Tracking People Across Disjoint Camera Views

A Thesis Submitted For the Degree of Doctorate of Philosophy

By

Christopher Madden

Faculty of Information Technology
University of Technology, Sydney
Australia

July 3, 2009

I, **Christopher Madden**, certify that the work in this thesis titled "**Tracking People Across Disjoint Camera Views**" has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text. I also certify that the thesis has been written by myself. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. The undersigned certify that they have read this thesis and that in their opinions it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

Production Note:
Signature removed prior to publication.

Christopher Madden

Date: July 3, 2009

Principal Supervisor: Massimo Piccardi

Acknowledgements

I would like to thank all those people who have been involved in some way with the creation of this thesis. My biggest thanks goes to Professor Massimo Piccardi for being my supervisor and assisting me with learning all of the areas I have needed to know for this work. And for filling me in on the areas that I still haven't mastered yet. To my loving wife and family thank you for all the caring support. I would like to promise that I will talk a little less about video surveillance from now on; however we both know that is unlikely to happen. I would also like to thank my friends and colleagues who have assisted me throughout my candidature. Especially those who have read, or even reread, this thesis and helped me to improve it to its current state. I would also like to thank you, the reader, for your time as this thesis is not only written to explore my knowledge of the area, but hopefully to pass some of that on to you.

Abstract

Tracking people around surveillance systems is becoming increasingly important in the current security conscious environment. This thesis presents a framework to automatically track the movements of individual people in large video camera networks, even where there are gaps between camera views. It is designed to assist security operators, or police investigations by providing additional information about the location of individuals throughout the surveillance area. Footage from an existing surveillance system has been used to test the framework under real conditions. The framework uses the similarity of robust shape and appearance features to match tracks. These features are extracted to build an object feature model as people move within a single camera view, which can be compared across cameras. The integration of matching similarities in the temporal domain increases the robustness to errors of many kinds. Frames with significant segmentation errors can be automatically detected and removed based upon their lack of similarity to the other models within the same track, increasing robustness.

The shape and appearance features used to generate the object models are based upon features humans habitually use for identifying individuals. They include a height estimate, a Major Colour Representation (MCR) of the individuals global colours, and estimates of the colours of the upper and lower portions of clothing. The fusion of these features is shown to be complementary, providing increased discrimination between individuals. The MCR colour features are improved through the mitigation of illumination changes using controlled equalisation, which improves the accuracy in matching colour under normal surveillance conditions and requires no training or scene knowledge. The incorporation of other features into this framework is also relatively straightforward.

This track matching framework was tested upon four individuals across two video cameras of an existing surveillance system. Existing infrastructure and actors were used to ensure that ground truth is available. Specific cases were constructed to test the limitations of the system when similar clothing is worn. In the data, the height difference ranges from 5 to 30 centimetres, and individuals may only be wearing 50% of similar clothing colours. The accuracy of matching an individual was as high as 91% with only 5% false alarms when all the system components were used. This may not become a fully automated system, but could be used in semi-automated or human assisted systems, or as the basis for further research into improved automated surveillance. Application areas range from forensic surveillance to the matching of the movements of key individuals throughout a surveillance network and possibly even target location.

Contents

1 Introduction **1**
- 1.1 Aim . . . 4
- 1.2 Scope . . . 5
- 1.3 Contribution . . . 6
  - 1.3.1 Publications . . . 7
- 1.4 Thesis Overview . . . 8

2 Literature Review **10**
- 2.1 Motion Detection and Object Segmentation Techniques . . . 10
  - 2.1.1 Background Modelling . . . 12
  - 2.1.2 Removal of Shadow and other Segmentation Noise . . . 16
- 2.2 Colour Space Research . . . 18
- 2.3 Statistical Similarity Measurements . . . 25
- 2.4 Object Tracking . . . 27
  - 2.4.1 Complexity Factors within Object Tracking . . . 30
  - 2.4.2 Feature-Based Tracking . . . 31
  - 2.4.3 Model-Based Tracking . . . 32
  - 2.4.4 Mean Shift-based Tracking . . . 35
  - 2.4.5 Summary of Tracking Literature . . . 36
- 2.5 Object Classification . . . 37
- 2.6 Current Disjoint Camera Tracking Methods . . . 38
- 2.7 Literature Summary . . . 47

