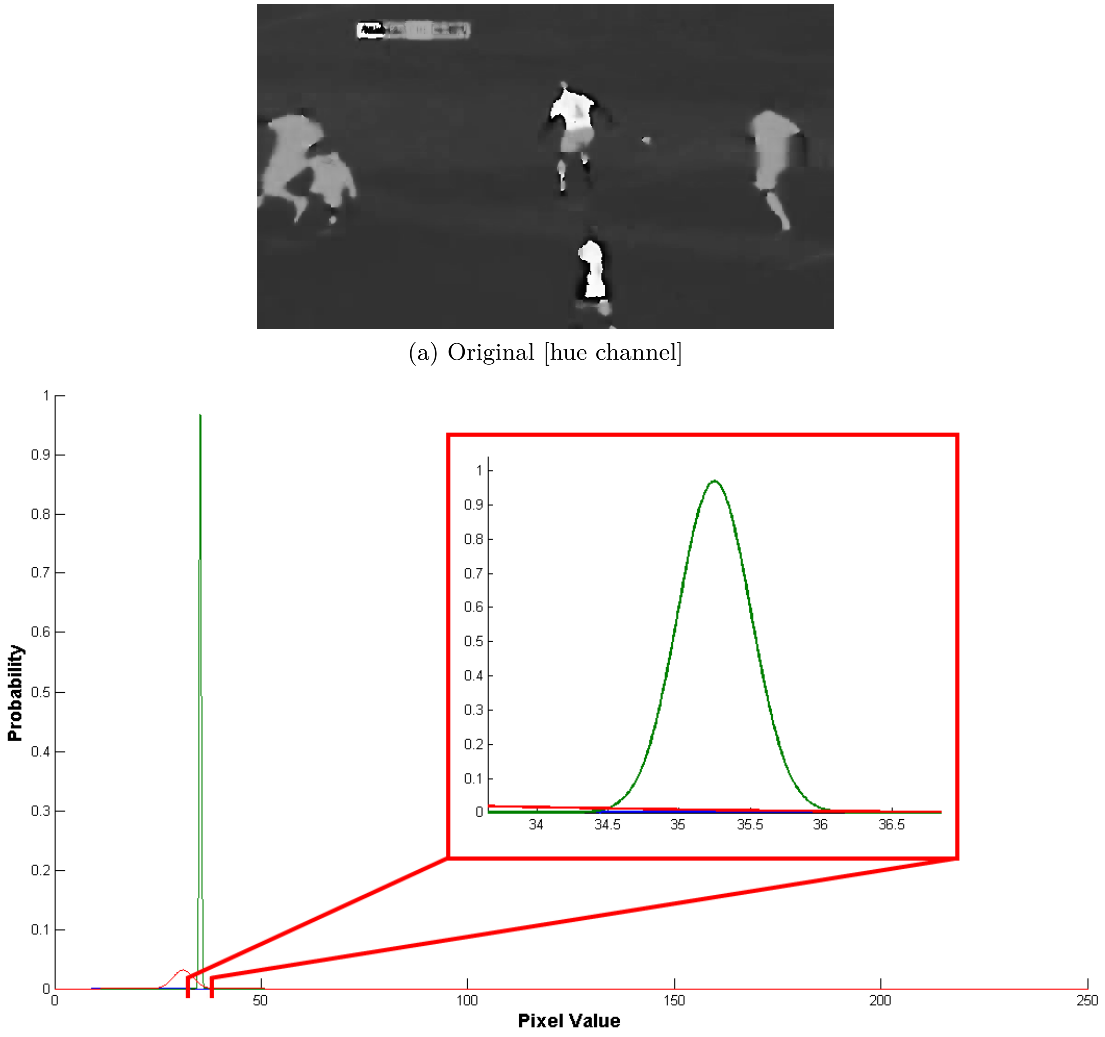
(a) Original [Hue Channel]

(b) Gaussian mixture model

Figure 3.13: Background model of example frame as modelled by a mixture of five Gaussian distributions

after it has been processed to include only the top three candidates. These results show the highest weighted Gaussian achieving a maximum certainty of 98%.

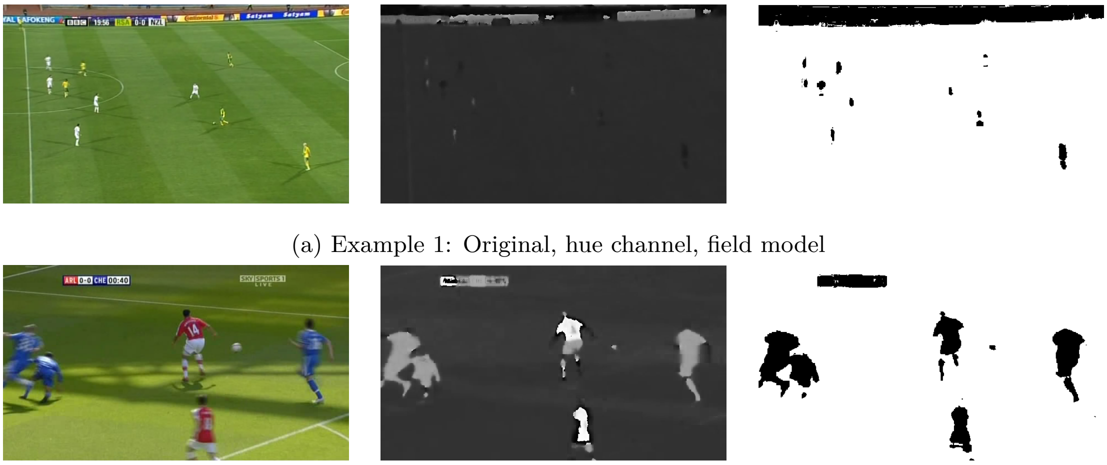
To determine the binary field mask for a given frame we iterated through the frame and compared each pixel with that of the field mask. If the pixel had a sufficiently high probability of belonging to the model then the pixel was marked as field, otherwise it was marked as foreground. For pixel $x$ the probability $p$ of it belonging to the background is $\max(G_k(x))$ where $G_k$ is the $k$th Gaussian distribution and $k \in 1..n$ for $n$ distributions (see Algorithm 3.6). The threshold value varies based on the characteristics of the video segmented. In Figure 3.15 a threshold of 1% ($p \geqq 0.1$) was used.



(a) Original [hue channel]

(b) Gaussian mixture model

Figure 3.14: Background model of example frame as modelled by a mixture of three Gaussian distributions generated using only the three largest candidate field regions



(a) Example 1: Original, hue channel, field model

(b) Example 2: Original, hue channel, field model

Figure 3.15: Field model generated using the mixture of Gaussians technique

**Algorithm 3.6** Field Mask for Gaussian mixture model

```
for each row
    for each col
        /*Get pixel value at position (row, col) in image*/
        x = image(row, col)

        /*Initilize mask as foreground*/
        mask(row, col) = 0

        /*Iterate through each Gaussian distribution*/
        for k = 1 : n

        /*Calculate the probability of the pixel with value x beloning
        to the background model*/
            p = $G_k(x) \longleftarrow w_k \cdot \left\| \frac{1}{\sigma_k\sqrt{2\pi}} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \right\|$

            /*Compare probability to the threshold*/
            if p >= threshold
                /*If probability is greater than threshold
                then set mask at position (row, col) as background*/
                mask(row, col) = 1
                continue
            end
        end
    end
end
```

In both the examples, Figure 3.15a and Figure 3.15b, only a single frame was used to model the background and the selected frames consisted mainly of field coloured pixels. However, for our work the model updates with each frame to maintain accuracy for changing conditions but this does cause concern as not all frames are dominated by field coloured pixels and of these frames a few may contain large homogeneously coloured regions. Depending on their frequency and positioning in the video sequence the impact on the field model can vary. It is possible to use the existing background model to further filter unsuitable candidates, however, this feedback loop relies heavily on the initial model being accurate, therefore the initial frames used to train the model need be carefully selected to ensure high levels of field pixels. This type of manual intervention was unnecessary for this work. The model's ability to recover from false positives generated by these types of regions will determine the necessity of this approach for situations not present in this work.

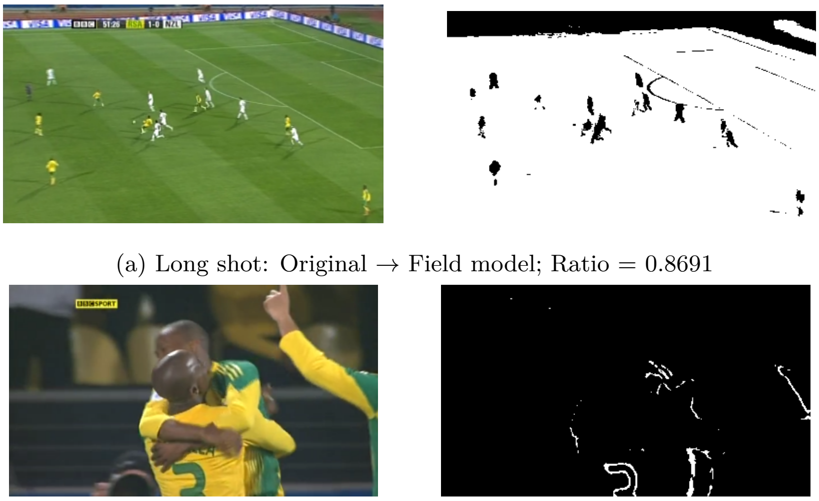
## 3.2 Shot Classification

### 3.2.1 Shot Features

Building upon a solid foundation created with the field model, various methods of extracting features from the model were investigated in order to assist with shot classification. Two broad types of features recur frequently in the literature, those directly establishing shot types from field properties and those which attempt to infer higher level features from this data [51]. For this work the focus was on investigating the link between field ratios and shot classes and then also at a higher level, established methods to estimate the field of view, or zoom level of a frame were also investigated.

#### 3.2.1.1 Field Ratios and Frame Composition

The ratio of field to non-field pixels in a shot is a computationally friendly method for determining certain shot characteristics [29]. Two shot classes which can be reliably separated by examining global field colour ratios are the out of field and long shots as they lie on opposing ends of expected values. Long shots contain high ratios of field coloured pixels while out of field shots typically demonstrate very low ratios [44, 48, 49] (see Figure 3.16).

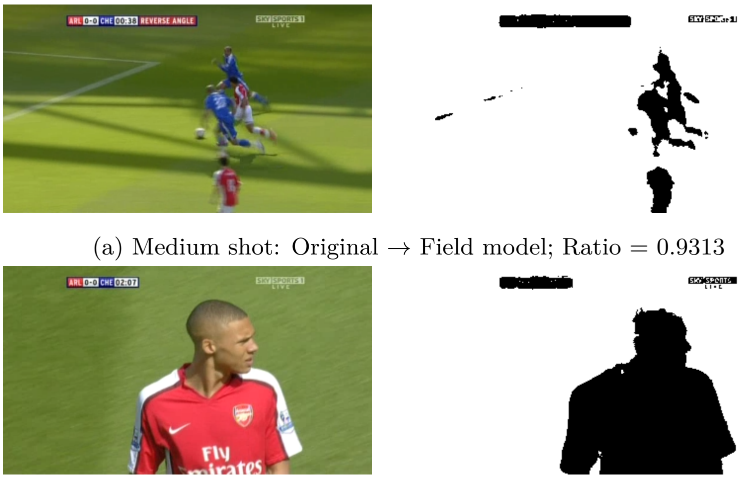


(a) Long shot: Original → Field model; Ratio = 0.8691

(b) Out of Field: Original → Field model; Ratio = 0.0119

Figure 3.16: Global field ratios established from field models for long and out of field shots

Medium and close-up shots provide a more difficult classification proposition as their field ratios can vary significantly depending on the composition of the specific frame and often contain a significant quantity of field coloured pixels, which make them difficult to separate from long shots (see Figure 3.17). To incorporate this frame composition variance, the image was segmented into zones and the field ratios of the varying zones were determined to link specific field ratios to a locality. This is useful because of the general consistent nature of production techniques used in broadcast soccer videos hence the composition of a frame can often yield pertinent information.

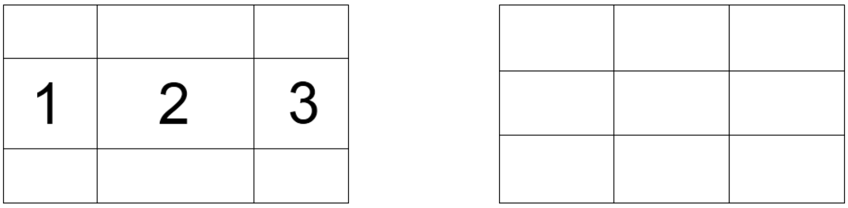


(a) Medium shot: Original → Field model; Ratio = 0.9313

(b) Close up: Original → Field model; Ratio = 0.7408

Figure 3.17: Global field ratios established from field models for medium and close up shots

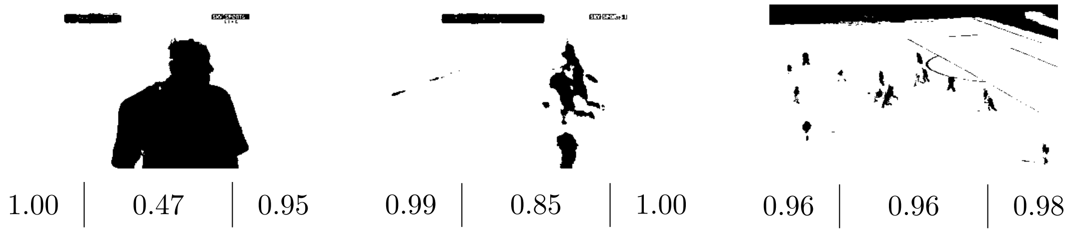
Two different spatial segmentation patterns were used on the field model to establish local field ratios for the various segments, shown in Figure 3.18. The golden section segmentation pattern uses a ratio of 3:5:3 for both the horizontal and vertical directions following common production strategies described in [34]. The evenly spaced pattern of nine segments was used as reference to create a feature pattern based on field ratio locations. Figure 3.19 shows the potential advantage gained in discriminatory power from calculating ratios based on predefined sections, as opposed to calculating a single ratio for the entire frame.



(a) Golden Section (3:5:3), only sections 1, 2 and 3 used

(b) Nine evenly distributed sections, all sections used

Figure 3.18: Spacial segmentation patterns

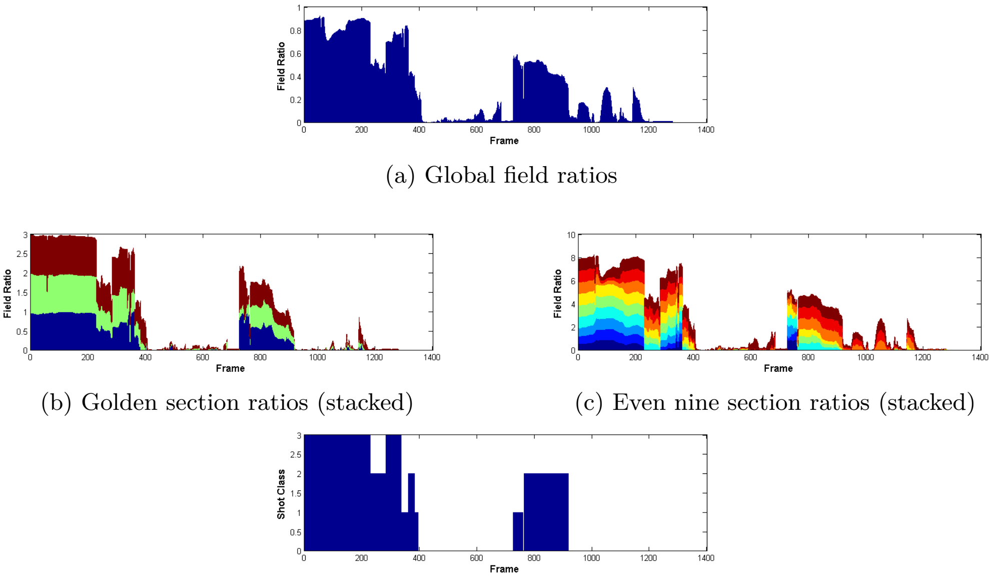


(a) Close (b) Medium (c) Long shot

Figure 3.19: Applying golden section segmentation pattern to close, medium and long shots

Field ratios were also segmented temporally for the purpose of establishing shot boundaries to inform classification procedures and to potentially reduce the number of classification decisions by only classifying certain frames per shot. Two basic techniques were used for this task. The first technique simply compares field ratios of the current frame to those of the previous frame and calculates the absolute difference. This is useful for the detection of shot boundaries which result from ‘cuts’ from one shot to another. The second technique compares field ratios of the current frame to the previous $K$ frames, where $K$ is the window size, to determine the average absolute difference between frames. This is useful for detecting shot boundaries generated by gradual transitions such as ‘fades’.

Figure 3.20 shows the ratios for each respective segmentation pattern (Figures 3.20a, 3.20b, 3.20c) over the course of a video sequence together with the manually defined shot class ‘mask’ (Figure 3.20d) for the sequence. In Figure 3.20a the golden field ratio has been plotted as it changes over the course of the 1278 frame sequence. Comparing this plot to that in Figure 3.20d we can see a certain level of correlation as the ratios change in response to the shot changes. This correlation can once again be seen with the golden and even ratios shown in Figures 3.20b and 3.20c. However, because ratios for individual areas are determined it is possible to see the changing relationship between these areas in a frame thus providing more information with which to determine the shot type. Additionally, large changes in successive ratios provide a strong indicator for shot boundaries and demonstrates the value of field ratios as a feature for shot classification.



(a) Global field ratios

(b) Golden section ratios (stacked)

(c) Even nine section ratios (stacked)

(d) Shot class mask for sequence; 3-Long, 2-Medium, 1-Close, 0-Out of field

Figure 3.20: Field ratios over the course of a 1278 frame video sequence

#### 3.2.1.2 Field of View

The field of view or zoom level of a shot can reasonably be considered as the predominant characteristic separating the three different types of in-field shots; long, medium and close. Therefore, being able to estimate a value for this characteristic is of great benefit to shot classification. Various methods have been used to generate this estimate and while the techniques are different, the core philosophy for all of them is similar, that of establishing a relationship between the field and the foreground.

The first technique used is described in [17]. It consists of generating a 1-dimensional field-player interaction curve using a vertical projection of field pixels (see Figure 3.21). By examining the characteristics of this curve, an estimate for the field of view of a frame is formed. The mean and standard deviation (Equations 3.16 and 3.17) are the features of the vertical projection used to classify the different shot types.

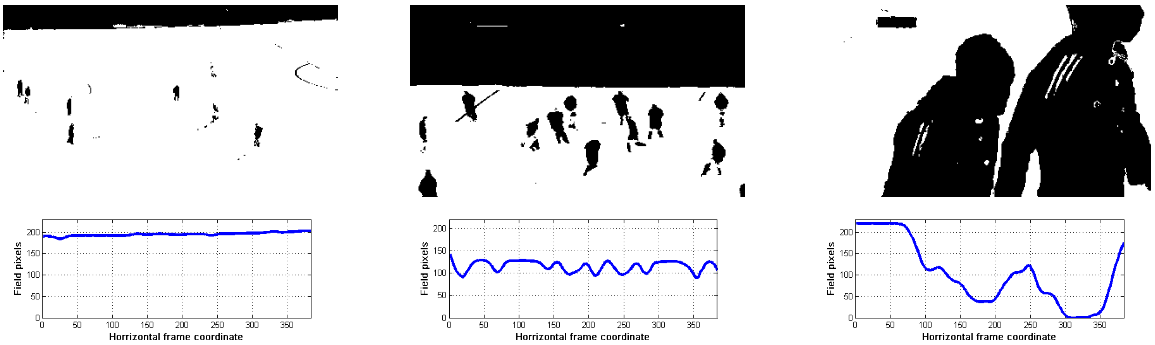
The mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.16}$$

where $x$ is the input vector and $n$ the total number of elements. The standard deviation $s$ is then given by

$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.17}$$

A cursory analysis of the data, together with observable trends in the literature, lead to the following hypothesis which would guide the experimentation. A high mean with small standard deviation suggests a long shot such as Figure 3.21a. A medium to high mean with a middling standard deviation suggests a medium shot such as Figure 3.21b. Lastly a low to medium mean with a large standard deviation suggests a close-up shot such as 3.21c. Of course these definitions of high, low, medium, small and large are subjective and vary from game to game but they do provide a useful illustration to convey the salient properties of this type of projection.