3 Colour-based Robust Appearance Features **49**
- 3.1 Appearance Feature Background . . . 50
- 3.2 MCR Colour Feature Extraction . . . 53
  - 3.2.1 Optimising MCR Using an Online k-means Algorithm . . . 56
- 3.3 Improving Robustness Using Incremental MCRs . . . 58
- 3.4 Comparing MCR or IMCR Appearance Features . . . 60
  - 3.4.1 Time Integration of Similarity . . . 63
- 3.5 Extracting Spatial MCR Colour Features . . . 65
- 3.6 Experimental Validation of MCR Appearance Features . . . 68
  - 3.6.1 Colour Experiments on Manually Segmented Individuals . . . 69
  - 3.6.2 Colour Experiments on Automatically Obtained Tracks . . . 72
- 3.7 Discussion of MCR Appearance Results . . . 77
- 3.8 Summary of MCR Appearance Features and Future Enhancements 79

4 Mitigating the Effects of Changes in Illumination **82**
4.1 Illumination Mitigation Background 83
4.2 Illumination Filtration 87
4.3 Histogram Stretching 88
4.4 Histogram Equalisation 90
4.5 Comparing Illumination Mitigation Techniques 93
4.6 Experimental Comparison of Mitigation Techniques 96
4.7 Discussion of Illumination Mitigation 100
4.8 Summary of Illumination Mitigation and Future Enhancements . . 101

5 Identification of Segmentation Errors **104**
5.1 Segmentation Error Identification Background 105
5.2 Identifying Segmentation Errors Through Changes in Bounding Box Height 106
5.3 Identifying Segmentation Errors Through Appearance Feature Analysis 107
5.3.1 Comparing Colour Features Between Frames 111
5.3.2 Typical MCR Patterns of Major Segmentation Errors . . . 112
5.4 Experimental Validation for the Identification of Major Segmentation Errors 114
5.5 Discussion of Segmentation Error Identification 116
5.6 Summary of Segmentation Error Identification and Future Enhancements 117

6 Height Based Robust Shape Feature **119**
6.1 Shape Feature Background 120
6.2 Obtaining Height Estimates Using Camera Calibration 124
6.3 Improved Automatic Monocular Height Estimates 128
6.4 Statistically Comparing Height Features 132
6.5 Experimental Verification of Height Estimation 134
6.5.1 Height Experiments Comparing Manual and Automatic Height Estimates 134
6.5.2 Height Experiments Using a Larger Dataset 137
6.6 Discussion of Height Results 139
6.7 Summary of Height Feature and Future Enhancements 140

7 Fusion Methods and Results for Combining Robust Features 142
7.1 Classifier-based Fusion Background 143
7.2 Classifier-based Fusion for Integrating Features across Differing Time Scales 146
7.3 Results From Fused Features 147
7.3.1 Evaluation of the Statistical Models 148
7.3.2 Evaluation of Fusing Features 150
7.4 Summary of Fused Features 153

8 Conclusions 156

List of Figures

1 Approaches incorporating spatial colour information 52
2 Major Colour Representation of 'tn_flower' 56
3 Original ore gold rose image (left) and reprojection of the 90% most frequent pixel clusters after 7 k-means iterations (right) . . . 57
4 MCR changes for 20 most significant colours with iterations of the k-means optimisation 57
5 MCR from three automatically detected people 59
6 IMCR matching of two tracks using time integration 64
7 Examples of upper and lower regions of segmented individuals . . 67
8 Same individuals observed in camera 3a and camera 5 69
9 Differing individuals observed in camera 3a and camera 5 71
10 Typical frames used for test cases 72
11 Typical backgrounds used for test cases 73
12 Four people of interest (Person's A, B, C, D from left) and good automatically segmented masks (from frames 775, 1095, 1542, 2044) 74
13 Poor segmentation in two sample cluttered frames 76
14 Accuracy of individual colour features 77
15 Sample people of interest and their red histograms under differing illumination conditions 83
16 Individuals R values before and after illumination filtration 88
17 Histogram stretching of the individual's pixels 90
18 Full equalisation of the individual's pixels 91

19 Controlled equalisation of the individual's pixels with varying k values 92
20 Intersection of the H0 and H1 curves of the height feature 95
21 Example of changes in bounding box height ratios where a large segmentation errors occurs and 5 sample frames from the track including the erroneous frame 108
22 Example of upper and lower regions from three segmentations of one person 110
23 Example of upper and lower regions from three segmentations of a second person 110
24 Three typical error patterns of frame based pairwise similarity comparisons given between 0 and 1 113
25 Four people of interest and automatically segmented masks of good quality 114
26 ROC curves of the height, colour and fused feature results 115
27 Accurate location of the bottom point improves height estimates . 125
28 Manually identified key image points for the track of one individual 127
29 Curvature values for a single individual showing curvature key points beginning at the top left most object pixel 130
30 Height estimates for 5 tracks demonstrating the manual height estimates 132
31 An individual's key points and height estimates using automatic and manual techniques on the poorly segmented track 13 135
32 ROC curves of the height feature results 138
33 Example $P(s_{UC}|H0)$ and $P(s_{UC}|H1)$ 145
34 Four people of interest 148
35 ROC curves of the height, MCR and fused feature results 151
36 ROC curves for fusing spatial colour MCR's, fusing all the colour MCR's, and fusing all the features 152
37 Pictorial storyboard summary of potentially matching tracks . . . 155

List of Tables

1 Results of IMCR Matching - same person 70
2 Results of IMCR Matching - differing people 72
3 Results of IMCR matching - differing people 73
4 Results of automated IMCR matching - 6 different cases 75

5 Global similarity measurements for matching and non-matching tracks 97
6 Upper MCR similarity for matching and non-matching tracks . . . 98
7 Lower MCR similarity for matching and non-matching tracks . . . 100
8 PD and PFA values of Bounding Box and MCR features for detecting segmentation errors 115
9 Auto height estimates of a 1710 mm individual over 15 tracks . . 136
10 Ground Truth of Participants 148
11 How variations to the optimum threshold affect% error rates . . . 149

1 Introduction

Computer vision-based object tracking is a difficult task that is mainly based upon shape, motion, and appearance features [47]. Motion features have tended to be widely utilised in human environments, such as within buildings, because of the previously limited camera resolution to exploit shape or appearance features effectively. Typical building or campus surveillance systems are created to assist human operators to view key locations around the surveillance area, whilst also considering the cost effectiveness of the security system. Thus they tend to consist of a relatively small number of cameras, sometimes of varying quality and differing camera properties. Such changing camera properties include colour saturation levels and shutter speed control, which may vary significantly across the cameras that are sparsely located around the surveillance area. Coverage of key security locations is sometimes improved through the installation of overlapping or near-overlapping cameras; however cost considerations make this uncommon throughout the wider surveillance system. The video data acquired from cameras throughout the surveillance system may also occur at different resolutions and frame rates. The resolutions are often low to minimise the data transfer, whilst still providing enough information change for a human operator to utilise the system. Further minimising the size of the data through image compression may also lead to considerable compression artefacts that can be problematic for automatic analysis. This creates a number of difficulties for automated computer vision using the same surveillance system as large gaps in coverage lead to unreliable motion cues across portions of the surveillance space, illumination changes between cameras or within a single camera over time, and internal camera parameters differ throughout the system.

Recent advances in affordable camera technology now provide increased camera resolution and quality. Although it can not cost-effectively satisfy the coverage problem for automated systems, it does provide improved image quality. This leads to improved information about the object, including shape and appearance features. The cost of upgrading the whole surveillance infrastructure can be considerable, leading to cameras of low resolution and quality often being used. Motion features are still extremely useful in local single camera views, or groups of overlapping or near overlapping camera view, where they are still the main tracking feature used [125]; however they are not as useful for tracking, or matching the tracks of individuals across regions where there is unreliable object movement information. Such regions occur where there are large gaps in coverage and the movements of humans may not reflect average motion. Indeed the

most important tracks may occur where an individual differs significantly from the average pattern. Improved shape and appearance features provide promise for an enhanced ability of automated security systems to mitigate some of the current problems, and allow reasonable accuracy of matching individuals in different cameras. Where the tracks of an individual are matched, they can be combined to locate where individuals have been viewed over time, which is effectively tracking individuals over the surveillance area. Articulated human movement adds extra difficulty as a wide range of shape changes can occur within the normal range of human movement, even with assumptions such as people generally walk upright. Even though there is a limited range of expected natural and artificial illumination conditions, these changes can still cause significant appearance changes. Effectively utilising these features promises to reduce the time and human effort that might be required to generate a set of tracks relating to how key individuals, or possibly even all individuals, might have moved around within the surveillance system.

This thesis presents research into a framework for the fusion of information across a multi camera surveillance system to analyse the movements of individual people using robust shape and appearance features. The presented work is based upon the definition of the surveillance session as 'a portion of one day where people enter the surveillance area from a known set of entry points to perform their activities before leaving through known exit points'. This definition leads to the following simplifying assumptions about the surveillance area, and the people viewed within that area:

1. All entry and exit points of the surveillance area are in view of a surveillance camera.

2. Individuals are unlikely to change their clothing or footwear; hence, many of their intrinsic shape and appearance features will remain relatively constant for the duration of the surveillance session.