(a) Long shot (b) Medium Shot (c) Close-up Shot

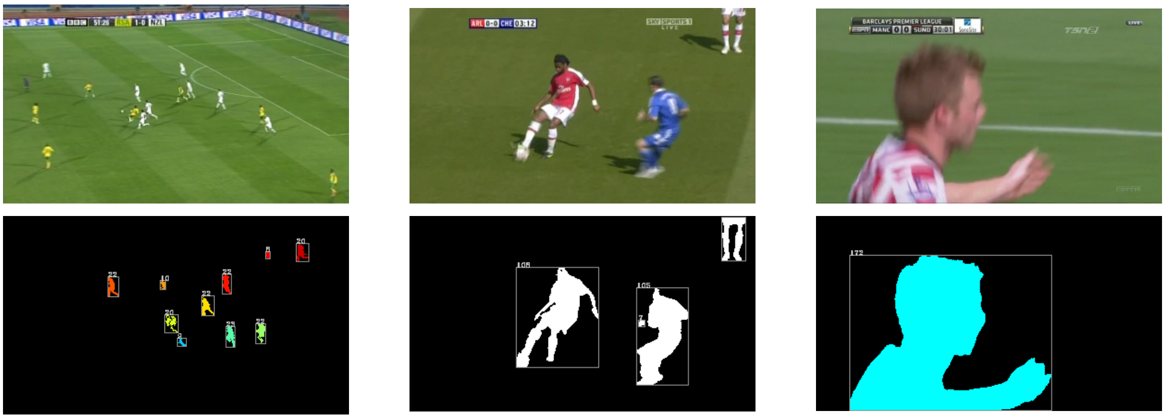
Figure 3.21: Vertical projection of field pixels (smoothed)

A second technique used, more common in the literature, is that of establishing a relationship between field and foreground by virtue of their respective sizes. This approximation was done using connected component analysis of the frame foreground and applying similar filtering rules as those used in section 3.1.1.4 where candidates which do not fit within expected dimensions are removed from consideration ($0.2 \leq \frac{width}{height} \leq 2$). By calculating these ratios of sizes we were able to generate an estimate for the field of view by assuming the smaller the foreground objects were in relation to the field the further out the camera was zoomed. Filtering rules were kept simple and only aspect ratio was considered under the assumption that most of the objects should correspond to players and players are almost always in an upright position. The size of a component is determined by the number of contributing pixels and ratios are calculated by a size weighted sum of height compared to the height of the frame (see equations 3.18 and 3.19). Figure 3.22 shows the results of connected component analysis of the foreground object detected in the three 'field backed' shot types. From the connected component analysis the heights of the varying objects are estimated by the bounding box heights of each component with larger object heights corresponding to a greater zoom level associated with the shot types. The weighted object height, H, and the height ratio, HeightRatio, are given by:

$$H = \frac{\sum_{i=1}^{n} w_i \cdot h_i}{\sum_{i=1}^{n} w_i} \tag{3.18}$$

$$HeightRatio = \frac{H}{height_{image}} \tag{3.19}$$

Where $w_i$ and $h_i$ correspond to the weight and height of object $i$ respectively.



(a) Long shot

(b) Medium shot

(c) Close shot

Figure 3.22: Size ratios of foreground objects

### 3.2.2 Shot Classifiers

Finally, the features established previously in Section 3.2.1 (field ratio, vertical projection and object size ratio) could be combined in various ways to perform classification for a frame. Shot classifications were initially calculated on each of the features separately to better understand the discriminatory power of each feature. Later they were combined to leverage the strengths of each. The decision tree was used as the classification tool for this work as it offered several suitable properties, such as the ease with which features can be combined and separated and the ability to observe the performance of individual decision nodes. The decision tree also offers flexibility by being able to combine other classifiers into the decision structure without much alteration.

#### 3.2.2.1 Field Ratios

The features investigated are the field ratios in the three different configurations; global, gold section and even nine. The goal here was to establish which ratios and in what order would produce the most reliable classification. Each set of ratios were evaluated separately. The global ratio is only a single value per frame so it is not possible to establish intra-relationships, such as those possible with the other two configurations. For the golden section both the weighted average ratio, $golden_{mean}$, of each section and the absolute difference, $golden_{diff}$, between the middle section ratio (R2) and the maximum of the two outer section ratios (R1 and R3) were used, given as;

$$golden_{mean} = \frac{3 \times R1 + 5 \times R2 + 3 \times R3}{3 + 5 + 3} \tag{3.20}$$

$$golden_{diff} = \max\left[(||R2 - R1||), (||R2 - R3||)\right] \tag{3.21}$$

This provided clues towards establishing the existence of a large foreground figure occupying the central area of the frame typically associated with close-up shots. Two features for the even configuration were defined, the average of the highest six section ratios, $even_{mean}$ (useful for eliminating noise introduced from certain boundary elements in a frame such as the stands or advertising boards) and the absolute difference between the outer and inner areas, $even_{diff}$. The outer area being those sections on the far left, $L$, and far right, $R$, of the frame with the inner area being those sections down the centre, $C$, of the frame. $even_{mean}$ and $even_{diff}$ are given as;

$$even_{mean} = \frac{\sum_{i=1}^{6} SR_i}{6} \tag{3.22}$$

$$L = \frac{R1 + R4 + R7}{3}$$

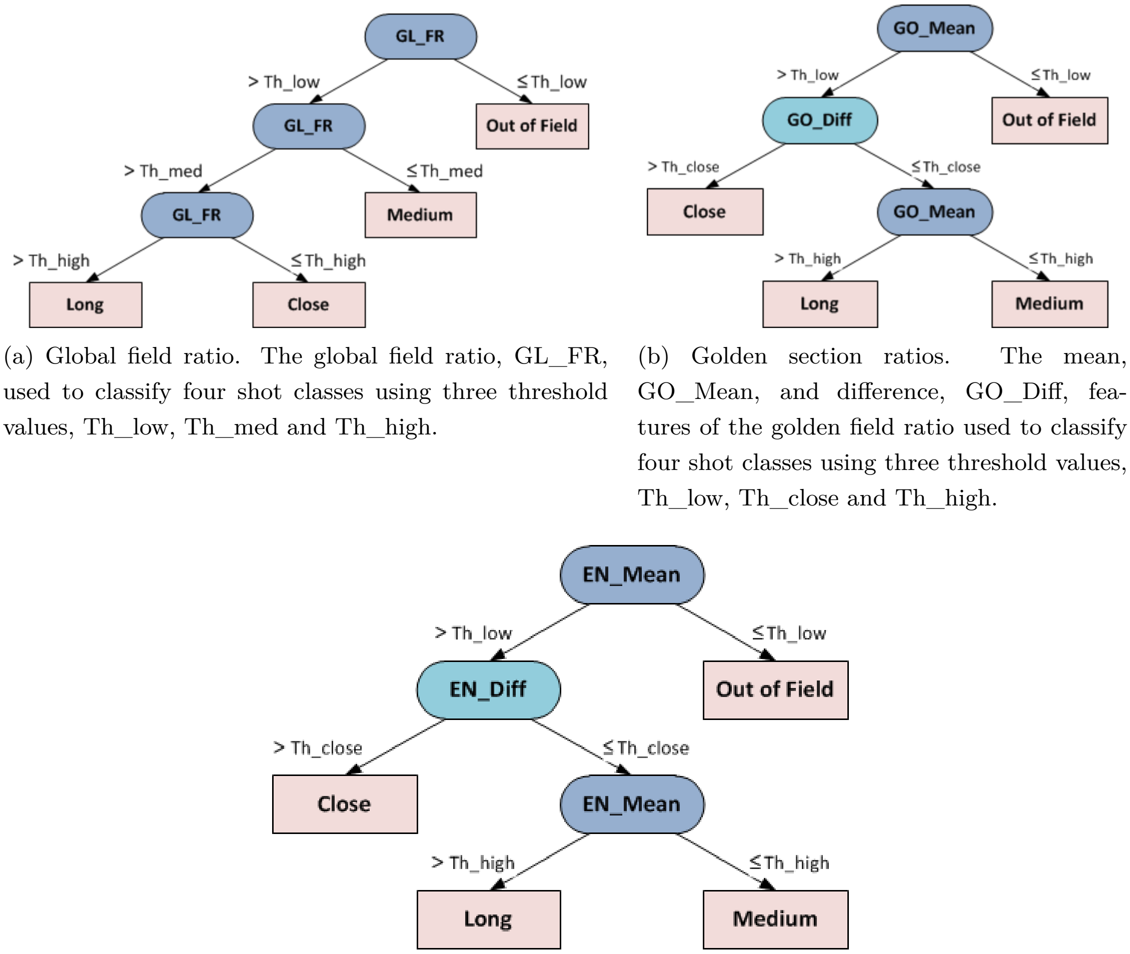
$$R = \frac{R3 + R6 + R9}{3}$$

$$C = \frac{R2 + R5 + R9}{3}$$

$$even_{diff} = \frac{||C - R|| + ||C - L||}{2} \tag{3.23}$$

where $SR$ is a vector of even field ratio sections sorted in descending order such that $SR_1$ contains the largest ratio and $SR_9$ the smallest

The even segmentation is used to describe the overall composition of the frame, in a similar way to the global ratio but with the ability to compensate for bias introduced from out of field elements typically found either on the top or bottom of a long shot. The golden sections account for the more nuanced differences between the shots such as large centrally located foreground objects.



(a) Global field ratio. The global field ratio, GL_FR, used to classify four shot classes using three threshold values, Th_low, Th_med and Th_high.

(b) Golden section ratios. The mean, GO_Mean, and difference, GO_Diff, features of the golden field ratio used to classify four shot classes using three threshold values, Th_low, Th_close and Th_high.

(c) Even nine ratios. The mean, EN_Mean, and difference, EN_Diff, features of the even nine field ratio used to classify four shot classes using three threshold values, Th_low, Th_close and Th_high.

Figure 3.23: Decision trees for field ratios

#### 3.2.2.2 Field of View

We have previously described, in Section 3.2.1.2, the two techniques used to estimate the field of view of a shot. Here we describe how these derived properties were combined in a decision tree structure to produce a shot classification.

The vertical projection technique produces two descriptors which were used for classification; mean projection height and standard deviation from mean. The mean projection height is analogous to the global field ratio but the smoothing of the projection removes certain outliers and reduces certain types of noisy data. The standard deviation provides information about the uniformity of the frame with the assumption that most disturbances will be caused by players on the field, identified as foreground objects, therefore the larger the disturbances, the larger the foreground objects hence the narrower the field of view. Incorporating this information, a decision tree (Figure 3.24) was constructed, where low mean values indicate out of field shots, large deviations indicate a close shot, a high mean with a small deviation indicates a long shot and a low mean with small deviations indicate a medium shot.

The next technique provides a single descriptor, the object/field ratio, which is best used in combination with other descriptors but for evaluation purposes was initially used separately. Similar to the standard deviation of the vertical projection, the object ratio provides information about the disturbances in the shot with the larger the disturbance the narrower the field of view. Therefore a large ratio was associated with a close-up shot, a small ratio with a long shot and anything in between determined as a medium shot (see Figure 3.25). Due to the nature of the feature, out of field shots were ignored/uncategorised.

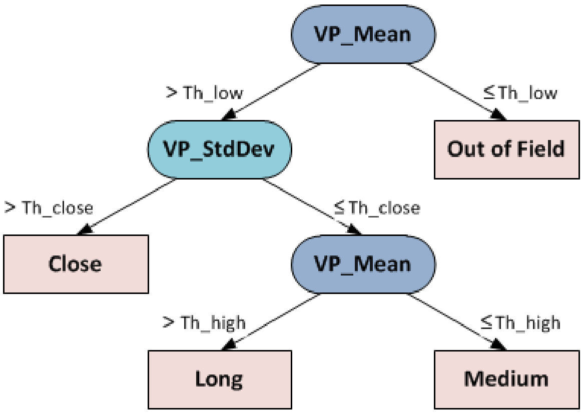


Figure 3.24: Vertical projection decision tree. The mean, VP_Mean, and standard deviation, VP_StdDev, features of the vertical projection are used to classify four shot classes using three threshold values, Th_low, Th_close and Th_high.

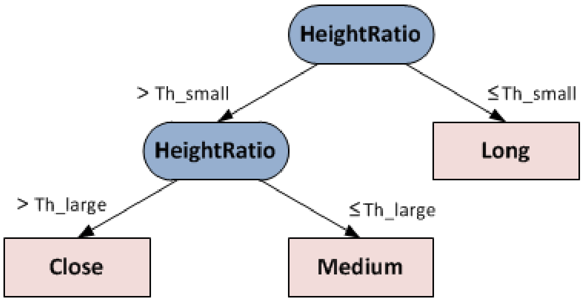


Figure 3.25: Object/Field size ratio decision tree. The height ratio is used to classify three shot classes using two threshold values, Th_small and Th_large.

#### 3.2.2.3 Feature Combination and the Neural Network Classifier

As a final stage, a selected combination of the previously described features were used in the classification procedure. Initially a simple decision tree structure was again employed to classify the shot types given the input feature set, however, because determining the performance of the individual features was no longer necessary a higher level classification technique, the multi-layer perceptron neural network (MLP-NN), was used instead. Using this technique generated results closer to those seen in other works where a combination of features is typically used, giving a better comparative reference.

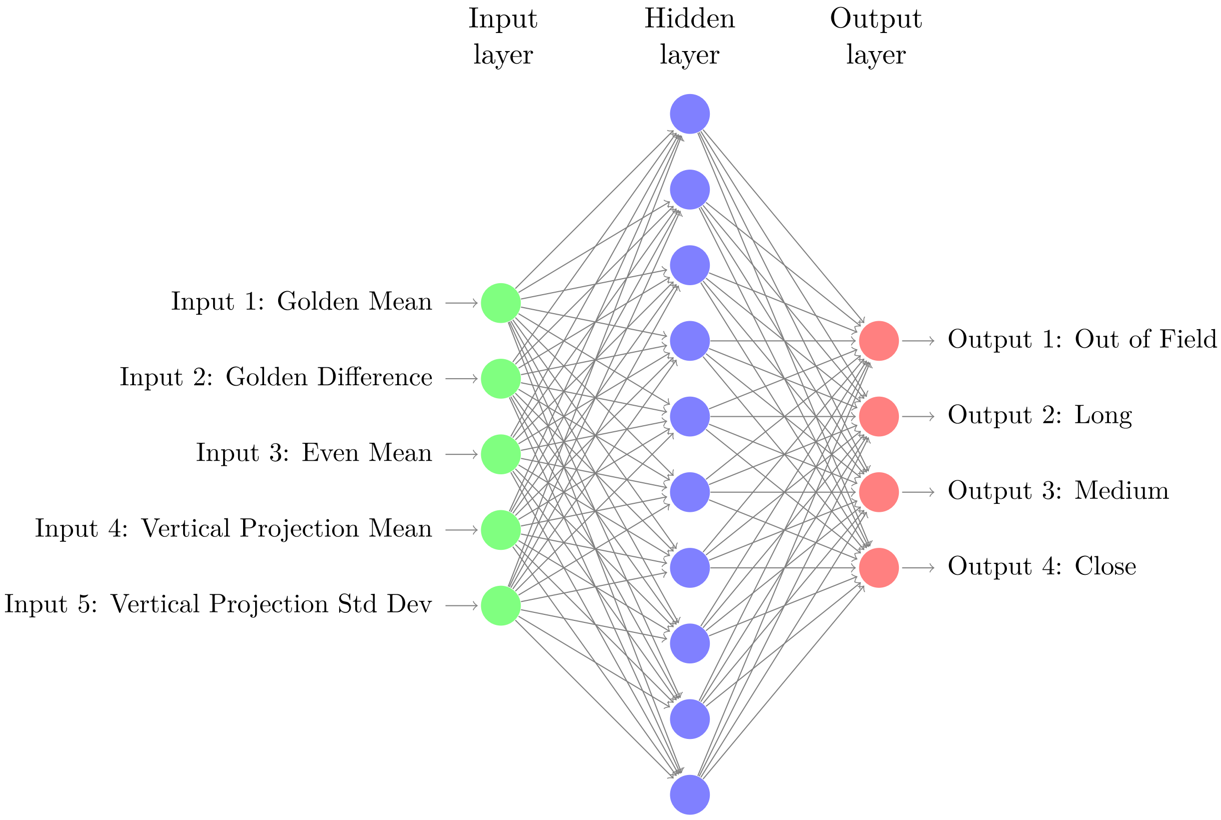
The features were selected based on a demonstrated potential for shot classification. This feature set was comprised of five previously described features; the mean and max difference features of the golden field ratio, the mean top six feature of the even nine field ratio and the mean and standard deviation of the vertical projection. This formed the 5-dimensional feature vector used as the input for the MLP-NN.

The MLP-NN was constructed as a two-layer, feed-forward network using a tan-sigmoid activation function for both hidden and output neurons. The scaled conjugate gradient back-propagation algorithm[2] was used to train the network. Figure 3.26 illustrates the basic neural network topology with input, hidden and output layers.

The training data was constructed as a subset of frames from each video sequence. The training subset consisted of 30 randomly selected feature vectors from each shot class, thus giving 120 [3] vectors used to train the network. The sample set was then segmented into training (80%), verification (15%) and testing (5%) data sets. The sample data was

[2] See [35] as implemented by MATLAB's[4] Neural Network toolbox

[3] Sequence 2 has less than 30 close shots therefore the total training sample was fewer than 120



*Note: Not all hidden layer neurons are shown

Figure 3.26: Multi-layer perceptron neural network

**Algorithm 3.7** Training Data Selection

**Input:** The feature matrix, **data**, of size $N \times K$ where $N$ is the number of feature vectors and $K$ the number of components per vector. The target class for each feature vector, **dataClass**, of size $1 \times N$ with values ranging from 1 to the total number of classes, $C$.

**Initialisation:** $\forall_{j=1..C}\ classCount_j = 0 \qquad \forall_{j=1..N}\ x_j = Random(N) \qquad count = 0$

```
for each $x_i$

    for j = 1 to C
        if( $classCount_j \geq 30$ )
            break

    if( $classCount_{dataClass_i} < 30$ )
        outputData($count$) = $data_i$ [$data_i$ = row $i$ of data]
        outputClass($count$, $dataClass_i$) = 1
        $classCount_{dataClass_i} = classCount_{dataClass_i} + 1$
        $count = count + 1$
    end
end
```

**Output:** The sample feature matrix, **outputData**, of size $count \times K$ with the number of sample feature vectors equal to $count$. The associated target class for each sample, **outputClass**, of size $count \times C$ in a format such that for each sample feature vector $i$ the target class is given by the column of **outputClass** at row $i$ containing the value 1.

randomly selected using Algorithm 3.7. The samples were randomly selected to increase the classifiers ability to generalise classification to a larger data set and to avoid over training specific samples.

### 3.2.3 Summary

This chapter has detailed the various methods used to classify shots in broadcast soccer video. The fundamental process of background modelling where two techniques, Colour Distance modelling and Gaussian Mixture modelling, were described first. Then the feature set comprising of three distinct features, field ratio, vertical projection and size ratio were described next. Finally the classification process driven by the decision tree model and a multi-layer perceptron neural network was described.

# Chapter 4

# Experimental Results and Discussion

This chapter details and discusses the results of the various techniques used to perform shot classification of broadcast soccer videos. The results for the two field colour (background) modelling techniques, colour distance modelling and Gaussian mixture modelling, are presented first followed by the results of shot classification using the three features described previously viz. field ratios, vertical projection and object size ratio and the combined classification using the MLP-NN. Each technique was applied to four different broadcast soccer video sequences of varying length from different matches and broadcast providers (see Table 4.1 for details). Experiments were performed on a Dell Inspiron 1525 laptop with a Intel Core 2 Duo T5750 CPU with 2GB of RAM, running Windows Vista 32-bit. The software tools comprised of Microsoft's Visual Studio 2009 IDE using the OpenCV computer vision library with MATLAB used for analysing results data.

| Sequence No. | Total Frames | Original Resolution | Processed Resolution | Shots* |
|---|---|---|---|---|
| 1 | 1550 | 640 x 368 | 384 x 220 | 11 |
| 2 | 948 | 640 x 368 | 384 x 220 | 6 |
| 3 | 3274 | 624 x 352 | 374 x 211 | 31 |
| 4 | 1846 | 640 x 360 | 384 x 216 | 12 |

*Number of continuous shot sequences

Table 4.1: Video Sequence Details

## 4.1 Field Colour Modelling

In this section we present the results for the colour distance modelling and Gaussian mixture modelling field modelling techniques. The results for each are given in the form of precision, recall and accuracy, first overall per sequence then by class per sequence. Precision, recall and accuracy are given as [19];

$$Precision = \frac{t_p}{t_p + f_p} \tag{4.1}$$

$$Recall = \frac{t_p}{t_p + f_n} \tag{4.2}$$

$$Accuracy = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \tag{4.3}$$

> where $t_p$ and $t_n$ are the true positives and true negatives respectively and $f_p$ and $f_n$ are the false positives and false negatives respectively.

The ground truth field model was manually generated for 10 to 15 frames per sequence (total of 46), with the frames extracted from throughout the sequence to sample the various shot types. These ground truth models were then compared with those generated by the background modelling techniques using Algorithm 4.1, generating the results in the form of true positive/negative, false positive/negative. A true positive is any background pixel which is labelled as background by the model. A true negative is any non-background pixel which is labelled as a non-background pixel by the model. A false positive is a non-background pixel which is labelled as a background pixel by the model. A false negative is a background pixel which is labelled as a non-background pixel by the model.

A full set of results in the form of confusion matrices is supplied in Appendix A.

**Algorithm 4.1** Ground Truth versus Model Compare

```
for row = 1 : size(groundtruth,1)
    for col = 1 : size(groundtruth,2)
        current_gt = groundtruth(row, col);
        current_model = model(row, col);

        if(current_gt == current_model)
            if(current_gt)
                truepos = truepos + 1;
            else
                trueneg = trueneg + 1;
            end
        end

        if(current_gt < current_model)
            falsepositive = falsepositive + 1;
        end

        if(current_gt > current_model)
            falsenegative = falsenegative + 1;
        end
    end
end
```

### 4.1.1 Colour Distance Modelling

Table 4.2 presents the results of the colour distance modelling process both with and without filtering performed by connected component analysis (described in Section 3.1.1.4) where field candidates were filtered based on size and aspect ratio. The mean and standard deviations for precision, recall and accuracy are shown per sequence. Table 4.3 on page 56 presents the resulting precision, recall and accuracy means separated by shot class (long, medium, close and out of field) per sequence.

The accuracy of the unfiltered performance shows the maximal difference across all sequences is less than 2% with an average standard deviation of less than 3%. This indicates the technique to be stable throughout a sequence as well as being robust to changes of venue, field conditions and broadcasting styles between the different sequences. Similarly the performance is consistent across all classes with a range of only 6% between the best and worst sequence/class performances, and on average a less than a 2% overall difference between classes. This shows the robustness of the technique in handling changes of camera angle and switching between the various cameras as seen between shots. Recall performance show similar values with a 2% range and a 4% standard deviation.

Precision values experienced the greatest fluctuation between sequences with a range of more than 7% with the worst performance seen in sequence 3 with 89.13% precision at a standard deviation of 15.23%, nearly 3 times higher than the other sequences. This poor performance is due mostly to a few frames containing a small number of field pixels, that tend to amplify the negative performance aspects of any false positive classifications. The per class comparison shows that the medium shot class has a precision of 77.14%, significantly lower than the 90.83% of the next worst performing sequence/class pair. The poorest performing frame in the sequence achieved a precision of 52.66% while still having a recall of 99.79% and an accuracy of 87.79%. This suggest that for a binary classification system, such as field modelling, where false positives and false negatives have an equal impact on the performance, accuracy is a more indicative measure of performance than precision and recall especially in situations where a large bias towards true positives or true negative may exist.

Overall, filtering increases the precision by +1.13% but at the expense of decreased recall and accuracy of -2.01% and -0.65% respectively. This was due to the filtering process increasing the mean number of false negatives while decreasing the mean number of false positives per frame. At this level of performance the increased computational cost of performing connected component analysis is not worthwhile. However, if a more advanced, accurate filtering process were to be employed it should be possible to universally increase the performance of the modelling technique.

Thus the colour distance model has been shown to be consistent and accurate across shot classes and video sequences, properties which are highly desirable for a field modelling technique. Trying to improve the results through post-processing via filtering did not have the desired effect and reduced the overall accuracy but this was more likely due to the implementation of the filtering process rather than the process itself. A more thorough process may indeed improve the results as required.

| | **Unfiltered** | | | **Filtered** | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **Accuracy** | **Precision** | **Recall** | **Accuracy** |
| **Sequence 1** | | | | | | |
| Mean | 96.36% | 97.57% | 96.25% | 97.15% | 93.51% | 94.45% |
| Std Dev | 3.89% | 2.56% | 2.72% | 4.13% | 14.52% | 7.16% |
| **Sequence 2** | | | | | | |
| Mean | 95.98% | 97.06% | 95.13% | 97.05% | 96.52% | 95.64% |
| Std Dev | 3.13% | 4.37% | 2.58% | 2.26% | 5.09% | 2.62% |
| **Sequence 3** | | | | | | |
| Mean | 89.13% | 99.59% | 95.27% | 91.70% | 97.85% | 95.11% |
| Std Dev | 15.23% | 0.44% | 3.51% | 9.57% | 0.45% | 1.66% |
| **Sequence 4** | | | | | | |
| Mean | 93.75% | 99.51% | 94.91% | 93.87% | 97.81% | 93.76% |
| Std Dev | 3.21% | 0.96% | 2.70% | 2.80% | 1.10% | 2.89% |
| **Average** | | | | | | |
| Mean | 93.81% | 98.43% | 95.39% | 94.94% | 96.42% | 94.74% |
| Std Dev | 6.36% | 2.08% | 2.88% | 4.69% | 5.29% | 3.58% |

Table 4.2: Colour Distance performance by sequence

| | Unfiltered | | | Filtered | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **Accuracy** | **Precision** | **Recall** | **Accuracy** |
| **Sequence 1** | | | | | | |
| Long | 97.96% | 99.17% | 97.56% | 97.80% | 98.93% | 97.33% |
| Medium | 98.06% | 96.90% | 97.13% | 98.75% | 86.22% | 91.96% |
| Close | 96.47% | 95.31% | 96.22% | 99.37% | 94.97% | 97.13% |
| **Sequence 2** | | | | | | |
| Long | 95.43% | 99.10% | 95.66% | 96.60% | 98.89% | 96.48% |
| Medium | 97.18% | 90.04% | 93.39% | 97.55% | 89.36% | 93.22% |
| Close | 97.55% | 93.87% | 94.24% | 98.80% | 91.86% | 93.85% |
| **Sequence 3** | | | | | | |
| Long | 95.20% | 99.49% | 96.55% | 95.57% | 97.91% | 95.74% |
| Medium | 77.14% | 99.65% | 92.15% | 83.93% | 97.94% | 93.62% |
| Close | 97.95% | 99.71% | 98.31% | 97.58% | 97.50% | 96.50% |
| **Sequence 4** | | | | | | |
| Long | 94.80% | 99.33% | 94.64% | 94.66% | 98.04% | 93.48% |
| Medium | 90.95% | 99.58% | 95.49% | 92.23% | 97.40% | 94.94% |
| Close | 93.40% | 99.99% | 95.15% | 93.17% | 97.53% | 93.40% |
| **Average** | | | | | | |
| Long | 95.85% | 99.27% | 96.10% | 96.16% | 98.44% | 95.76% |
| Medium | 90.83% | 96.54% | 94.54% | 93.12% | 92.73% | 93.44% |
| Close | 96.34% | 97.22% | 95.98% | 97.23% | 95.47% | 95.22% |

Table 4.3: Colour Distance performance by shot class

### 4.1.2 Gaussian Mixture Model

The Gaussian Mixture model (GMM) is the second field modelling technique implemented. Table 4.4 shows the results in terms of precision, recall and accuracy of the Gaussian mixture model technique applied to the four video sequences, firstly showing the overall means and standard deviations and then showing the mean values by class.

The accuracy of the Gaussian mixture model proved somewhat less consistent than the colour distance model with a range across video sequences of 5.54% and an overall standard deviation of 7.13%. The per class accuracy showed a range of over 10% between best and worst with an overall difference of 1.13% between classes. Much of this variation can be attributed to the need to alter the model parameters for each sequence, namely the learning rate of the model and the initial standard deviation for newly created Gaussians. In certain circumstances, where frames which were not field colour dominated existed near the beginning of the sequence or lasted for an extended duration the GMM would occasionally learn non-field colours as primary e.g. player uniform colours. This had a severe negative impact on the performance. With a better understanding of the relationship between model parameters and sequence properties it should be possible to increase both the stability and the performance of the model by setting appropriate values for the Gaussian initialization parameters and especially the learning rate. The initialization parameters set the properties of newly formed Gaussian in the mixture, which occur when replacing the previous lowest valued Gaussian in the Mixture, and the learning rate determines how quickly the model will adjust to new data.

| | Precision | Recall | Accuracy |
|---|---|---|---|
| **Sequence 1** | | | |
| Mean | 89.25% | 99.13% | 92.09% |
| Std Dev | 9.40% | 1.68% | 4.08% |
| **Sequence 2** | | | |
| Mean | 95.03% | 97.71% | 94.57% |
| Std Dev | 3.45% | 2.05% | 2.77% |
| **Sequence 3** | | | |
| Mean | 89.72% | 99.67% | 88.63% |
| Std Dev | 8.18% | 0.60% | 16.86% |
| **Sequence 4** | | | |
| Mean | 92.98% | 99.84% | 94.18% |
| Std Dev | 5.34% | 0.21% | 4.81% |
| **Average** | | | |
| Mean | 91.74% | 99.09% | 92.37% |
| Std Dev | 6.59% | 1.13% | 7.13% |

(a) by sequence

| | Precision | Recall | Accuracy |
|---|---|---|---|
| **Sequence 1** | | | |
| Long | 91.72% | 99.19% | 92.69% |
| Medium | 84.28% | 99.92% | 90.64% |
| Close | 96.77% | 95.67% | 94.87% |
| **Sequence 2** | | | |
| Long | 94.32% | 97.77% | 93.68% |
| Medium | 95.17% | 96.19% | 95.50% |
| Close | 98.42% | 98.96% | 98.11% |
| **Sequence 3** | | | |
| Long | 87.71% | 99.95% | 91.54% |
| Medium | 88.82% | 99.17% | 94.75% |
| Close | 96.08% | 99.73% | 96.93% |
| **Sequence 4** | | | |
| Long | 95.04% | 99.89% | 95.44% |
| Medium | 94.26% | 99.54% | 96.62% |
| Close | 85.53% | 99.99% | 87.95% |
| **Average** | | | |
| Long | 92.20% | 99.20% | 93.34% |
| Medium | 90.63% | 98.70% | 94.38% |
| Close | 94.20% | 98.59% | 94.47% |

(b) by class

Table 4.4: GMM performance

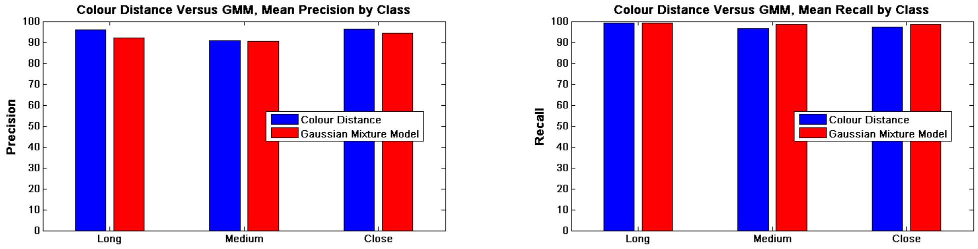
### 4.1.3 Comparison with related works

Figures 4.1a and 4.1b show the combined mean values for precision and recall, respectively, across all four sequences per class for both the colour distance model and the GMM. Both techniques offer a similar level of performance with the colour distance model offering better results for precision but with the Gaussian mixture model having a marginally stronger recall performance. In terms of accuracy the colour distance model had a strong performance of 95.39%, a +3.02% increase from the GMM's 92.37% accuracy. The colour distance model also proved more consistent with the standard deviation of accuracy at 2.88% compared to the Gaussian model deviation of 7.13%, a +4.25% increase.

A comparison of these results with those of similar techniques found in the literature, as seen in Table 4.5, shows the performance to be within $1 \sim 2\%$. The two techniques from literature were chosen based on their use of the colour distance modelling technique, technique A[18], and the Gaussian mixture modelling technique, technique B[24].

Technique A uses a fusion of two colour spaces, HSI and L*a*b, to create a control space and primary space used to model the field colour. The data set comprised of 6050 frames spread over four clips, processed at a resolution of 88x60, with 41 frames manually annotated and used to generate the results. Technique B uses the expectation maximization algorithm to construct the GMM in the hue-luminance colour space (creating a 2-Dimensional model), post-process filtering occurs via the use of a region growing technique. Details about the data set were not given beyond the use of three clips from various sources.

The implemented colour distance model slightly outperformed that of A, the difference is however small enough to be accounted for by variance in the different data sets. The significant difference in resolutions between that used to generate the results of this work and the resolutions used to generate the results of A may require further investigation as to their impact on the results. The implemented GMM performed worse than that of B by a large enough margin to suggest possible inadequacies of the implemented technique.



(a) Precision

(b) Recall

Figure 4.1: Colour Distance Versus GMM, comparison by class

| | **CD** | **GMM** | **A**[1] | **B**[2] |
|---|---|---|---|---|
| **Accuracy** | **95.39%** | 92.37% | 95% | 94.34% |

[1] Colour Distance model based technique by [18].

[2] Gaussian Mixture model based technique by [24].

Table 4.5: Comparison of proposed works to those in literature for field colour modelling

## 4.2 Shot Classification

In this section, we present the results of the three features used for shot classification viz. field ratios, vertical projection and object size ratio; the multi-layer perceptron neural network classification results and finally the aggregated results of the classification techniques compared with results obtained from notable works in the literature. The results were generated by manually labelling each frame in the sequence as belonging to one of four shot classes; long, medium, close and out of field, and then using an automated procedure to compare the manually given labels to the labels given by each of the classification techniques. Results are given in the form of true positive (correct), false positive and false negative, followed by the subsequent precision and recall values for each class per sequence. A true positive occurs when the classifier correctly identifies a frame as belonging to the target class, thus the output class equals the target class. A false positive occurs when the classifier incorrectly identifies a frame from a different target class as belonging to the current output class. A false negative occurs when the classifier incorrectly identifies a frame from the current target class as belonging to a different output class.

A full set of results in the form of confusion matrices is supplied in Appendix B.

### 4.2.1 Field Ratios

The field ratio features are the features which establish a field to non-field ratio for the each shot in three different configurations; global, golden and even nine. The performance for each field ratio set is given individually showing the complete breakdown per sequence for each shot class along with both the precision and recall rates. Tables 4.6, 4.7 and 4.8 (pages 63, 64 and 65) show the performance of the global, golden and even field ratios respectively. Figures 4.2 and 4.3 (page 66) present a side by side comparison of the three ratio sets by shot class per video sequence for precision and recall respectively. Finally Figures 4.4 and 4.5 (page 67) show the 2-dimensional scatter plots per sequence for the golden ratio and even nine ratio set which plots the two classification inputs of mean and max ratio difference for each ratio set. This demonstrates how the classes are distributed in this space and provides insights into how classes may be separated using these inputs.

The results for the field ratio feature set show the even nine field ratio to have the highest overall performance with precision and recall rates of 81% and 83% while both global and golden ratios performed less accurately with precision and recall rates of 59%/64% and 79%/84% respectively. Comparing the even nine and the golden section ratios on a class by class basis, the golden section showed the highest performance for the long and medium shot classes by a slight margin while the even nine ratio performed significantly better with the out of field shot class. The close shot class was split between the two with the even nine ratio having a significantly better precision and the golden ratios having a significantly better recall. As the golden field ratio analyses only the central portion of the frame, this would suggest that the entire frame needs to be analysed, such as that done in the even nine ratios, when trying to identify shots with very few field coloured pixels, e.g. out of field shots.

Both the golden section and even nine ratios outperformed the global field ratio for every shot class, leaving the global ratio as an obsolete feature as better results can be achieved by the other techniques with little to no disadvantage.

Two particular individual cases stand out: the close shot class for sequence 2 and the medium shot class for sequence 3. The sequence 2 close shot class can be labelled as outlier data as a result of the relatively low number of close shots in the sequence compared to other shot types which has the effect of making its classification unreliable. So even though this class may have false positives comparable to other classes in the sequence the subsequent precision will be vastly lower due to low number of close shots and thus maximum number of possible true positives. The sequence 3 medium shot class is however, of greater concern. The best performing feature for this combination, the golden section, managed a precision of only 15.57%, over 40% below the average of 58.14% for that class. This suggested there were properties of this sequence which required further inspection.

Examining the sequence 3 2-D scatter plots in Figures 4.4 and 4.5 reveals the likely cause of such poor performance. Sequences 1, 2 and 4 show a degree of separation between the shot classes but in sequence 3 there is a significant degree of overlap between the medium and long shot classes. This makes it very difficult to separate the two classes in a classification process using only a 2-dimensional feature set, especially with a decision tree classifier. This suggests both the use of a higher dimensional feature set and a more advanced classifier. Figure 4.6 on page 68 expands the region occupied by the medium and long shot classes for both the golden section and even nine field ratios in order to further highlight the degree of overlap. Sequence 1 shows the least amount of overlap between classes and thus has the most distinct shot classes and consequently is the best performing sequence of the four.