3. Individuals are tracked reasonably accurately for a reasonable duration whilst within the view of any of the system's cameras, such that the correct object regions are associated to a track.

4. Individuals are segmented from the background into a single blob, or singly label group, but not necessarily accurately.

5. Individuals are often observed at a distance from the camera, so biometric features such as faces may not be generally available.

6. Where cameras are significantly disjoint, motion features may vary unpredictably between those cameras as individuals are allowed free motion.

7. Illumination varies significantly, but within the limited range typical of natural or artificial lighting for public premises.

The above assumptions suggest that tracking is of a reasonable quality and duration. These assumptions hold for current tracking software on video where traffic is sparse. Some fragmentation of track can occur when conditions become difficult or the level of traffic becomes high; however such cases can be treated as extra tracks to be merged within the framework. They also suggest shape and appearance features should be used rather than motion features to generate accurate matching of information about object tracks within any surveillance system where camera views do not overlap. Although motion features maybe reliable for some portions of the system [123], these are only considered as a possible feature and are not fully explored for enhancing the systems results. Unfortunately, due to the articulated motion of people, few shape features other than height or gait are likely to remain stable during walking. Most appearance features are likely to remain stable within the extent of a surveillance session, although they may also be affected by articulation. For these reasons, the track matching framework is based upon extraction, comparison and fusion of upper clothing, lower clothing, and global colour appearance as well as height feature information without the use of motion models or expected transitions between camera views. Whilst exceptional cases may be easily constructed for this feature set to fail, it is designed to provide sufficient discrimination (at the ground truth level) for a large majority of real cases. Even where many people may be of similar appearance, it can reduce the amount of manual footage revision that would be required for human operators to perform the task alone.

The features identified can be made more robust by analysing for large changes in features along the track obtained from a single camera view. Where large changes are automatically detected from individual or small groups of frames, they can be removed to reduce the impact of these feature errors and mitigate their propagation into the surveillance system. This uses a small training set to determine the statistical likelihood of matching and non-matching cases based upon the features similarity values, which are fused using Bayes theorem. Where frames of a track are determined to be non-matching they can be discarded as they are likely to have significant errors. A similar Bayesian classification process into matching and non-matching classes is also used to determine the tracks which are from a single individual. Other key features, such as facial information, or camera

transitions can be incorporated within this framework where they are available or reliable. This could be achieved by extending the fusion framework to fuse the additional feature or features; however the achieved accuracy without these features is still adequate for a semi-automated system. It should also be noted that for a human to search through and observe comparisons of every track is very time consuming, but it much simpler to review potential matches that are incorrect and separate them for rematching correctly. This should be taken into consideration with a semi-automated version to optimise the system.

1.1 Aim

This thesis aims to address a shortage of work investigating tracking the movements of individuals across real surveillance systems. Such systems cannot achieve high accuracy by simply relying upon motion cues for tracking between cameras with large blind areas between camera views. Therefore this thesis aims to explore the usage of shape and appearance features that have recently become available with increased camera resolution. The specific aims of this thesis are investigating the research areas of:

1. Extracting invariant or quasi-invariant appearance features that are tolerant to illumination changes - This explores extracting a variety of appearance features focusing primarily upon colours and spatial colour components. Methods that can be used to mitigate illumination changes are also investigated in order to more accurately evaluate their stability both between and within camera views, as well as their accuracy, discriminative ability, and how complementary they are to other possible features.

2. Extracting invariant or quasi-invariant shape features - This explores extracting a variety of possible shape features, including height estimation, to evaluate their usefulness for matching the tracks of individuals in terms of stability, accuracy, discriminative ability, and complementary effect to other features.

3. Identifying frames within tracks which have significant errors - This investigates the most effective methods that can be used to identify frames which carry significant segmentation errors, so that they can be removed from the feature extraction process to create more robust feature sets.

4. Integrate the results of the frame level features along the temporal axis - This aims to compensate for minor changes that may occur due to small

segmentation errors, and increases the robustness of individual feature measurements.

5. Fusing features to improve the track matching accuracy - This investigates some of the possible feature fusion processes for their suitability in a track matching system. This evaluates the computational complexity, expandability and accuracy of the system, as well as the training required.