The medium shot class of sequence 4 also appears to exhibit some interesting behaviour where two distinct groups have formed, more evident with the even nine ratio 2D scatter plot in Figure 4.5. One group with a low mean (0.2-0.4) and another group with a high mean (0.8-1.0). This suggests the consideration of a new class or subclass may be useful for certain broadcasts such as separating the medium class into two subclasses, the medium, field back and the medium, non-field backed class. The medium, field backed class would essentially represent the shot taken from a high angle but zoomed in to an area of the field which would still frame the players within the field and the medium, non-field backed class would represent the shot taken from field level looking across the field which would often include a large portion of audience or advertising boards in the background.

| **Class** | **Count** | **Correct** | **False** | | **Precision** | **Recall** |
|---|---|---|---|---|---|---|
| | | | ***Negative*** | ***Positive*** | | |
| **Sequence 1** | | | | | | |
| **Long** | 278 | 258 | 20 | 20 | 92.81% | 92.81% |
| **Medium** | 231 | 231 | 0 | 82 | 73.80% | 100.00% |
| **Close** | 70 | 17 | 53 | 20 | 45.95% | 24.29% |
| **Out of field** | 699 | 649 | 50 | 1 | 99.85% | 92.85% |
| **Sequence 2** | | | | | | |
| **Long** | 342 | 254 | 88 | 52 | 83.01% | 74.27% |
| **Medium** | 173 | 78 | 95 | 79 | 49.68% | 45.09% |
| **Close** | 7 | 6 | 1 | 98 | 5.77% | 85.71% |
| **Out of field** | 426 | 384 | 78 | 33 | 92.09% | 83.12% |
| **Sequence 3** | | | | | | |
| **Long** | 2552 | 1468 | 1084 | 252 | 85.35% | 57.52% |
| **Medium** | 174 | 5 | 169 | 314 | 1.57% | 2.87% |
| **Close** | 310 | 165 | 145 | 912 | 15.32% | 53.23% |
| **Out of field** | 238 | 158 | 80 | 0 | 100.00% | 66.39% |
| **Sequence 4** | | | | | | |
| **Long** | 840 | 568 | 272 | 134 | 80.91% | 67.62% |
| **Medium** | 282 | 212 | 70 | 539 | 28.23% | 75.18% |
| **Close** | 593 | 10 | 583 | 238 | 4.03% | 1.69% |
| **Out of field** | 131 | 131 | 0 | 14 | 90.34% | 100.00% |
| | | | **Average** | | | |
| | | | Long | | 85.52% | 73.05% |
| | | | Medium | | 38.32% | 55.78% |
| | | | Close | | 17.77% | 41.23% |
| | | | Out of Field | | 95.57% | 85.59% |
| | | | **Overall** | | 59.29% | 63.91% |

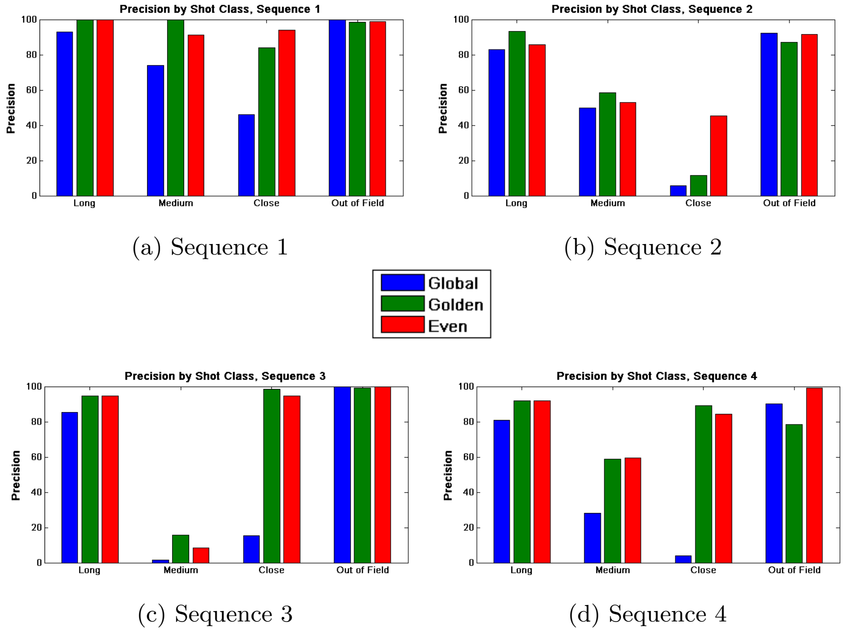
Table 4.6: Field Ratio classification performance, global field ratio

| **Class** | **Count** | **Correct** | **False** | | **Precision** | **Recall** |
|---|---|---|---|---|---|---|
| | | | ***Negative*** | ***Positive*** | | |
| **Sequence 1** | | | | | | |
| **Long** | 278 | 276 | 2 | 0 | 100.00% | 99.28% |
| **Medium** | 231 | 218 | 13 | 0 | 100.00% | 94.37% |
| **Close** | 70 | 68 | 2 | 13 | 83.95% | 97.14% |
| **Out of field** | 699 | 691 | 8 | 12 | 98.29% | 98.86% |
| **Sequence 2** | | | | | | |
| **Long** | 342 | 304 | 38 | 22 | 93.25% | 88.89% |
| **Medium** | 173 | 91 | 82 | 65 | 58.33% | 52.60% |
| **Close** | 7 | 7 | 0 | 53 | 11.67% | 100.00% |
| **Out of field** | 426 | 353 | 73 | 53 | 86.95% | 82.86% |
| **Sequence 3** | | | | | | |
| **Long** | 2552 | 2317 | 235 | 133 | 94.57% | 90.79% |
| **Medium** | 174 | 45 | 129 | 244 | 15.57% | 25.86% |
| **Close** | 310 | 292 | 18 | 5 | 98.32% | 94.19% |
| **Out of field** | 238 | 236 | 2 | 2 | 99.16% | 99.16% |
| **Sequence 4** | | | | | | |
| **Long** | 840 | 757 | 83 | 67 | 91.87% | 90.12% |
| **Medium** | 282 | 183 | 99 | 129 | 58.65% | 64.89% |
| **Close** | 593 | 509 | 84 | 61 | 89.30% | 85.83% |
| **Out of field** | 131 | 110 | 21 | 30 | 78.57% | 83.97% |
| | | | **Average** | | | |
| | | | Long | | 94.92% | 92.27% |
| | | | Medium | | 58.14% | 59.43% |
| | | | Close | | 70.81% | 94.29% |
| | | | Out of Field | | 90.74% | 91.21% |
| | | | **Overall** | | 78.65% | 84.30% |

Table 4.7: Field Ratio classification performance, golden field ratio

| **Class** | **Count** | **Correct** | **False** | | **Precision** | **Recall** |
|---|---|---|---|---|---|---|
| | | | ***Negative*** | ***Positive*** | | |
| | | | **Sequence 1** | | | |
| **Long** | 278 | 275 | 3 | 0 | 100.00% | 98.92% |
| **Medium** | 231 | 230 | 1 | 22 | 91.27% | 99.57% |
| **Close** | 70 | 61 | 9 | 4 | 93.85% | 87.14% |
| **Out of field** | 699 | 677 | 22 | 9 | 98.69% | 96.85% |
| | | | **Sequence 2** | | | |
| **Long** | 342 | 328 | 14 | 54 | 85.86% | 95.91% |
| **Medium** | 173 | 86 | 87 | 77 | 52.76% | 49.71% |
| **Close** | 7 | 5 | 2 | 6 | 45.45% | 71.43% |
| **Out of field** | 426 | 359 | 67 | 33 | 91.58% | 84.27% |
| | | | **Sequence 3** | | | |
| **Long** | 2552 | 2224 | 328 | 129 | 94.52% | 87.15% |
| **Medium** | 174 | 31 | 143 | 331 | 8.56% | 17.82% |
| **Close** | 310 | 305 | 5 | 17 | 94.72% | 98.39% |
| **Out of field** | 238 | 237 | 1 | 0 | 100.00% | 99.58% |
| | | | **Sequence 4** | | | |
| **Long** | 840 | 727 | 113 | 65 | 91.79% | 86.55% |
| **Medium** | 282 | 210 | 72 | 143 | 59.49% | 74.47% |
| **Close** | 593 | 480 | 113 | 89 | 84.36% | 80.94% |
| **Out of field** | 131 | 131 | 0 | 1 | 99.24% | 100.00% |
| | | | | **Average** | | |
| | | | Long | | 93.04% | 92.13% |
| | | | Medium | | 53.02% | 60.39% |
| | | | Close | | 79.59% | 84.48% |
| | | | Out of Field | | 97.38% | 95.18% |
| | | | **Overall** | | 80.76% | 83.04% |

Table 4.8: Field Ratio classification performance, even field ratio



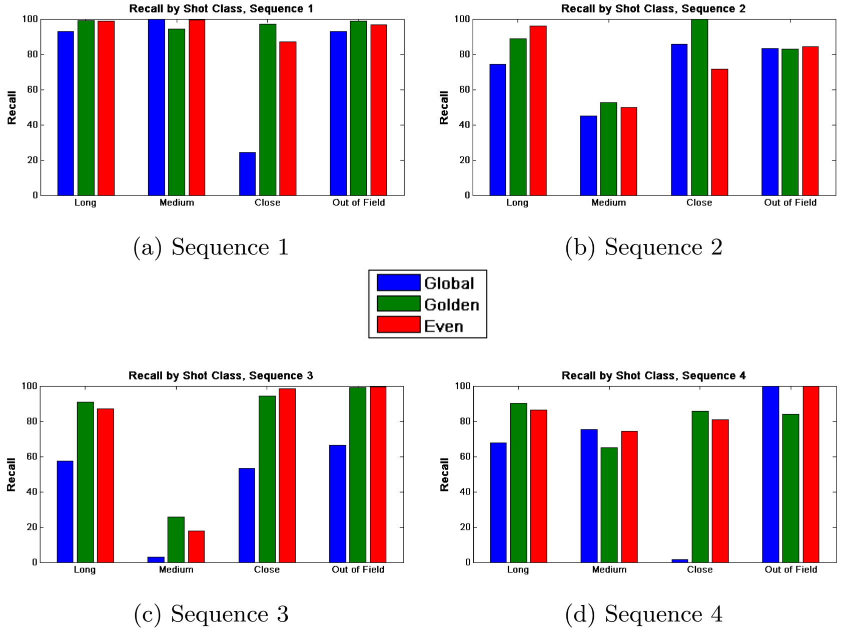
(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Figure 4.2: Precision by shot class for field ratios



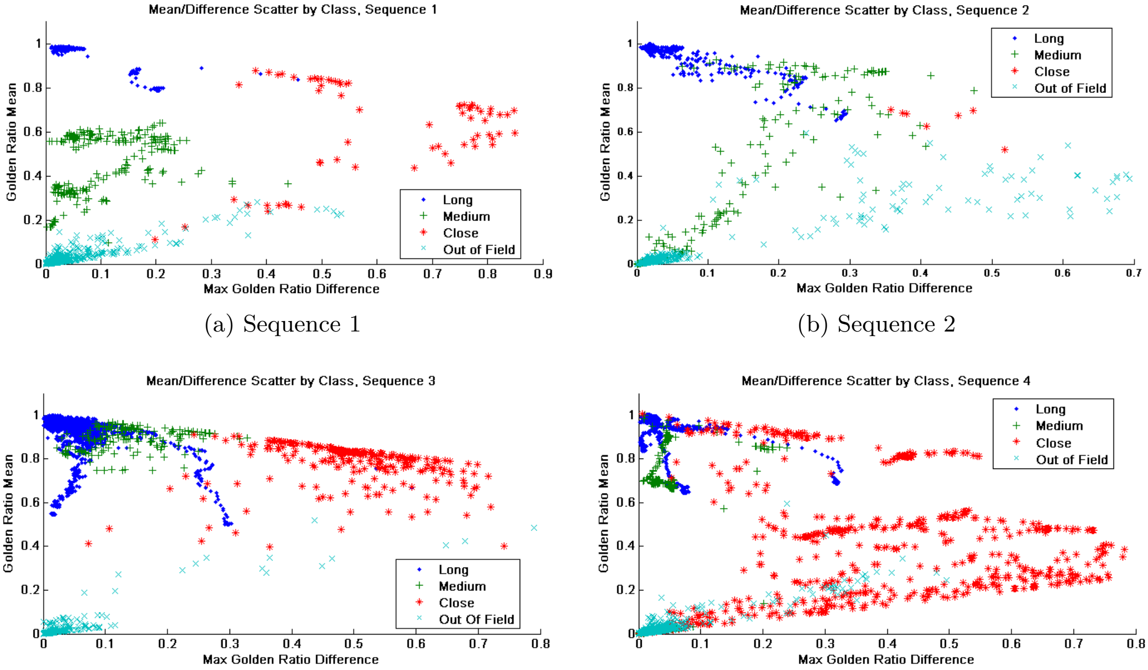
(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Figure 4.3: Recall by shot class for field ratios



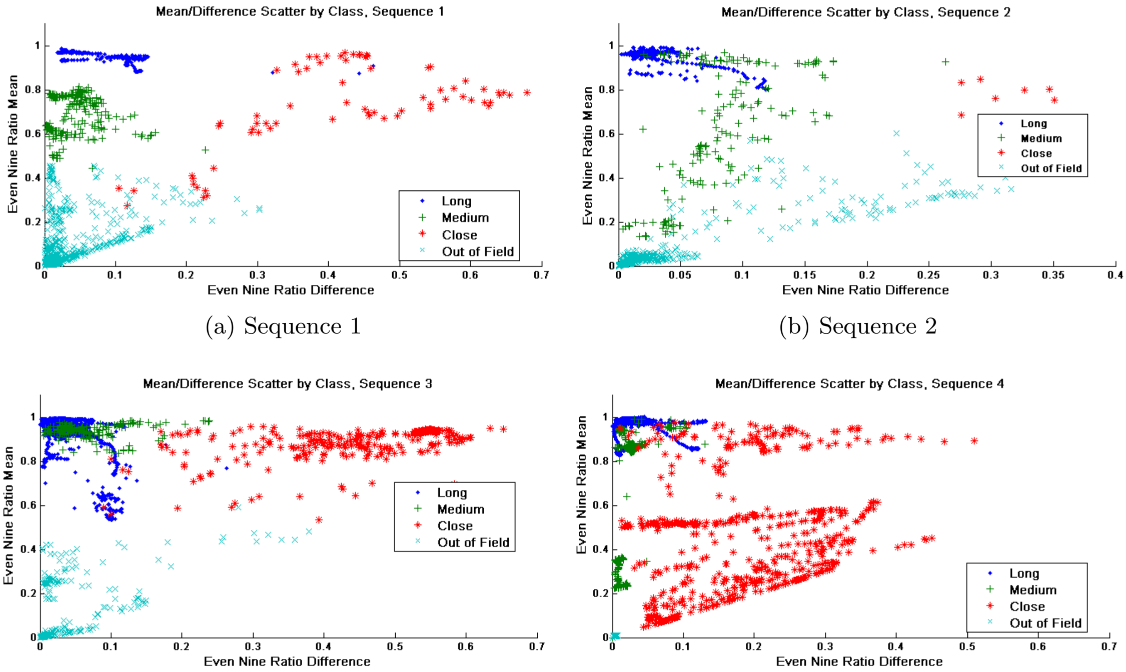
(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Figure 4.4: Mean/Difference 2D scatter plots per sequence by class for the golden section field ratios



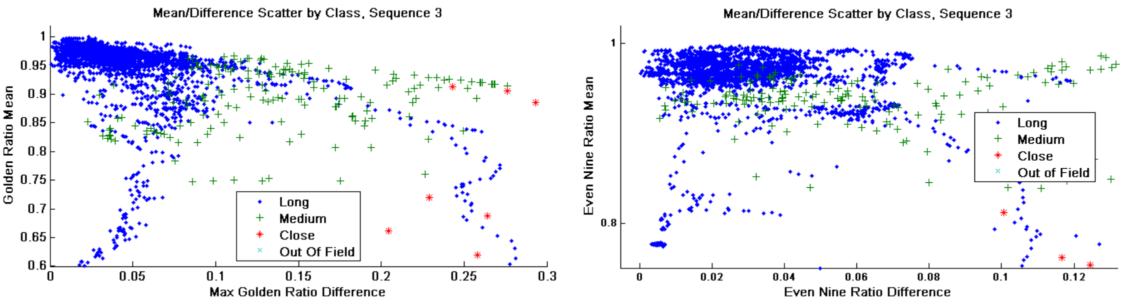
(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Figure 4.5: Mean/Difference 2D scatter plots per sequence by class for the even nine field ratios



(a) Golden Section Ratio

(b) Even Nine Ratio

Figure 4.6: Mean/Difference 2D scatter plots for sequence 3, medium/long shot class highlight

### 4.2.2 Vertical Projection

The vertical projection feature projects each horizontal line of pixels in the field mask on to a single vector which is then used to extract the mean projection and standard deviation of the projection. Table 4.9 presents the full performance breakdown, including precision and recall, across all video sequences. Figure 4.7 shows a side by side comparison of the respective precision and recall values for each sequence per class. Finally Figure 4.8 presents the 2-dimensional scatter plots of the vertical projection mean versus standard deviation, highlighting the regions of this space occupied by the different classes.

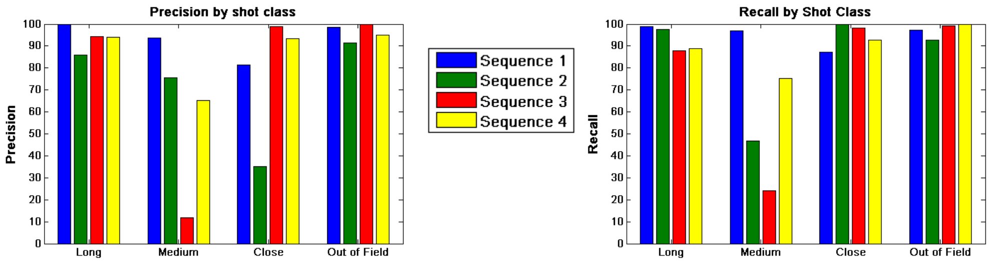
The vertical projection feature set provided fairly strong results with an overall accuracy of 82%/86% for precision and recall, giving it a better performance than the field ratios. The majority of gains achieved by the vertical projection resulted from the improved classification accuracy of the medium and close shot types over the even nine field ratio, showing a 9% increase in precision for the medium shot class and a 10% increase in recall for the close shot class.

Comparing the 2-dimensional scatter plots seen in Figures 4.4 (golden field ratio) and 4.8 (vertical projection) we can see a more prominent clustering of classes in the latter graphs corresponding to the vertical projection features than we see with the two features associated with the golden field ratio. This is a good indicator that the vertical projection and it's subsequent features may prove more powerful for shot classification. Given additional features and a higher order classification technique such as a SVM, the vertical projection may emerge as a strong candidate as a feature for more accurate shot classification.

In the same manner as the field ratio features, the medium shot class of sequence 3 remains problematic for the vertical projection. However, when observing the 2-D scatter plot for sequence 3 in Figure 4.8 the degree of class separation is far more prominent, and the degree of overlap far less than for the field ratio but this has not resulted in a better classification performance. In fact the vertical projection performed worse for this combination than the golden section field ratio. Seeing the classification procedure in Figure 3.24 in Chapter 3 shows the long and medium shot class to be separated solely by the vertical projection mean, corresponding to the y axis of the projection. For the other sequences this proves effective however for sequence 3 it is clear that such a strategy will not work and instead it would be more appropriate to separate the medium and long shots for the sequence by the standard deviation for the vertical projection, corresponding to the x axis. Doing so increases the classification performance of both classes significantly with the precision and recall increasing from 94%/87% to 98%/97% for long shots and from 12%/24% to 63%/75% for medium shots. Because a decision tree classifier is constructed using only general domain knowledge it can not adapt to or learn specific distributions in a way other techniques are able which again indicates the necessity of including more sophisticated techniques in the decision process.

| Class | Count | Correct | False | | Precision | Recall |
|---|---|---|---|---|---|---|
| | | | *Negative* | *Positive* | | |
| | | | **Sequence 1** | | | |
| **Long** | 278 | 275 | 3 | 0 | 100.00% | 98.92% |
| **Medium** | 231 | 224 | 7 | 15 | 93.72% | 96.97% |
| **Close** | 70 | 61 | 9 | 14 | 81.33% | 87.14% |
| **Out of field** | 699 | 680 | 19 | 9 | 98.69% | 97.28% |
| | | | **Sequence 2** | | | |
| **Long** | 342 | 334 | 8 | 55 | 85.86% | 97.66% |
| **Medium** | 173 | 81 | 92 | 26 | 75.70% | 46.82% |
| **Close** | 7 | 7 | 0 | 13 | 35.00% | 100.00% |
| **Out of field** | 426 | 395 | 31 | 37 | 91.44% | 92.72% |
| | | | **Sequence 3** | | | |
| **Long** | 2552 | 2240 | 312 | 138 | 94.20% | 87.77% |
| **Medium** | 174 | 42 | 132 | 311 | 11.90% | 24.14% |
| **Close** | 310 | 304 | 6 | 3 | 99.02% | 98.06% |
| **Out of field** | 238 | 236 | 2 | 0 | 100.00% | 99.16% |
| | | | **Sequence 4** | | | |
| **Long** | 840 | 747 | 93 | 48 | 93.96% | 88.93% |
| **Medium** | 282 | 212 | 70 | 113 | 65.23% | 75.18% |
| **Close** | 593 | 549 | 44 | 39 | 93.37% | 92.58% |
| **Out of field** | 131 | 131 | 0 | 7 | 94.93% | 100.00% |
| | | | **Average** | | | |
| | | | Long | | 93.51% | 93.32% |
| | | | Medium | | 61.64% | 60.78% |
| | | | Close | | 77.18% | 94.45% |
| | | | Out of Field | | 96.26% | 97.29% |
| | | | **Overall** | | 82.15% | 86.46% |