1.2 Scope

The scope of this thesis has been carefully chosen to primarily explore those areas required for matching the tracks of individuals across cameras. This area has not yet received much attention, and uses unrealistic assumptions about the individuals or the environment. Much of the research conducted for this thesis is based upon a number of underlying technologies which are described in detail throughout Section 2. These technologies, which include object segmentation and tracking within a single camera, among others, have developed to a stage where they provide reasonably reliable results upon which advanced methods can be developed. The scope of this thesis therefore assumes the reasonable reliability of many of these reasonably accurate underlying technologies, primarily in the areas of object segmentation and tracking. Areas such as segmentation can be a source of error in extracted features; however rather than redeveloping them, this thesis investigates techniques to identify and mitigate these errors and generate more robust object features. Illumination changes are also a significant error source; hence this thesis investigates the improved application of fast colour invariance techniques to mitigate its effect.

In addition to utilising particular features that have already been used successfully, this work also looks to limit the usage of features and assumptions that use general case statistics to improve their results. Although such statistical features can improve the overall accuracy of a tracking system, they tend to dramatically increase the error rate for anomalous cases. This occurs because the usage of priors favours outcomes from the class or case that is more frequent, rather than relying fully upon the feature information. This can become a problem because within a surveillance scenario the anomalous cases are often the most important, as people who commit offences are likely to exhibit anomalous behaviour patterns. In order to reduce the possibility of errors with anomalous cases, this work has been designed to stay 'prior-neutral'. Thus it does not utilise popular statistics such as path transitions or inter-camera walking time [123] in order to increase

the overall system results. A second motive is reducing the need to generate and maintain a detailed map of the specific camera locations and transitions, which are yet to be proven for large surveillance installations with non-overlapping camera views.

The methods used to fuse the results of the similarity of features have also been explored to investigate both the lowest level of errors and the ease with which other features can be included. The investigation of the results also looks at the limitations of the proposed methods and how these are affected by the limitations of reusing existing surveillance systems or the installation of new systems. One important aspect that has not been fully explored, and therefore is mentioned minimally is the real-time requirements of surveillance systems. Full implementation and optimisation has been outside the scope of this thesis due to time constraints. Hence the system is being investigated for forensic or after the event analysis, where real-time implementation is not as crucial.

1.3 Contribution

There are many contributions of this thesis to the wider research community. These are generated specifically for the disjoint camera tracking problem, but could also be used to improve results in other areas. These are the development and automatic application of:

1. A non-parametric flexible clustered colour representation, Major Colour Representation (MCR). This can be applied to provide accurate colour information about individual objects in a single frame region within a compact notation that retains the three dimensional colour information. This thesis currently focuses upon individual humans, however this method should be suitable for general objects from single images or image sequences. A k-means process is used to improve the accuracy of the MCR clusters to better represent the pixels which are associated to them.

2. Averaging the MCR across a small window of frames corresponding to at least half the gait period, creates IMCR features. This window reduces the impact of articulated shape changes upon the appearance features.

3. Investigation of the usage of spatial colour features to represent object appearance in a spatially discriminative way. This led to the development of spatial MCR and IMCR features, which can be used to extract spatial information for matching individuals.

4. Development of a symmetric similarity measurement for comparing MCR features that is complementary to the Kolmogorov divergence [70].

5. Development of a 'controlled equalisation' technique than can be used to mitigate illumination effects upon the colour histogram of an object. This method makes an object's colour information more matchable across a variety of illumination conditions, whilst still retaining the general profile of an object's colour histogram. It is also object dependent and does not require either training, or prior scene knowledge.

6. Time-integrated matching of MCR or IMCR features along the track sequence of any two objects. This utilises a Bayesian fusion technique where features to be integrated are dependent. This technique improves the robustness of the matching results, especially where sections of the track may have minor segmentation errors, or appearance changes according to the pose.

7. Identification of frames with major segmentation errors through an analysis of the changes in features, especially using MCR based appearance features.

8. Improved height estimation for a moving human from a single calibrated camera view. this is achieved through improved location of the feet and the top of the head in a single image.

9. Development of a statistical similarity measurement for improving comparisons of frame based height estimates obtained from two tracks.

10. Investigation of a framework for the Bayesian fusion of multiple features on a track level to achieve the maximum accuracy with a limited amount of training. These features may be either dependent, where a weighted sum is appropriate, or independent, where the product rule is most appropriate. This framework also includes a bias term that allows for the selection of the optimum operating point to either minimise the total error, or minimise a function. Such a cost based approach is important in a system where the cost of false positives differs from the cost of false alarms.