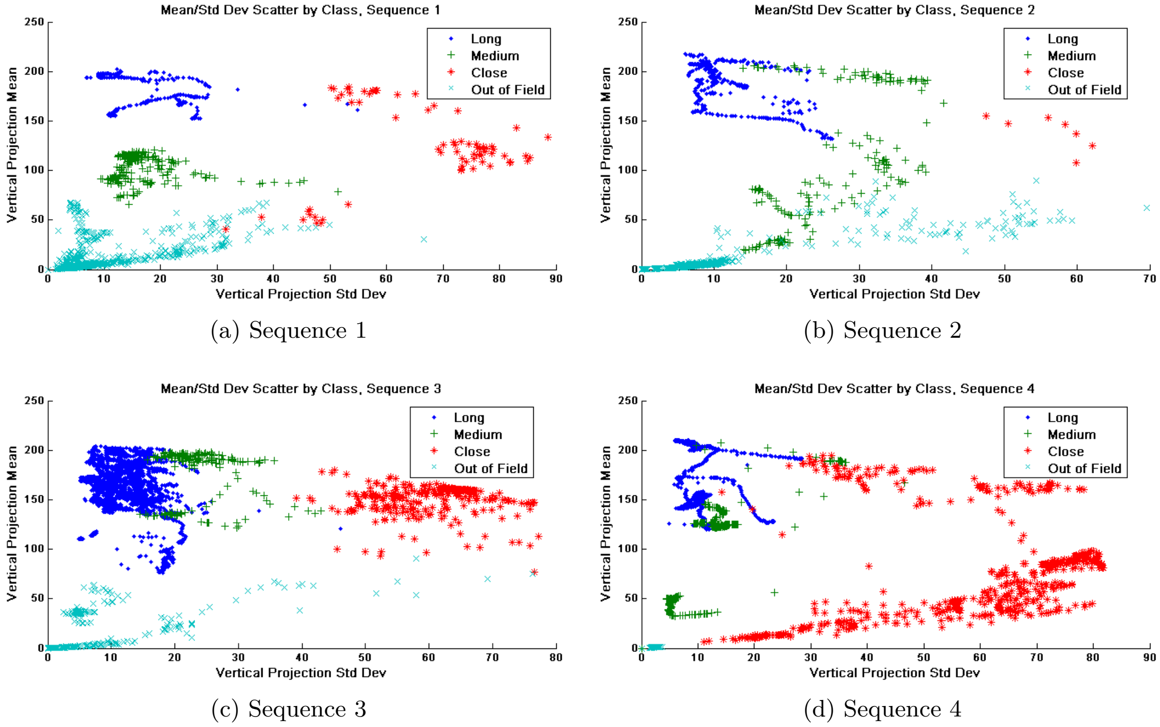
Table 4.9: Vertical Projection performance



(a) Precision

(b) Recall

Figure 4.7: Performance by shot class for vertical projection



(a) Sequence 1

(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

Figure 4.8: Mean/Std Dev 2D scatter plots per sequence by class for vertical projections

### 4.2.3 Size ratios

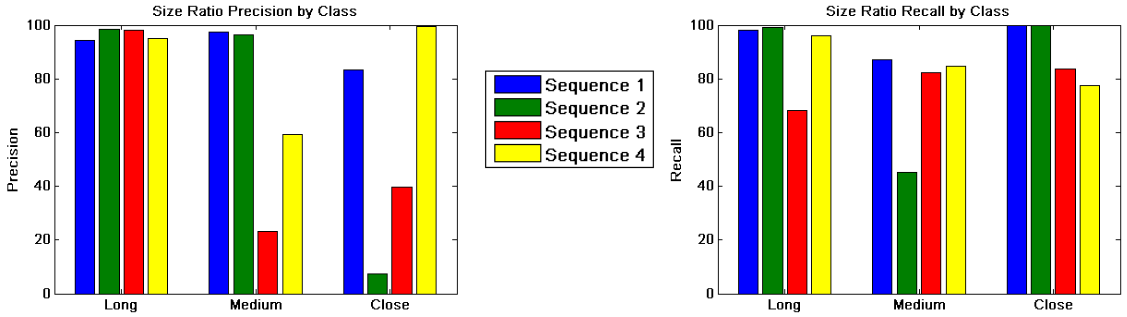
The size ratio feature estimates field of view by examining components in the field mask and comparing the component heights in relation to the frame height. Table 4.10 presents the full performance breakdown, including precision and recall, across all video sequences for the size ratio feature. Figure 4.9 shows a side by side comparison of the respective precision and recall values for each sequence per class.

The object size ratio feature again offers fair levels of accuracy with overall precision and recall values of 74% and 85% respectively. It's important to note that object size ratio is not capable of classifying out of field shots due to the lack of field coloured pixels and possible increase in noise, therefore these values are based on the classification of only three classes.

The medium shot class of sequence 3 again proved problematic with a low precision rate of 23%, although a recall rate of 82% was achieved which is significantly higher than that achieved by other features. The close shot class of sequence 3 also sees a marked decrease in precision when compared with other feature sets, similarly the recall performance of the medium shot class in sequence 2. Unlike the vertical projection feature the size ratio feature offers only a single dimension with which to classify shots, therefore there are no avenues to pursue the improvement of the performance for this instance. But given the performance attained using only a single value, the size ratio feature merits consideration for inclusion in combination with other features when classifying shots.

| **Class** | **Count** | **Correct** | **False** | | **Precision** | **Recall** |
|---|---|---|---|---|---|---|
| | | | ***Negative*** | ***Positive*** | | |
| **Sequence 1** | | | | | | |
| **Long** | 278 | 273 | 5 | 16 | 94.46% | 98.20% |
| **Medium** | 231 | 201 | 30 | 5 | 97.57% | 87.01% |
| **Close** | 70 | 70 | 0 | 14 | 83.33% | 100.00% |
| **Sequence 2** | | | | | | |
| **Long** | 342 | 339 | 3 | 5 | 98.55% | 99.12% |
| **Medium** | 173 | 78 | 95 | 3 | 96.30% | 45.09% |
| **Close** | 7 | 7 | 0 | 90 | 7.22% | 100.00% |
| **Sequence 3** | | | | | | |
| **Long** | 2552 | 1738 | 814 | 33 | 98.14% | 68.10% |
| **Medium** | 174 | 143 | 31 | 470 | 23.33% | 82.18% |
| **Close** | 310 | 259 | 51 | 393 | 39.72% | 83.55% |
| **Sequence 4** | | | | | | |
| **Long** | 840 | 807 | 33 | 42 | 95.05% | 96.07% |
| **Medium** | 282 | 239 | 43 | 165 | 59.16% | 84.75% |
| **Close** | 593 | 460 | 133 | 2 | 99.57% | 77.57% |
| | | | **Average** | | | |
| | | | Long | | 96.55% | 90.37% |
| | | | Medium | | 69.09% | 74.76% |
| | | | Close | | 57.46% | 90.28% |
| | | | **Overall** | | 74.37% | 85.14% |

Table 4.10: Size ratio performance



(a) Precision

(b) Recall

Figure 4.9: Performance by shot class for size ratios

### 4.2.4 Neural Network

Table 4.11 presents the full performance breakdown, including precision and recall, across all video sequences for the combined feature vector using a multi-layer perceptron neural network (MLP-NN).

This technique managed consistent levels of performance across all sequences and classes with precision values above 73% and recall values above 81% with an average precision and recall across all classes of 91.43% and 95.69% respectively. The classifier behaved consistently with only the expected minor variations between multiple training/classification runs of the same sequence. Sequence 2 proved challenging allowing for instances of a poorly trained classifier due to the very limited number of close shots in the sequence. This type of scenario is unlikely to occur in a typical full match sequence.

When looking at individual class performance per sequence the medium shot class of sequence 3 still remains the most difficult to classify but is only a 5% decrease in precision from the next most difficult sequence. This still shows a greater than 50% increase in precision and recall for that combination, in most cases, when compared to using features individually. Figure 4.10 shows that nearly all the misclasifications for this sequence were between the medium and long shot classes. This suggests there are still some shot properties not accounted for in the feature set, and that the discovery and extraction of these properties will increase the performance of the shot classification process. The most promising is likely the accurate detection of the players themselves as they are often the most relevant objects in the frame thus providing the most salient information regarding a shot composition.

| **Class** | **Count** | **Correct** | **False** | | **Precision** | **Recall** |
|---|---|---|---|---|---|---|
| | | | ***Negative*** | ***Positive*** | | |
| **Sequence 1** | | | | | | |
| **Long** | 278 | 275 | 3 | 0 | 100.00% | 98.92% |
| **Medium** | 231 | 230 | 1 | 0 | 100.00% | 99.57% |
| **Close** | 70 | 69 | 1 | 23 | 75.00% | 98.57% |
| **Out of field** | 699 | 697 | 20 | 2 | 99.71% | 97.21% |
| **Sequence 2** | | | | | | |
| **Long** | 342 | 306 | 36 | 1 | 99.67% | 89.47% |
| **Medium** | 173 | 169 | 4 | 49 | 77.52% | 97.69% |
| **Close** | 7 | 7 | 0 | 2 | 77.78% | 100.00% |
| **Out of field** | 426 | 412 | 14 | 2 | 99.52% | 96.71% |
| **Sequence 3** | | | | | | |
| **Long** | 2552 | 2507 | 45 | 32 | 98.74% | 98.24% |
| **Medium** | 174 | 142 | 32 | 50 | 73.96% | 81.61% |
| **Close** | 310 | 305 | 5 | 9 | 97.13% | 98.39% |
| **Out of field** | 238 | 229 | 9 | 0 | 100.00% | 96.22% |
| **Sequence 4** | | | | | | |
| **Long** | 840 | 783 | 57 | 36 | 95.60% | 93.21% |
| **Medium** | 282 | 253 | 29 | 69 | 78.57% | 89.72% |
| **Close** | 593 | 566 | 27 | 1 | 99.82% | 95.45% |
| **Out of field** | 131 | 131 | 0 | 7 | 94.93% | 100.00% |
| | | | **Average** | | | |
| | | | Long | | 98.50% | 94.96% |
| | | | Medium | | 82.51% | 92.15% |
| | | | Close | | 87.43% | 98.10% |
| | | | Out of Field | | 98.54% | 97.54% |
| | | | **Overall** | | 91.75% | 95.69% |

Table 4.11: Combined performance using a MLP-NN classifier

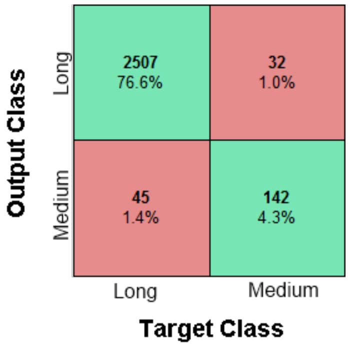


Figure 4.10: Isolated confusion matrix for the medium and long shot classes of sequence 3

### 4.2.5 Comparison

Table 4.11 combines all the previously discussed feature set classification results (**F**ield **R**atio[1], **V**ertical **P**rojection and **S**ize **R**atio) along with the MLP-NN classification results. The results have been aggregated across all video sequences and compared with the results obtained in four notable papers using similar techniques [18, 36, 44, 45]. Some techniques did not classify out of field shots and thus have been marked as NA.

Various data sets were used in these papers with only [18] and [44] appearing to share the same or similar data sets. Both papers use video with a resolution of 352x288, processed at a resolution of 88x60. [18] further details the data set as consisting of 6050 frames, with 41 frames used to evaluate performance. [44] mentions only the length of the video being 49mins. [36] uses video with resolutions ranging from 512x384 to 480x360, together making up 4000 frames and 1000 shots, 25% of which are used for training purposes. [45] provides only the frame count of 14000 and source, 2002 FIFA World Cup, of the data set.

The results show the MLP-NN using the 5 feature set combination as performing very well compared to using each feature set separately within a decision tree structure and even compared to those techniques in literature. The MLP-NN achieved the best precision for both the long and out of field shot classes and achieved the best recall for the close and out of field shot classes. Second best precision and recall rates were also achieved for the medium shot class. Comparing the MLP-NN to the individual features an average increase in precision can be seen across all classes by +10.78% and an increase in recall by +9.88% with the medium shot class showing the largest single class increase with precision and recall going up by +21.26% and +26.84% respectively. The individual features managed a strong performance for the long and out of field shot class outperforming other methods in both precision and recall for the out of field shot class and offering comparable or better precision and recall for the long shot class.

This comparison serves to illustrate the virtues of the proposed methods but is still largely cursory and a more thorough analysis of the various techniques would be required to definitely address the advantages and disadvantages of each technique. A comparison of this nature would need to establish a standardized definition for each shot class and process identical data sets with each technique. This is outside the scope of this work but would be a candidate for future endeavours.

[1] For the field ratio only the even nine field ratio results have been displayed

| | **Precision** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | FR | VP | SR | MLP-NN | [18] | [44] | [36] | [45] |
| **Long** | 93.04% | 93.51% | 96.55% | **98.50%** | 83.48% | 84.7% | 96.5% | 83.53% |
| **Medium** | 53.02% | 61.64% | 69.09% | 82.51% | 81.25% | 67.53% | **93.0%** | 81.57% |
| **Close** | 79.59% | 77.18% | 57.46% | 87.43% | 96.63% | 93.9% | 94.0% | **97.12%** |
| **Out of Field** | 97.38% | 96.26% | NA | **98.54%** | NA | NA | 92.0% | 92.58% |
| | **Recall** | | | | | | | |
| **Long** | 92.13% | 93.32% | 90.37% | 94.96% | 96.0% | 69.3% | **96.5%** | 92.03% |
| **Medium** | 60.39% | 60.78% | 74.76% | 92.15% | 78.0% | 76.4% | **94.7%** | 84.85% |
| **Close** | 84.48% | 94.45% | 90.28% | **98.10%** | 86.0% | 73.9% | 92.0% | 91.75% |
| **Out of Field** | 95.18% | 97.29% | NA | **97.54%** | NA | NA | 94.0% | 89.8% |

Table 4.12: Comparison of proposed works to those in literature for shot classification

# Chapter 5

# Conclusion and Future Work

This chapter concludes this dissertation by summarising the ideas and achievements in this work for furthering the goal of shot classification in broadcast soccer video. Various aspects of the field colour modelling techniques are discussed regarding how each technique performed relative to each other and the techniques in literature, the strength and weakness of each technique and finally if these techniques may be used in other sports or to solve similar problems. The shot classification discussion focuses on three areas; the level of success achieved in classifying the different shot types, the performance of each feature/classifier and the possible limitations of the features when applying them to other sports. A final retrospective is given along with suggestions for avenues of exploration to improve shot classification in broadcast soccer video.

## 5.1 Field Colour Modelling

This work has shown how two techniques, the colour distance model and Gaussian mixture model, can be used to model dominant colours for the purpose of modelling the play-field area in broadcast soccer videos. The results in Section 4.1 show both the field colour modelling techniques to have achieved a high level of accuracy (92% ~ 95%) on par (within 2%) with similar techniques found in the literature. The colour distance model did however, slightly outperform the Gaussian mixture model in terms of both accuracy and consistency across sequences and shot classes.

There are numerous specific factors to consider when evaluating the performance of these techniques. This specific implementation of the Gaussian mixture modelled only the hue channel due to limitations of computation time. If these concerns are not relevant, the model can be further expanded to encompass all colour channels thus increasing accuracy.

The video sequences themselves were selected based on a varied distribution of shot types and only formed a portion of the full length of a typical soccer match. Thus the robustness of the techniques were not rigorously tested regarding changing global illumination over the course of a match. The adaptability of the Gaussian mixture model technique does provide an advantage in that it can incorporate global illumination changes. However, this is an area of concern as determining the rate at which the model adapts to new data, such as gradual changes, affects the way noisy or transitory data alters the model. The higher the learning rate the better the model adapts to change but as a consequence the model will also learn noisy data more quickly and thus possibly deteriorate the model's accuracy. The GMM requires various parameters to be configured correctly which requires knowledge of the modelled data. Such a requirement would not be feasible for real world production. Therefore to consider such a technique would also require the development of methods to facilitate the automatic configuration of the GMM.

The colour distance model as implemented lacks an update process and instead incorporates a training phase which may address certain issues arising from global illumination changes if the training data included samples from various portions of the sequence but it is uncertain what impact that may have on overall accuracy. The need for a training process is itself a disadvantage since the training sequence needs to be manually selected. Alternatively certain update schemes were discussed in Section 3.1.1 which could potentially add an element of adaptability to this technique. Although the Euclidean distance measure was employed for this work it is not the only distance that can be used. Mahalanobis distance can be considered as it incorporates the data distribution into the distance measurement. Compared to the GMM, the colour distance model is relatively simple to configure with only a single parameter, the leniency factor, which required adjustment between sequences. While it is possible to adjust other model parameters this was not required to produce any of the results seen in this work. These parameters were assumed to have only minor impact on the performance and were thus not investigated but it would be prudent to confirm this assumption in future work.

Both techniques provide a solid foundation from which further processing and analysis may proceed as both offer a reliable, accurate and extensible solution to the task of field modelling.

Extending this work in field modelling beyond broadcast soccer video to video of other sports, broadcast or otherwise, is both feasible and natural. Any sport which is played on a large uniformly coloured field, e.g. tennis, cricket, (field) hockey, are candidate sports for field colour modelling. Indeed numerous papers have been written which examines the applications of field modelling for these sports [7, 17, 18, 24, 31, 37]. The key limitation for these techniques would be the requirement for the dominant view of the sequence to include as it's main constituent a uniformly coloured background, typically for sports this represents the playing field. The required degree of uniformity is dependant on the technique used, as certain techniques, such as GMM, are able to represent multi-modal distributions and therefore have the ability to model backgrounds with more than one dominant colour.

With regard to general background modelling in a fixed view, fixed camera environment it may be possible to adapt these techniques to work along side or supplemental to traditional background modelling techniques. Perhaps incorporating a windowing scheme to establish areas of local colour dominance and using the information to separate not only foreground objects from background objects but even separate background or foreground objects from each other. The ability of these techniques to establish pixel relationships through colour space could be a property worth considering for applications of general background modelling.

## 5.2 Shot Classification

Three sets of features (field ratio, vertical projection and object size ratio) have been evaluated regarding their ability to perform shot classification in broadcast soccer video. Each feature set was evaluated independently using a decision tree classifier and then combined for use with a multi-layer perceptron neural network (MLP-NN).

The long shot class proved easier to classify than other shot types with all the features showing precision and recall rates greater than 90%. This may be due to the more consistent nature of long shots as typically there are fewer variables influencing the shot's composition, or at least the changes are more subtle. That is because of the 'zoomed' out nature of long shots, objects in the field generally occupy only a small area of the shot compared to the area occupied by the field therefore objects entering or leaving the shot have a significantly smaller impact on the shot composition for long shots than for other shot types. The out of field shot class also managed precision and recall rates near or above 90%, a strong performance but it is worth noting that certain higher level event classification techniques require further classification of out of field shots into audience and player shots which could have a different semantic relevance. So while identifying out of field shots may be reliable using these techniques, further classification of this shot type would require new features or the current features would need to be adapted for this problem.

The medium and close shot types proved the most difficult shots to classify, largely due to the varied nature of the shots and the many variables which may affect the shot's composition. This was exemplified by the difficulty in classifying the medium shot class of sequence 3 using any of the features. Unlike long shots which are typically captured from cameras in a fixed position at a fixed height from the field, medium and close shot are often captured from multiple cameras at multiple locations around the field and located at varying heights and possibly changing position between shots. The limited field of view also means small changes in the scene can have a huge effect on the composition of the shot, e.g. if a player walks into a close shot the number of field pixels can change by as much as 100%.

Of the three different field ratio feature sets the even nine ratio set achieved the highest overall performance with precision and recall values of 81% and 83% respectively. The golden section ratio achieved slightly lower overall results with precision and recall values of 79% and 84% respectively but in specific instances outperformed the even nine ratio. The global ratio performed poorly, managing overall precision and recall rates of only 59% and 64%, significantly lower than the other two ratio sets and showing no instances in which it out performs the others. Therefore the use of the even nine or golden section ratio sets will always be preferable to that of the global ratio.

The vertical projection feature set managed to outperform the field ratio feature sets by achieving an overall precision and recall of 82% and 86%. Crucially the vertical projection feature set improved classification performance over the field ratio feature sets for the medium and close shot classes. Additionally observing the 2-D scatter plots for both the field ratios and the vertical projection, showed the vertical projection to have a more pronounced separation of shot classes in the feature space and thus a greater potential for shot classification given an appropriate classifier.

The object size ratio feature achieved an overall precision and recall of 74% and 85%, surprisingly high given shot classes were separated by only a single value. As previously mentioned this feature set was not used to classify out of field shots and therefore the results reflect only the classification of the long, medium and close shot classes. Considering this features ability to classify shots even within its 1-dimensional feature space, including this feature as part of a higher dimensional feature set is recommend.

When the features were combined and used as an input to a MLP-NN, performance across all shot classes increased for both precision and recall by an average of 10.78% and 9.88% respectively. The largest increase was seen by the medium shot class with an increase in precision of 21.26% and an increase in recall of 26.84% which would suggest that the features offered greater discriminatory power than the specific configuration of the decision trees would allow for this shot class. This technique proved comparable to results obtained from similar techniques shown in Section 4.2.5, demonstrating the highest precision and recall rates for a number of classes and achieving an overall precision and recall rate of 91.75% and 95.69% respectively.