1.3.1 Publications

This work has been well received by the international community, as can be seen by the number of accepted publications. These have been accepted into a variety of conferences and journals:

- Christopher Madden, Eric Dahai Cheng, Massimo Piccardi, "Tracking People across Disjoint Camera Views by an Illumination-Tolerant Appearance Representation" *Machine Vision and Applications Journal*, Vol. 18, pp 234-778, 2007.

- C. Madden and M. Piccardi, "Comparison of Techniques for Mitigating Illumination Changes on Human Objects in Video Surveillance", *In Proceedings of the International Symposium on Visual Computing* (ISVC), 2007.

- C. Madden and M. Piccardi, "A Framework for Track Matching Across Disjoint Cameras using Robust Shape and Appearance Features", *In Proceedings of the International Conference on Advanced Video and Signal Based Surveillance* (AVSS), 2007.

- C. Madden and M. Piccardi, "Detecting Major Segmentation Errors for a Tracked Person Using Colour Feature Analysis", *In Proceedings of the International Conference on Image Analysis and Processing* (ICIAP), 2007.

- E. Cheng, C. Madden, M. Piccardi, "Mitigating the Effects of Variable Illumination for Tracking Across Disjoint Camera Views", *In Proceedings of the International Conference on Advanced Video and Signal Based Surveillance* (AVSS), pp 32-38, 2006.

- C. Madden and M. Piccardi, "Height Measurement as a Session-based Biometric for People Matching Across Disjoint Camera Views", *In Proceedings of the Image and Vision Computing New Zealand* (IVCNZ), pp. 282-286, 2005.

1.4 Thesis Overview

This thesis is organised into eight chapters, each focussing upon a component of extracting and using robust object features for tracking through track matching across a wide-area surveillance system. The chapters are organised to logically follow the progression of the system. Chapter 2 is the literature review, which is presented in six sections. The underlying technologies are described through the first five sections, as these form the basis of the ability to build and analyse the potential features of a tracked individual. The next section of the literature review looks more specifically, and in more depth, at the literature on the techniques used in applications proposed for tracking across disjoint camera systems.

A final summary is provided for clarity. Chapter 3 explores how to extract the appearance features of an individual, including the incorporation of spatially based upper and lower clothing colours, as well as global colours. It also investigates how to make these features more robust to pose changes through integration over time. Chapter 4 then looks at the mitigation of illumination effects upon the appearance of an observed individual including an analysis of a novel 'controlled equalisation' technique. Chapter 5 describes the methods developed to identify segmentation errors that might affect feature extraction, so that more robust features can be utilised. Chapter 6 investigates the use of height as a stable shape feature, and how an improved height estimate can be extracted from calibrated monocular cameras. It also investigates how the height estimate can be analysed along a track sequence to make it both more robust, and reduce the impact of gait effects. Chapter 7 then investigates the different methods to fuse the results of the robust features in order to distinguish between two individuals. This fusion also explores the fusion of features in the temporal domain so that the features can be fused and compared at the same level of time integration, whether that is frame level, across a window of frames, or along the entire track. This provides many results that are used to compare different fusion methods based upon the results of the individual features. Finally Chapter 8 concludes the thesis by summarising the results of the project including future research that would be required to fully design and implement a working system to effectively match two individuals based upon their tracks in any of the camera views throughout the system.

2 Literature Review

This literature review is separated into six distinct sections, with a final section summarising the literature relating to tracking or track matching across disjoint cameras. The first five sections explore components of literature that relate to the underlying technologies that are important to surveillance. The broad categories summarised in these sections are motion detection and object segmentation, colour representation techniques, statistical similarity measurements, object tracking and object classification. These sections are ordered specifically to reflect their interrelation, with those provided later in the literature often building upon aspects of the previous sections. This is especially true with many object tracking and classification applications. Section 2.6 details a number of systems that have been proposed in this area. The literature described builds upon advances in the underlying technologies of the previous sections. Advances in the underlying sections are therefore likely to improve results in these systems. It is also expected that new research areas into higher level applications, such as the extraction of longer term behaviour patterns may build further upon this research.

This thesis considers the areas above to comprise the most relevant literature on underlying technologies upon which the higher level tracking, or track matching across disjoint cameras is built. Other research areas such as behaviour analysis or high resolution personal identification are generally outside of the scope of this thesis, as it is focussed upon reliable information from existing surveillance systems. This thesis does not aim to advance these underlying technologies, rather it looks at mitigating the errors that propagate into higher applications from them. The current literature related to tracking across disjoint cameras is explored in detail within section 2.6. This examines how this relatively new research area applies the techniques that have been developed in the wider tracking literature specifically to tracking across multiple cameras that are not overlapping or near-overlapping.