The limitation of these features is that they all operate on the field model which is appropriate for broadcast soccer video but careful consideration would need to be given before extending these features to other broadcast sports video. Considerations such as how the field and actors within the field are framed in different shots or if the features adequately represent the distinct properties of the desired shots? Even the use of the field as a background model can be challenged and should techniques exist to generate a background model incorporating other aspects of a shot, it could instead be possible to extract the features discussed in this dissertation from this new model.

Extending beyond sports video and extending these features for use while classifying shots in general video may be possible but such a broad topic is beyond the scope of this work and warrants its own discussion.

## 5.3 Future Work

All the techniques evaluated have potential for further improvement either through a richer analysis or from expanding on the ideas already established.

Both of the background modelling techniques proved very capable at modelling the play-field although lacking in certain aspects, aspects which could be built into these techniques to improve performance. The colour distance modelling technique would benefit from a strong update procedure which could help the technique maintain a high level of performance throughout an extended video sequence. The update procedure would, however, have to take care not to learn noisy or incorrect data which would reduce the accuracy of the model. The need for a manually selected training sequence is one of the disadvantages of this technique, therefore any improvement that would remove this requirement or automate the process would be advantageous. The Gaussian mixture model can be expanded by including not just the hue channel but also other colour channels which may increase the accuracy at the expense of computational complexity. A major difficulty of the Gaussian mixture model is the selection of model parameters, which determine the attributes of newly formed Gaussians and the rate at which new observations increase the weight of matching Gaussians (the learning rate). Properly configuring these parameters requires knowledge of the modelled data, which is an unreasonable restriction in most cases. Thus methods need to be investigated which are capable of either automatically configuring or assisting in the configuration of such parameters.

Each of these techniques focus on modelling the play-field through the colour space but supplementing the model with other descriptors such as shape and texture information may increase the performance of the models. Filtering based on shape has been discussed in this work but it was shown that simple aspect ratio with component size filtering was not sufficient to adequately separate field and non-field areas. Further research is needed on this topic with a focus on more detailed shape descriptors. Texture filtering has not been mentioned in this work but texture information may be useful both for pre- and post-filtering potential field candidates.

As previously mentioned, the colour distance model and the Gaussian mixture model operate within the colour space to model the field colour. In sports videos, however, the field colour is not the only semantically important colour. The colours of player's uniforms are generally distinct between teams and thus conveys information. Using colour modelling techniques to model and identify player uniform colours would prove useful in the task of content analysis. Once player uniforms can be modelled and identified it then becomes possible to attempt tracking of players on the field. For both modelling and tracking a player the type of shot analysed would no doubt become important.

Three shot classification feature sets were used and all but the global field ratio proved useful. However, the vertical projection feature set did display the highest performance and showed the greatest potential for improvement with numerous other features being readily extractable from the projection. These features were all extracted from the field model but there are other features which may be used for shot classification such as features extracted from the play-field lines or advertising boards. Such features could assist in shot localisation and subsequently provide information about the type of shot. Another feature which would benefit from additional information is the object size ratio. Because the connected component analysis was performed merely on the foreground mask produced by the field modelling techniques the analysis was limited in its ability to separate objects not framed by the field. If it were possible to identify and separate players (and referees) from audience and non-field background a far more accurate size ratio would be possible.

The use of the MLP-NN in conjunction with a combined feature vector provided a significant classification performance increase over the single feature set with decision tree combination. This indicates that further investigation and use of higher order classification techniques such as SVMs, Bayesian classifiers and neural networks is a prudent and sensible next step in evolving a shot classification system.

A shot type or rather a pseudo shot type which was not explored in this work is the slow-motion or action replay shot. While not strictly a separate shot type the use of these production techniques can be very useful for event classification as the presence of a replay will often indicate the proximity of an important event.

Beyond broadcast soccer video, expanding this work into other types of sports video would be a worthwhile endeavour which could lead to a better understanding of the features and possibly the relationship between the various sports. The ultimate goal though for shot classification is as a facilitator for event detection in sports video, and thus investigating classification outputs as inputs into an event detection system would be of crucial importance to future work.

# References

[1] 3D Color Inspector/Color Histogram, 03 2012. URL imagej.nih.gov/ij/plugins/color-inspector.html.

[2] Hawk-Eye, 08 2012. URL www.hawkeyeinnovations.co.uk.

[3] IBM VideoAnnEx, 07 2012. URL www.research.ibm.com/VideoAnnEx/.

[4] MATLAB, 08 2012. URL www.mathworks.com/products/matlab/.

[5] OpenCV developer portal, 03 2012. URL code.opencv.org.

[6] OpenCV wiki, 03 2012. URL opencv.willowgarage.com.

[7] J. Assfalg, M. Bertini, C. Colombo, A. Del Bimbo, and W. Nunziati. Semantic annotation of soccer videos: automatic highlights identification. *Computer Vision and Image Understanding*, 92(2-3):285–305, 2003.

[8] J. Assfalg, M. Bertini, C. Colombo, and A.D Bimbo. Semantic annotation of sports videos. *IEEE Multimedia*, 9(2):52–60, 2002.

[9] L. Bai, S. Lao, W. Zhang, and G. Jones. A semantic event detection approach for soccer video based on perception concepts and finite state machines. *Image Analysis for Multimedia Interactive Services*, pages 2–5, 2007.

[10] F.N. Bezerra and E. Lima. Low cost soccer video summaries based on visual rhythm. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 71–78. ACM, 2006. ISBN 1595934952.

[11] D. Brezeale and D.J. Cook. Automatic video classification: A survey of the literature. *Systems, Man, and Cybernetics*, pages 1–16, 2008.

[12] F. Chang, C.J. Chen, and C.J. Lu. A linear-time component-labeling algorithm using contour tracing technique. *Computer Vision and Image Understanding*, 93(2):206–220, 2004.

[13] A. Chianese, R. Miscioscia, V. Moscato, S. Parlato, and A. Picariello. A fuzzy approach to video scene detection and its application for soccer matches. In *Proceedings of the 4th International Conference on Intelligent Systems Design and Application*, 2004.

[14] F. Coldefy and P. Bouthemy. Unsupervised soccer video abstraction based on pitch, dominant color and camera motion analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 268–271, 2004.

[15] D. Comaniciu and P. Meer. Mean shift analysis and applications. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1197–1203. IEEE, 1999.

[16] L.Y. Duan, M. Xu, T.S. Chua, Q. Tian, and C.S. Xu. A mid-level representation framework for semantic sports video analysis. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 33–44, 2003.

[17] L.Y. Duan, M. Xu, and Q. Tian. Semantic shot classification in sports video. In *Proc. SPIE Storage and Retrieval for Media Database*, pages 300–313, 2003.

[18] A. Ekin, A.M. Tekalp, and R. Mehrotra. Robust dominant color region detection with applications to sports video analysis. In *Proceedings of IEEE ICIP*, volume 1, pages 21–24, 2003.

[19] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874, June 2006. ISSN 01678655.

[20] A.R.J. François and G.G. Medioni. Adaptive color background modeling for real-time segmentation of video streams. In *Proceedings of the International Conference on Imaging Science, Systems, and Technology*, pages 227–232, 1999.

[21] T. Gevers and A.W.M. Smeulders. Color-based object recognition. *Pattern Recognition*, 32:453–464, 1999.

[22] C.L. Huang, H.C. Shih, and C.Y. Chao. Semantic analysis of soccer video using dynamic Bayesian network. *Multimedia, IEEE Transactions on*, 8(4):749–760, 2006.

[23] J.L. Jian, M.H. Hung, C. Hsieh, and Y. Chang. Real-time scene classification for baseball videos. In *18 th IPPR Conf. on Computer Vision, Graphics and Image Processing*, 2005.

[24] S. Jiang, Q. Ye, and W. Gao. A new method to segment playfield and its applications in match analysis in sports video. *Proceedings of the 12th Annual ACM*, 1:292–295, 2004.

[25] Y.L. Kang, J.H. Lim, Q. Tian, and M.S. Kankanhalli. Soccer video event detection with visual keywords. 3:1796–1800, 2003.

[26] S.H. Khatoonabadi and M. Rahmati. Automatic soccer players tracking in goal scenes by camera motion elimination. *Image and Vision Computing*, 27(4):469–479, March 2009.

[27] D.S. Lee. Effective Gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):827–832, 2005.

[28] J. Li, T. Wang, W. Hu, M. Sun, and Y. Zhang. Soccer highlight detection using two-dependence Bayesian network. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1625–1628. IEEE, 2006.

[29] T. Lin and H.J. Zhang. Automatic video scene extraction by shot grouping. *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, 4:39–42, 2000.

[30] J. Liu, X. Tong, W. Li, T. Wang, Y. Zhang, and H. Wang. Automatic player detection, labeling and tracking in broadcast soccer video. *Pattern Recognition Letters*, 30(2): 103–113, January 2009.

[31] Y. Liu, S. Jiang, Q. Ye, W. Gao, and Q. Huang. Playfield detection using adaptive GMM and its application. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 421–424, 2005.

[32] M. Luo, Y.F. Ma, and H.J. Zhang. Pyramidwise structuring for soccer highlight extraction. In *Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 2, pages 945–949. IEEE, 2003.

[33] A.M. McIvor. Background subtraction techniques. *Proc. of Image and Vision Computing*, (1), 2000.

[34] G. Millerson. *The Technique of Television Production (Library of Communication Techniques).* Focal Pr, 1990. ISBN 0240512898.

[35] M.F. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525 – 533, 1993. ISSN 0893-6080.

[36] N. Nan, G. Liu, and X. Qian. An SVM-based soccer video shot classification scheme using projection histograms. *Advances in Multimedia Information*, pages 883–886, 2008.

[37] S. C. Pei and F. Chen. Semantic scenes detection and classification in sports videos. In *Proceedings of IPPR Conference on Computer Vision, Graphics and Image Processing (CVGIP)*, pages 210–217, 2003.

[38] M. Piccardi. Background subtraction techniques: a review. *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, 4:3099–3104, 2004.

[39] R. Ren and J. M. Jose. Football video segmentation based on video production strategy. *Advances in Information Retrieval*, pages 433–446, 2005.

[40] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2, 1999.

[41] H. Sun, J.H. Lim, and Q. Tian. Semantic labeling of soccer video. In *Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1787–1791 vol.3, 2003.

[42] K. Suzuki, I. Horiba, and N. Sugie. Linear-time connected-component labeling based on sequential local operations. *Computer Vision and Image Understanding*, 89(1):1–23, 2003.

[43] Y. Tabii and R. Oulad Haj Thami. A framework for soccer video processing and analysis based on enhanced algorithm for dominant color extraction. *International Journal of Image Processing (IJIP)*, 3(4):131–142, 2009.

[44] A. M. Tekalp and A. Ekin. Automatic soccer video analysis and summarization. *Image Processing, IEEE Transactions on*, 12(7):796–807, 2003.

[45] X. Tong, Q. Liu, and H. Lu. Shot classification in broadcast soccer video. *ELCVIA*, 7 (1), 2008.

[46] V. Tovinkere and R.J. Qian. Detecting semantic events in soccer games: Towards a complete solution. In *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, volume 1, pages 833–836, 2001.

[47] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 255–261, 1999. ISBN 0769501648.

[48] L. Wang, B. Zeng, S. Lin, G. Xu, and H.Y. Shum. Automatic extraction of semantic colors in sports video. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 3, pages 617–620, 2004. ISBN 0780384849.

[49] B. Yang, L.F. Sun, F. Wang, Peng Wang, and S.Q. Yang. Mid-level descriptors extraction of soccer video with domain knowledge. In *Systems, Man and Cybernetics, 2006. SMC'06. IEEE International Conference on*, volume 6, pages 4937–4941, 2006.

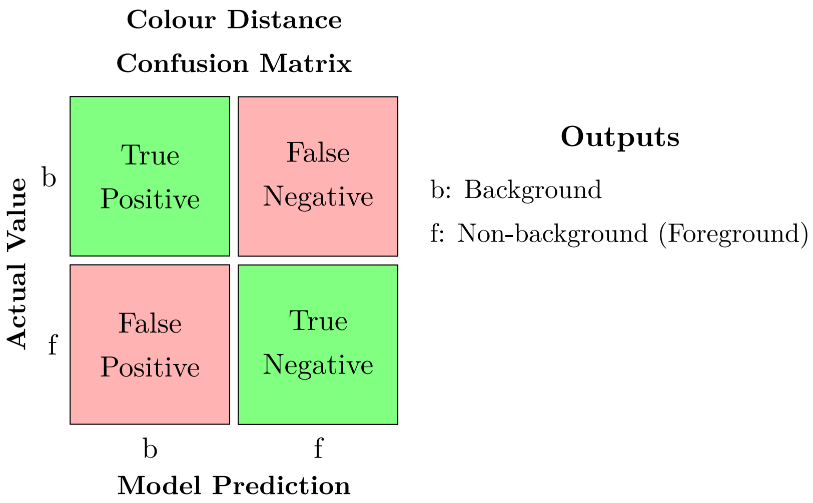
[50] H.S. Yoon, Y.J. Bae, and Y. Yang. A soccer image sequence mosaicking and analysis method using line and advertisement board detection. *ETRI*, 24(6):443–454, 2002.

[51] X. Yu and D. Farin. Current and emerging topics in sports video processing. In *IEEE International Conference on Multimedia and Expo, 2005. ICME 2005*, pages 526–529, 2005.

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

# Appendix A

# Background Modelling Performance



**Colour Distance Confusion Matrix**

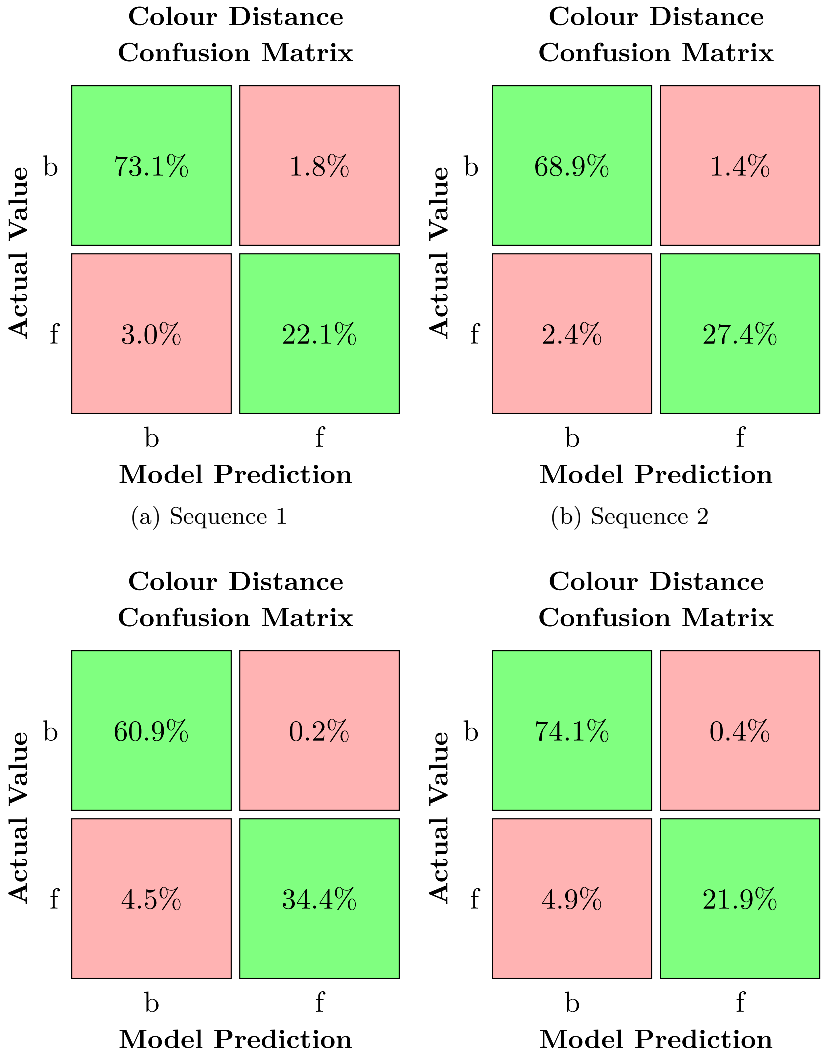
| Actual Value | | Model Prediction | |
|---|---|---|---|
| | | b | f |
| | b | True Positive | False Negative |
| | f | False Positive | True Negative |

**Outputs**

b: Background

f: Non-background (Foreground)

Figure A.1: Confusion Matrix Template



**Colour Distance Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 73.1% | 1.8% |
| f | 3.0% | 22.1% |

(a) Sequence 1

**Colour Distance Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 68.9% | 1.4% |
| f | 2.4% | 27.4% |

(b) Sequence 2

**Colour Distance Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 60.9% | 0.2% |
| f | 4.5% | 34.4% |

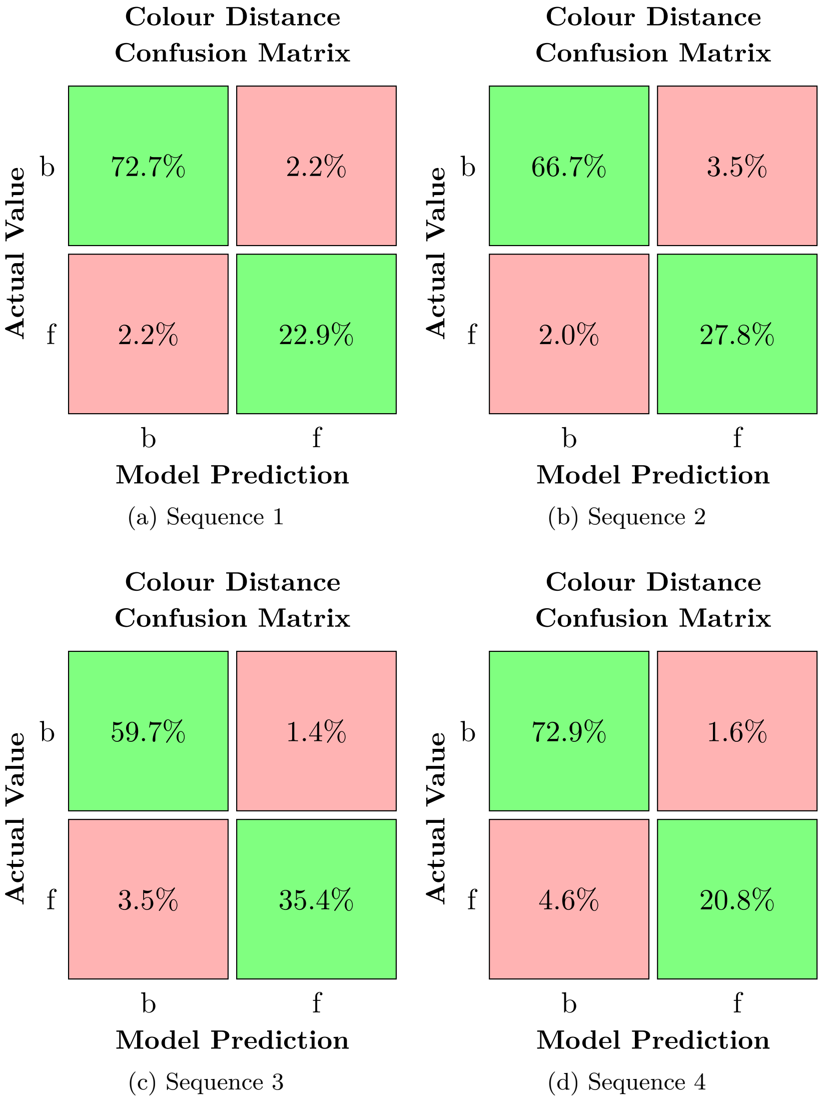
(c) Sequence 3

**Colour Distance Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 74.1% | 0.4% |
| f | 4.9% | 21.9% |

(d) Sequence 4

Figure A.2: Colour Distance Confusion Matrices



Colour Distance Confusion Matrix

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 72.7% | 2.2% |
| f | 2.2% | 22.9% |

(a) Sequence 1

Colour Distance Confusion Matrix

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 66.7% | 3.5% |
| f | 2.0% | 27.8% |