2.1 Motion Detection and Object Segmentation Techniques

Human brains and computers are not capable of detecting, tracking, and analysing every pixel level change that occurs within the camera views of a wide area surveillance network. In order to reduce the complexity of the information in the system to a manageable size, some assumptions are made. The most widely used assumption in video surveillance is that objects of interest within the field of view are either moving, or once were moving. This is becoming increasingly

important due to the increasing resolution of cameras. Although this improves the amount of pixels upon the objects of interest, it also increases the amount of pixels in the background that are not really of interest. In other areas, such as image analysis, there is only a single image to be analysed rather than an image sequence. Segmenting objects in this area can not use object motion, so it is reliant upon region analysis and edge analysis techniques. These segmentation techniques are not considered widely in this literature review as the speed of motion based segmentation is important. Most methods of motion-based object segmentation fall within the following three categories: background subtraction, temporal differencing, or clustering optical flow. The current state of the art techniques all generate a degree of errors in the objects that are segmented [96], hence a number of post-processing techniques are also used to reduce the amount of pixel noise and to remove object shadows as though they are moving with the object. The shadows and noise are actually not part of the object and therefore are often not interesting for further analysis. The most common of these techniques are outlined in Section 2.1.2.

Background subtraction is a popular method that compares the current image to the background model. Once this model of the background scene has been developed, pixels that are changing from the background can be detected and combined to form motion blobs, or connected regions of foreground pixels. The blobs identified can then be analysed in further processes as regions of interest. Developing and updating the background model is explored in detail in section 2.1.1 as it is a difficult and active research area. This approach has proven very good for static backgrounds, but recent research has focused upon improving the background models [28, 43, 64, 112] to better handle lighting changes and background movement.

Temporal Differencing also uses pixel differences computed between two or more consecutive frames to determine whether the pixels are changing over time [98]. It is computationally simple and adapts well to dynamic environments, such as changing lighting conditions; however it suffers from the foreground aperture problem, where it usually misses detecting some pixels in the middle of an object [47]. This approach is also reliant upon objects continually moving as they will be incorporated into the background once they have paused even for just a few frames. Background motion such as moving trees are problematic as they will be classified as foreground regions, so it tends to be used in indoor environments with simple backgrounds.

Optical flow based segmentation uses flow vectors of changes over time to determine regions of movement. Vectors such as displacement vectors [78] can be

used as the basis of contour tracking algorithms, or used for advanced tracking of articulated bodies. Independent analysis of limb motion can be also be provided and used for areas such as gait analysis, where the relative movements of components of objects may provide extra information. Barron *et al.* [5] provides a good overview of the theoretical basis of optical flow, and early work in the field, with recent work [118] showing that this method can produce clear motion segmentation results and may provide very useful results for the analysis of how the different limbs of a person move. Optical flow could also help with the use of moving cameras [107] as it allows for the estimation and removal of the camera movement information; however it is extremely prone to data-association noise and tends to be computationally intensive.

Background subtraction is the most popular technique for object segmentation because it seems to offer greater accuracy than the simpler temporal differencing technique. It is also more widely understood and considered faster and more reliable than the optical flow techniques for stationary cameras. Optical flow seems to be becoming more popular as both techniques are computationally complex, but optical flow may be able to provide more information about how portions of an object are moving relative to each other, such as limb movements and be able to work with moving cameras. Other approaches have been proposed which combine the different techniques, or extend these techniques to model more than just the background. These include extending the mixture of Gaussian model using the Expectation Maximisation (EM) algorithm to classify pixels into background, foreground, or shadows [36]. This model is continually updated based upon the classification results. The Video Surveillance and Monitoring (VSAM) project [12] combined a simple adaptive background subtraction technique with a three frame differencing technique to create a fast algorithm that seems effective for moving object detection. The trade-off between the accuracy of the information provided and the algorithm speed required for an application is a researcher's most important consideration at this stage; yet it still has not been quantitatively analysed in the literature.

2.1.1 Background Modelling

To correctly identify objects of interest moving in a frame using background subtraction, it is first necessary to model the environment, or background, within which those objects are moving. This usually makes environment modelling the first step of any visual surveillance algorithm. The type of camera and camera motion is important as moving cameras have to be treated differently to fixed

cameras. This is because their background is moving and can be more difficult to model [107, 109]. Though some work has been conducted into recreating 3-D environments [16], this is a computationally expensive technique that is not typically used. Outdoor environments can also have significant motion in the background due to trees and other objects that increase the difficulty of 3-D reconstructions. Illumination modelling within this environment can also provide for colour constancy within the scene; however the complex interplay of multiple and time varying illumination sources on complex 3-D objects of previously unknown composition can make such models computationally intensive and ultimately unrealistic when applied to real surveillance images [72].