(b) Sequence 2

Colour Distance Confusion Matrix

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 59.7% | 1.4% |
| f | 3.5% | 35.4% |

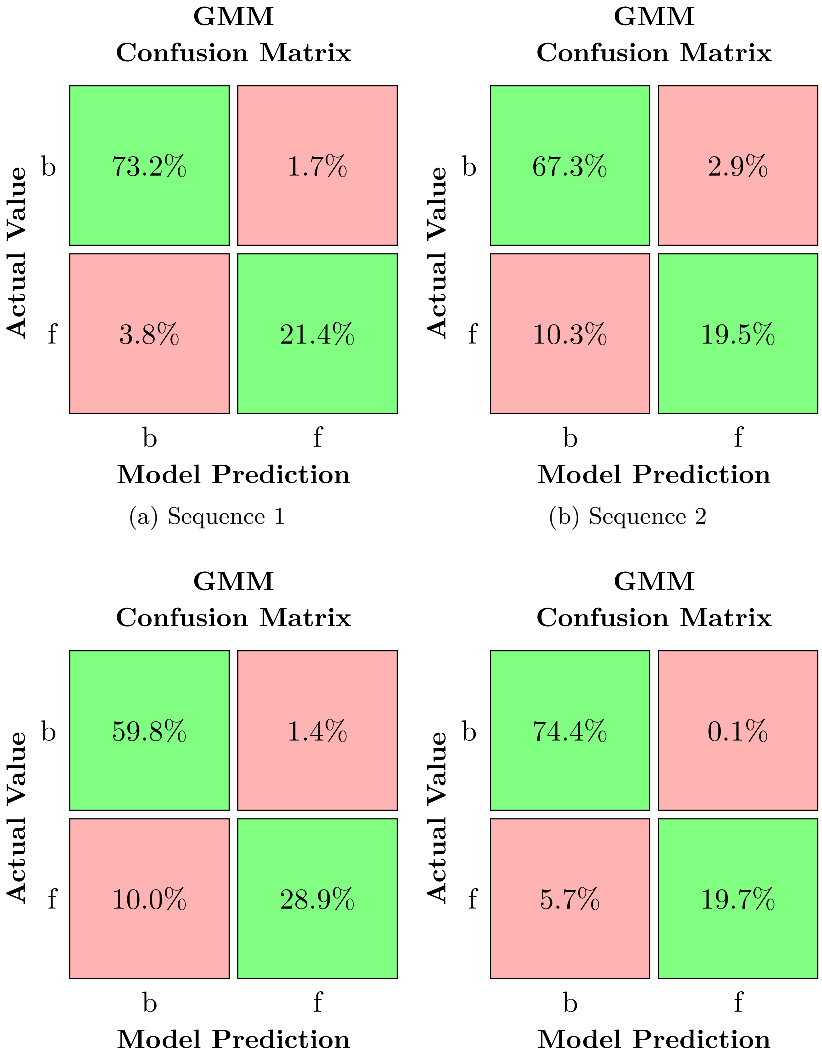
(c) Sequence 3

Colour Distance Confusion Matrix

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 72.9% | 1.6% |
| f | 4.6% | 20.8% |

(d) Sequence 4

Figure A.3: Colour Distance Confusion Matrices, Filtered



**GMM Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 73.2% | 1.7% |
| f | 3.8% | 21.4% |

(a) Sequence 1

**GMM Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 67.3% | 2.9% |
| f | 10.3% | 19.5% |

(b) Sequence 2

**GMM Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 59.8% | 1.4% |
| f | 10.0% | 28.9% |

(c) Sequence 3

**GMM Confusion Matrix**

| Actual Value \ Model Prediction | b | f |
|---|---|---|
| b | 74.4% | 0.1% |
| f | 5.7% | 19.7% |

(d) Sequence 4

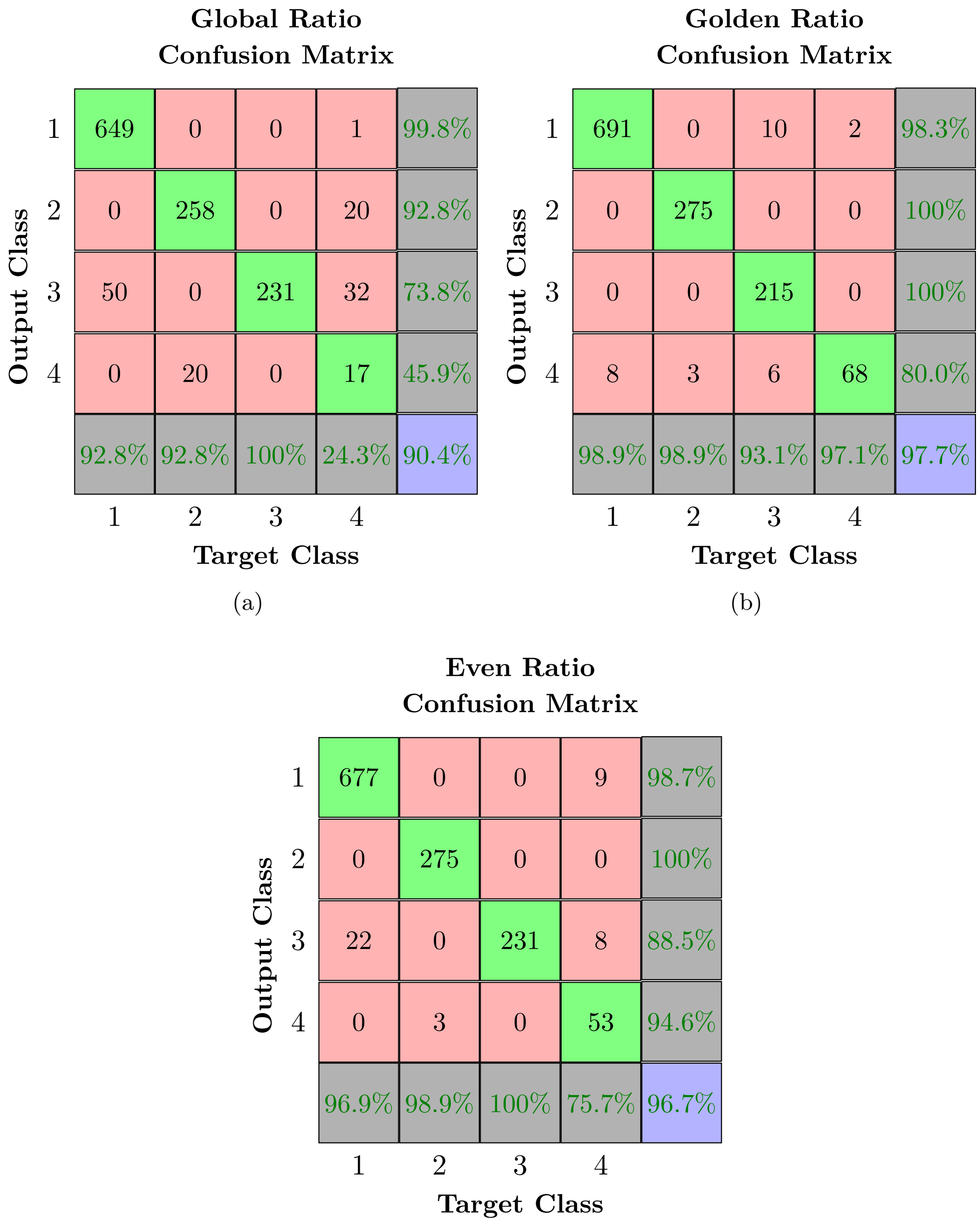
Figure A.4: GMM Confusion Matrices

# Appendix B

# Shot Classification Performance

| | **Type** |
|---|---|
| 1. | Out of Field |
| 2. | Long |
| 3. | Medium |
| 4. | Close |

Table B.1: Shot types



**Global Ratio Confusion Matrix**

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 649 | 0 | 0 | 1 | 99.8% |
| 2 | 0 | 258 | 0 | 20 | 92.8% |
| 3 | 50 | 0 | 231 | 32 | 73.8% |
| 4 | 0 | 20 | 0 | 17 | 45.9% |
| | 92.8% | 92.8% | 100% | 24.3% | 90.4% |

Target Class

(a)

**Golden Ratio Confusion Matrix**

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 691 | 0 | 10 | 2 | 98.3% |
| 2 | 0 | 275 | 0 | 0 | 100% |
| 3 | 0 | 0 | 215 | 0 | 100% |
| 4 | 8 | 3 | 6 | 68 | 80.0% |
| | 98.9% | 98.9% | 93.1% | 97.1% | 97.7% |

Target Class

(b)

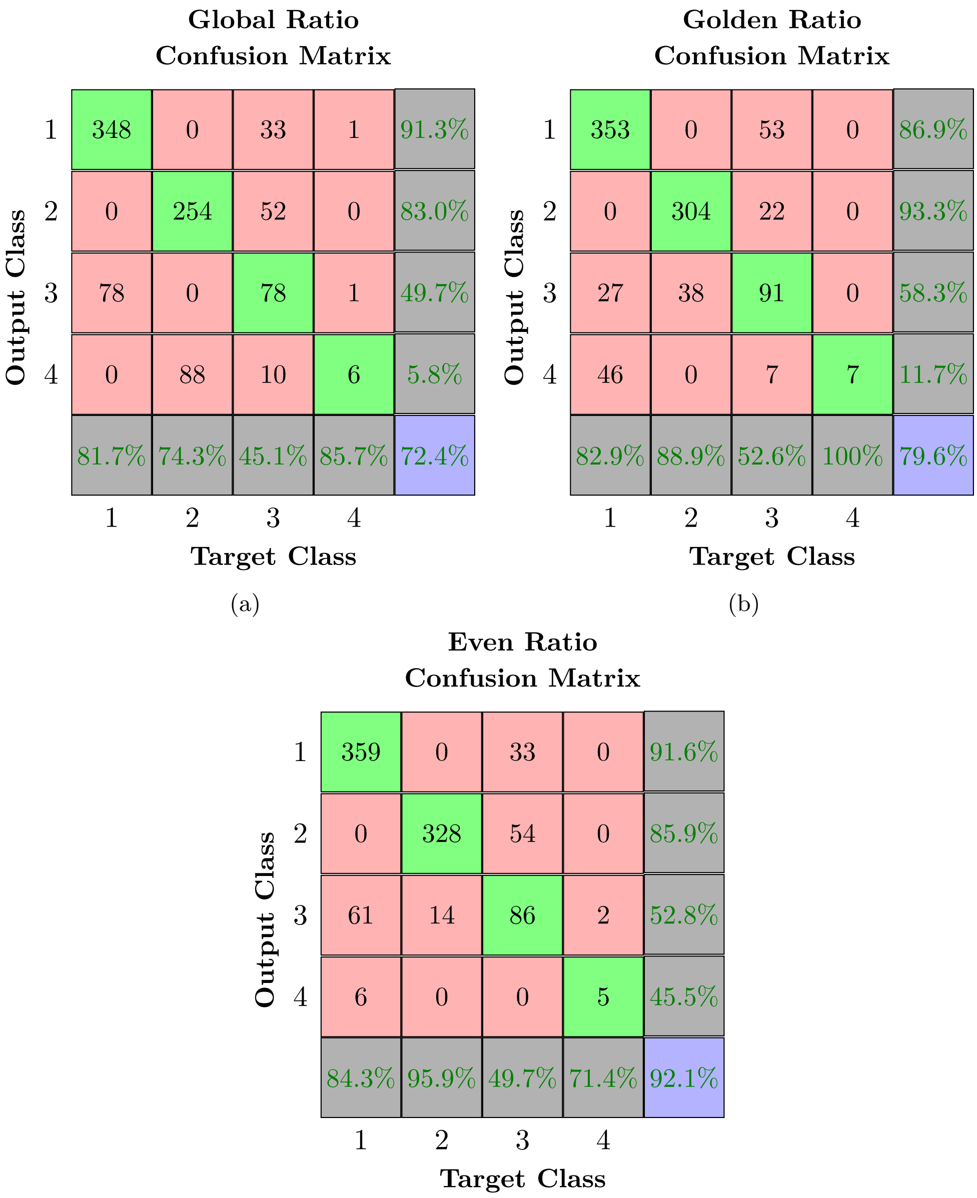
**Even Ratio Confusion Matrix**

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 677 | 0 | 0 | 9 | 98.7% |
| 2 | 0 | 275 | 0 | 0 | 100% |
| 3 | 22 | 0 | 231 | 8 | 88.5% |
| 4 | 0 | 3 | 0 | 53 | 94.6% |
| | 96.9% | 98.9% | 100% | 75.7% | 96.7% |

Target Class

(c)

Figure B.1: Field Ratio Confusion Matrices, Sequence 1



**Global Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 348 | 0 | 33 | 1 | 91.3% |
| 2 | 0 | 254 | 52 | 0 | 83.0% |
| 3 | 78 | 0 | 78 | 1 | 49.7% |
| 4 | 0 | 88 | 10 | 6 | 5.8% |
| | 81.7% | 74.3% | 45.1% | 85.7% | 72.4% |

(a)

**Golden Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 353 | 0 | 53 | 0 | 86.9% |
| 2 | 0 | 304 | 22 | 0 | 93.3% |
| 3 | 27 | 38 | 91 | 0 | 58.3% |
| 4 | 46 | 0 | 7 | 7 | 11.7% |
| | 82.9% | 88.9% | 52.6% | 100% | 79.6% |

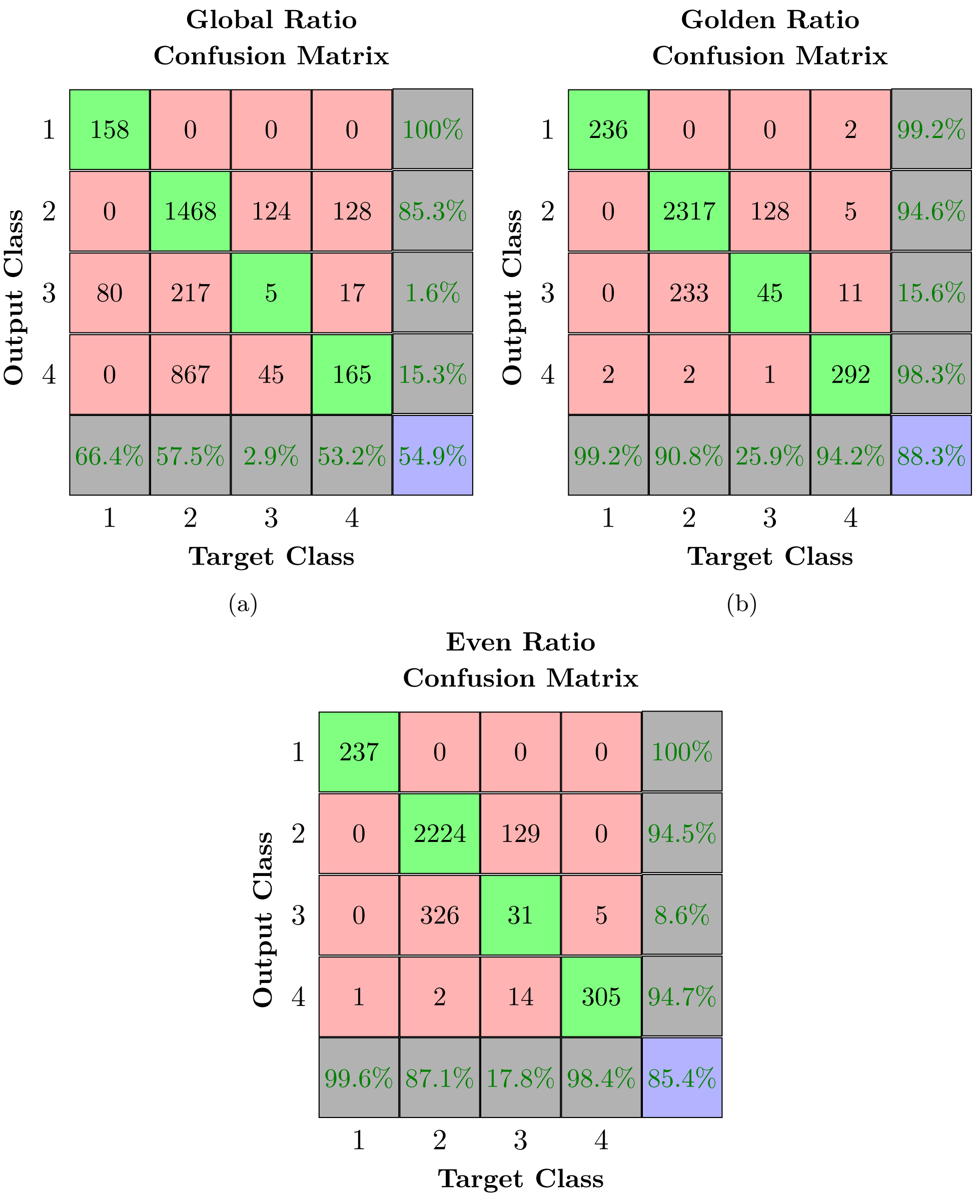
(b)

**Even Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 359 | 0 | 33 | 0 | 91.6% |
| 2 | 0 | 328 | 54 | 0 | 85.9% |
| 3 | 61 | 14 | 86 | 2 | 52.8% |
| 4 | 6 | 0 | 0 | 5 | 45.5% |
| | 84.3% | 95.9% | 49.7% | 71.4% | 92.1% |

(c)

Figure B.2: Field Ratio Confusion Matrices, Sequence 2



**Global Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 158 | 0 | 0 | 0 | 100% |
| 2 | 0 | 1468 | 124 | 128 | 85.3% |
| 3 | 80 | 217 | 5 | 17 | 1.6% |
| 4 | 0 | 867 | 45 | 165 | 15.3% |
| | 66.4% | 57.5% | 2.9% | 53.2% | 54.9% |

(a)

**Golden Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 236 | 0 | 0 | 2 | 99.2% |
| 2 | 0 | 2317 | 128 | 5 | 94.6% |
| 3 | 0 | 233 | 45 | 11 | 15.6% |
| 4 | 2 | 2 | 1 | 292 | 98.3% |
| | 99.2% | 90.8% | 25.9% | 94.2% | 88.3% |

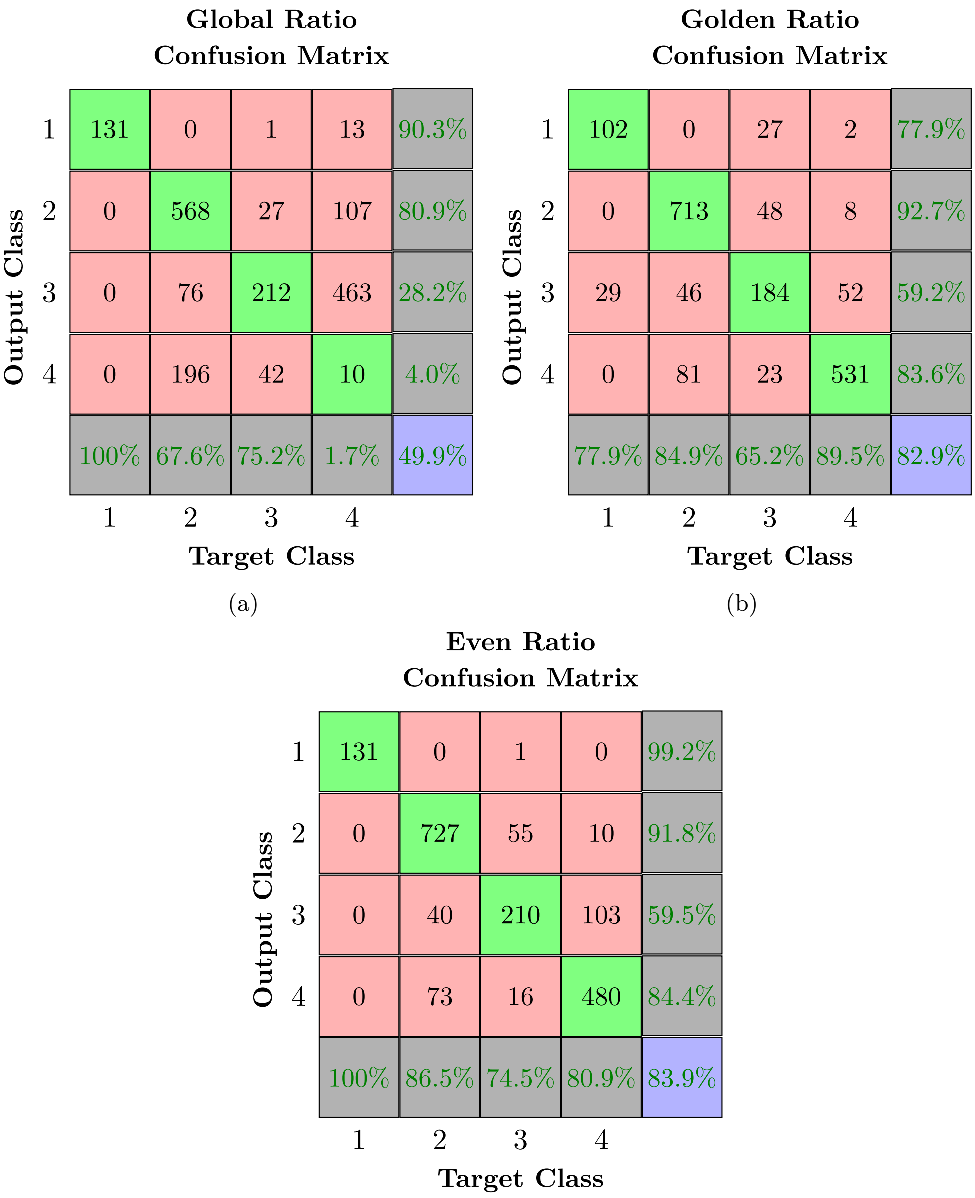
(b)

**Even Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 237 | 0 | 0 | 0 | 100% |
| 2 | 0 | 2224 | 129 | 0 | 94.5% |
| 3 | 0 | 326 | 31 | 5 | 8.6% |
| 4 | 1 | 2 | 14 | 305 | 94.7% |
| | 99.6% | 87.1% | 17.8% | 98.4% | 85.4% |

(c)

Figure B.3: Field Ratio Confusion Matrices, Sequence 3



**Global Ratio Confusion Matrix**

| Output Class | | | | | |
|---|---|---|---|---|---|
| 1 | 131 | 0 | 1 | 13 | 90.3% |
| 2 | 0 | 568 | 27 | 107 | 80.9% |
| 3 | 0 | 76 | 212 | 463 | 28.2% |
| 4 | 0 | 196 | 42 | 10 | 4.0% |
| | 100% | 67.6% | 75.2% | 1.7% | 49.9% |
| Target Class | 1 | 2 | 3 | 4 | |

(a)

**Golden Ratio Confusion Matrix**

| Output Class | | | | | |
|---|---|---|---|---|---|
| 1 | 102 | 0 | 27 | 2 | 77.9% |
| 2 | 0 | 713 | 48 | 8 | 92.7% |
| 3 | 29 | 46 | 184 | 52 | 59.2% |
| 4 | 0 | 81 | 23 | 531 | 83.6% |
| | 77.9% | 84.9% | 65.2% | 89.5% | 82.9% |
| Target Class | 1 | 2 | 3 | 4 | |

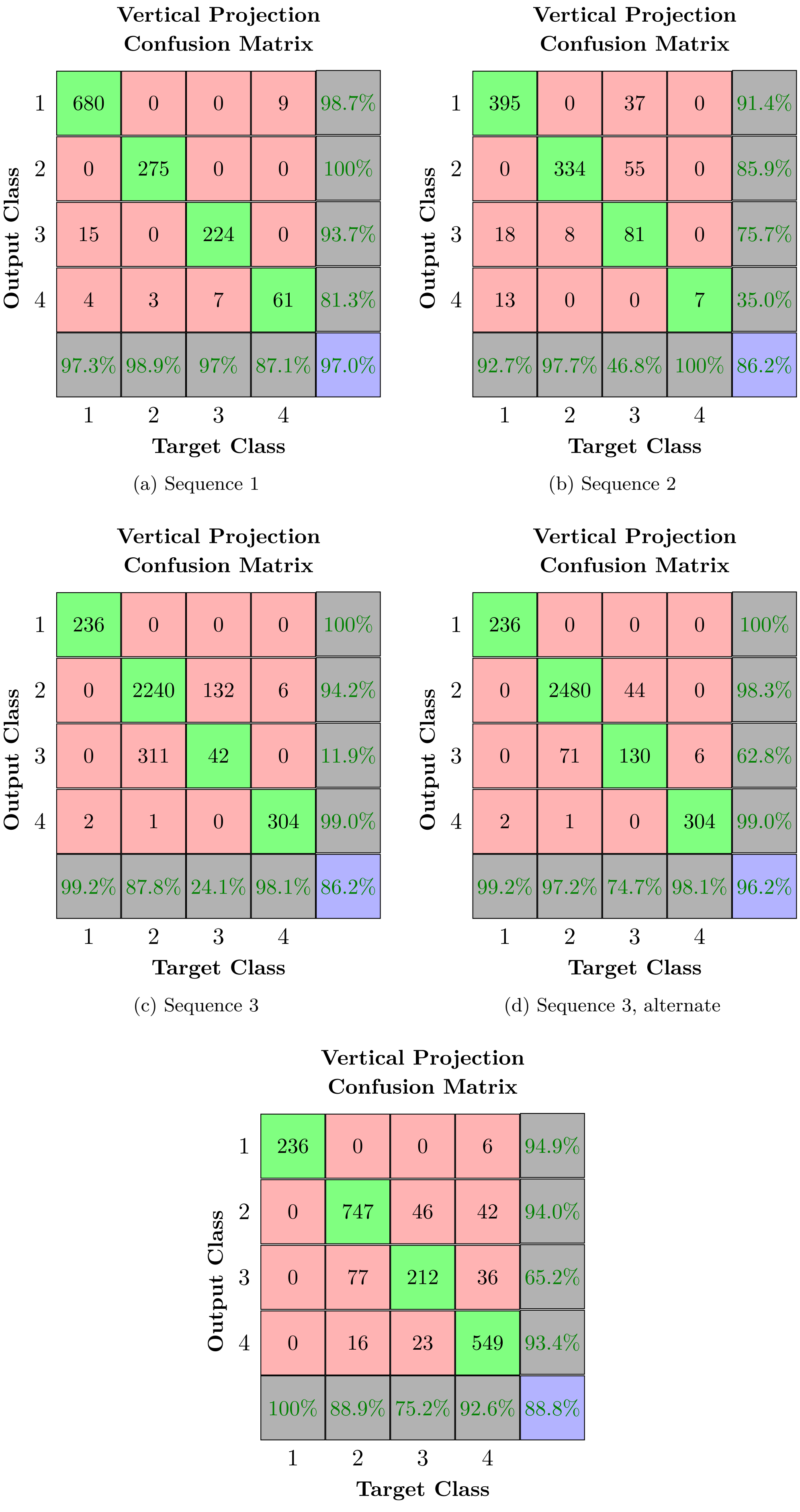
(b)

**Even Ratio Confusion Matrix**

| Output Class | | | | | |
|---|---|---|---|---|---|
| 1 | 131 | 0 | 1 | 0 | 99.2% |
| 2 | 0 | 727 | 55 | 10 | 91.8% |
| 3 | 0 | 40 | 210 | 103 | 59.5% |
| 4 | 0 | 73 | 16 | 480 | 84.4% |
| | 100% | 86.5% | 74.5% | 80.9% | 83.9% |
| Target Class | 1 | 2 | 3 | 4 | |

(c)

Figure B.4: Field Ratio Confusion Matrices, Sequence 4



Vertical Projection Confusion Matrix

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 680 | 0 | 0 | 9 | 98.7% |
| 2 | 0 | 275 | 0 | 0 | 100% |
| 3 | 15 | 0 | 224 | 0 | 93.7% |
| 4 | 4 | 3 | 7 | 61 | 81.3% |
| | 97.3% | 98.9% | 97% | 87.1% | 97.0% |

(a) Sequence 1

Vertical Projection Confusion Matrix

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 395 | 0 | 37 | 0 | 91.4% |
| 2 | 0 | 334 | 55 | 0 | 85.9% |
| 3 | 18 | 8 | 81 | 0 | 75.7% |
| 4 | 13 | 0 | 0 | 7 | 35.0% |
| | 92.7% | 97.7% | 46.8% | 100% | 86.2% |

(b) Sequence 2

Vertical Projection Confusion Matrix

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 236 | 0 | 0 | 0 | 100% |
| 2 | 0 | 2240 | 132 | 6 | 94.2% |
| 3 | 0 | 311 | 42 | 0 | 11.9% |
| 4 | 2 | 1 | 0 | 304 | 99.0% |
| | 99.2% | 87.8% | 24.1% | 98.1% | 86.2% |

(c) Sequence 3

Vertical Projection Confusion Matrix

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 236 | 0 | 0 | 0 | 100% |
| 2 | 0 | 2480 | 44 | 0 | 98.3% |
| 3 | 0 | 71 | 130 | 6 | 62.8% |
| 4 | 2 | 1 | 0 | 304 | 99.0% |
| | 99.2% | 97.2% | 74.7% | 98.1% | 96.2% |

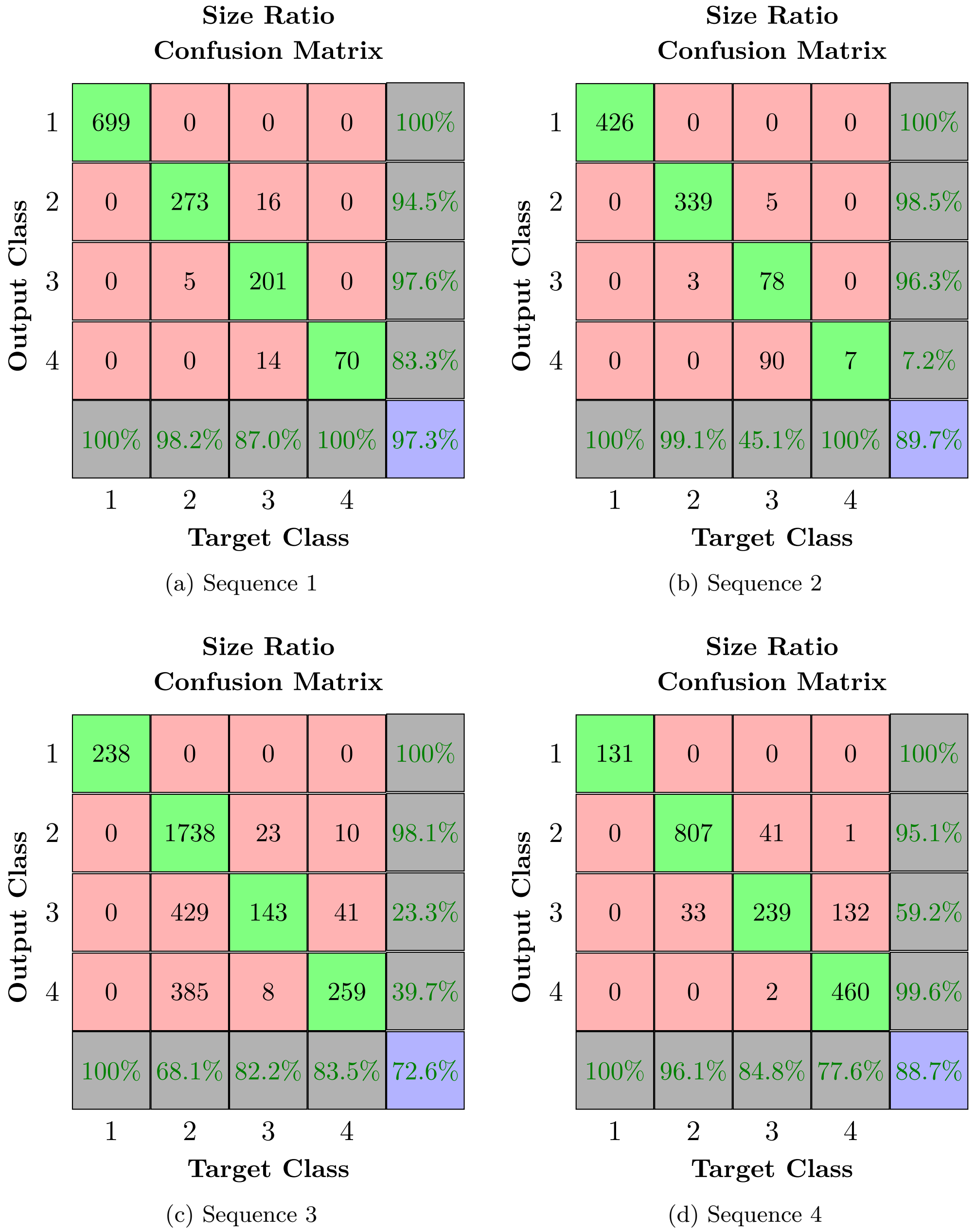
(d) Sequence 3, alternate

Vertical Projection Confusion Matrix

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 236 | 0 | 0 | 6 | 94.9% |
| 2 | 0 | 747 | 46 | 42 | 94.0% |
| 3 | 0 | 77 | 212 | 36 | 65.2% |
| 4 | 0 | 16 | 23 | 549 | 93.4% |
| | 100% | 88.9% | 75.2% | 92.6% | 88.8% |

(e) Sequence 4

Figure B.5: Vertical Projection Confusion Matrices



**Size Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 699 | 0 | 0 | 0 | 100% |
| 2 | 0 | 273 | 16 | 0 | 94.5% |
| 3 | 0 | 5 | 201 | 0 | 97.6% |
| 4 | 0 | 0 | 14 | 70 | 83.3% |
| | 100% | 98.2% | 87.0% | 100% | 97.3% |

(a) Sequence 1

**Size Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 426 | 0 | 0 | 0 | 100% |
| 2 | 0 | 339 | 5 | 0 | 98.5% |
| 3 | 0 | 3 | 78 | 0 | 96.3% |
| 4 | 0 | 0 | 90 | 7 | 7.2% |
| | 100% | 99.1% | 45.1% | 100% | 89.7% |

(b) Sequence 2

**Size Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 238 | 0 | 0 | 0 | 100% |
| 2 | 0 | 1738 | 23 | 10 | 98.1% |
| 3 | 0 | 429 | 143 | 41 | 23.3% |
| 4 | 0 | 385 | 8 | 259 | 39.7% |
| | 100% | 68.1% | 82.2% | 83.5% | 72.6% |

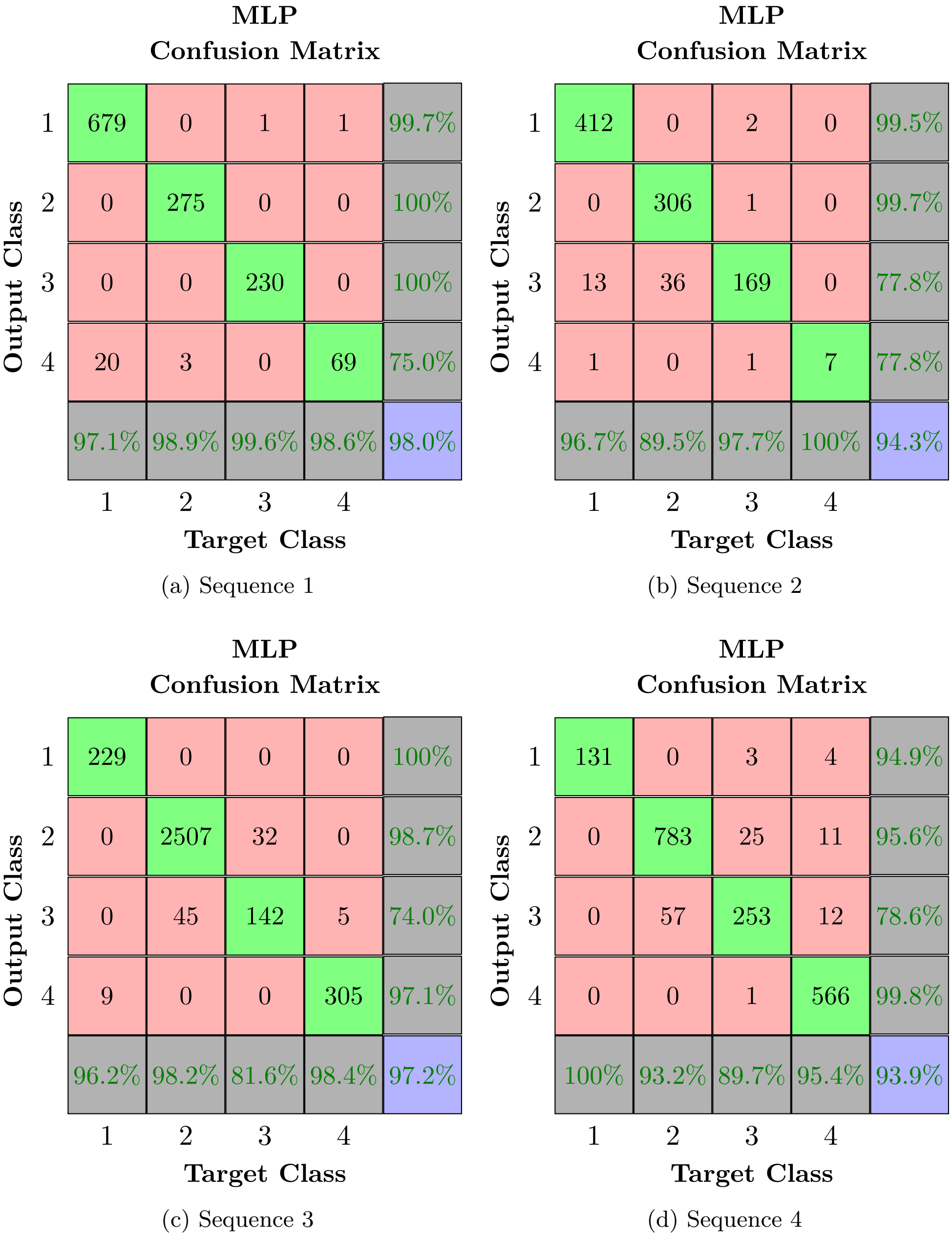
(c) Sequence 3

**Size Ratio Confusion Matrix**

| Output Class \ Target Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 131 | 0 | 0 | 0 | 100% |
| 2 | 0 | 807 | 41 | 1 | 95.1% |
| 3 | 0 | 33 | 239 | 132 | 59.2% |
| 4 | 0 | 0 | 2 | 460 | 99.6% |
| | 100% | 96.1% | 84.8% | 77.6% | 88.7% |

(d) Sequence 4

Figure B.6: Size Ratio Confusion Matrices



MLP Confusion Matrix

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 679 | 0 | 1 | 1 | 99.7% |
| 2 | 0 | 275 | 0 | 0 | 100% |
| 3 | 0 | 0 | 230 | 0 | 100% |
| 4 | 20 | 3 | 0 | 69 | 75.0% |
| | 97.1% | 98.9% | 99.6% | 98.6% | 98.0% |
| | 1 | 2 | 3 | 4 | |

Target Class

(a) Sequence 1

MLP Confusion Matrix

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 412 | 0 | 2 | 0 | 99.5% |
| 2 | 0 | 306 | 1 | 0 | 99.7% |
| 3 | 13 | 36 | 169 | 0 | 77.8% |
| 4 | 1 | 0 | 1 | 7 | 77.8% |
| | 96.7% | 89.5% | 97.7% | 100% | 94.3% |
| | 1 | 2 | 3 | 4 | |

Target Class

(b) Sequence 2

MLP Confusion Matrix

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 229 | 0 | 0 | 0 | 100% |
| 2 | 0 | 2507 | 32 | 0 | 98.7% |
| 3 | 0 | 45 | 142 | 5 | 74.0% |
| 4 | 9 | 0 | 0 | 305 | 97.1% |
| | 96.2% | 98.2% | 81.6% | 98.4% | 97.2% |
| | 1 | 2 | 3 | 4 | |

Target Class

(c) Sequence 3

MLP Confusion Matrix

| Output Class | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 131 | 0 | 3 | 4 | 94.9% |
| 2 | 0 | 783 | 25 | 11 | 95.6% |
| 3 | 0 | 57 | 253 | 12 | 78.6% |
| 4 | 0 | 0 | 1 | 566 | 99.8% |
| | 100% | 93.2% | 89.7% | 95.4% | 93.9% |
| | 1 | 2 | 3 | 4 | |

Target Class

(d) Sequence 4

Figure B.7: MLP Confusion Matrices