More common techniques aim to automatically create a 2-D model of the background in the image plane from a video sequence. This model is usually stored in the form of a background probability density function(pdf) for each pixel location. If it can update for illumination changes and background motion, then it can provide better foreground object segmentation [64]. Early models included many pixel centred techniques such as temporal averaging techniques [60] and adaptive Gaussian estimation [59]. These models worked well under specific conditions, but some aspects of complex backgrounds generate significant errors. The factors that make a background complex were addressed by Toyama *et al.* [112] where they identified the following eight factors:

1. Bootstrapping, where there is no training period without moving foreground objects to fully initialise the background model.
2. Foreground aperture, where the centre of objects may be detected incorrectly as background.
3. Background motion, where objects in the background, such as trees, move.
4. Gradual illumination changes, such as the sun moving through the day.
5. Sudden illumination changes, such as lights being switched on in a room.
6. Objects that have moved becoming stationary, and returning to the background
7. Camouflage of foreground objects, where a moving object looks similar to the background.
8. Shadows effects which are not actual foreground objects are still often detected as moving foreground regions, changing the objects appearance.

The significance of these effects is widely discussed through the literature [12, 28, 43, 45, 47, 90, 96] in the context of their effects on a particular author's video sequence. These discussions are aimed at providing additional motivation for the author's techniques and how they overcome the factors that occur within their video sequence. This demonstrates which techniques can be used to overcome particular factors; however a full analysis of how these factors affect the percentages of moving pixels detected has not been properly conducted. The more commonly used techniques address many of the complexity factors, but Toyama *et al.* [112] also raised the suggestion that some of these factors, especially shadow removal, may be more appropriately addressed in higher models than the background model. This can simplify the motion detection aspect of the process at the expense of adding an extra shadow compensation process; however in practise this seems to be producing more useful results.

Grimson *et al.* [41] first introduce the widely popular Mixture Of Gaussian (MOG) model to allow the background to adapt to small illumination changes and small background movements, such as waving tree branches. This background adaptation can also resolve bootstrapping, and allows for stationary objects that were once moving to shift into the background when they have stopped. It tries to find the most accurate match of a small set of Gaussian curves to the pdf of a background pixel. This is useful as multiple background colours can occur at a single location in the background when it moves slightly, such as a tree swaying. In this instance a leaf may be showing at the pixel during one frame, a branch showing in the next frame, and a wall in the following frame. Three different curves could thus model these three different possible background colours. Much research has been conducted into identifying the number of Gaussians that will provide the most accurate model [90]. Determining the ideal number of Gaussians for each pixel in a view is too computationally complex to run at real time speeds given the current hardware. Typically values between 3 and 5 Gaussian curves have been found to be sufficient in practical applications [41].

Other methods have used multiple levels, or a multiple parameter approach. The Wallflower algorithm, proposed by Toyama *et al.* [112], performs three levels of background maintenance for background subtraction. These are the pixel level, region level, and the whole frame level. This level approach tries to incorporate information about the local, regional and global changes in a video sequence. In the W4 system developed by Haritaoglu *et al.* [43] the three parameters of minimum intensity, maximum intensity, and maximum intensity difference between consecutive frames form a pixel based statistical background model. This system is designed to run in a real-time outdoor environment to perform person

tracking with small background motions and illumination changes. The adaptive background model developed by McKenna *et al.* [77] utilises colour and gradient information to minimise the effects of shadows and poor colour determination effects. The adaptive nature of the model also allows it to compensate for small lighting variations. Many other models [26, 41] can also utilise colour spaces that involve chromaticity, or some other supposedly illumination independent colours, in order to limit the effects of illumination changes, and shadow effects.

Elgammal *et al.* [25, 28] proposed a per pixel non-parametric statistical model based upon Kernel Density Estimation (KDE). This model is adaptive, like the MOG model, so it does not strictly require a training period without moving objects, and it can compensate for some background motion, moving objects becoming stationary and illumination changes. Unlike the MOG method where a small number of Gaussians are used to accurately match the colour histogram, this model uses a small Gaussian kernel for each value found in the history. With typically 50 to 100 samples, this method allows the model to adapt to match the background probability function with a high degree of accuracy, even where the pixel's pdf has many modes, or no obvious modes. Updating the KDE model is also proposed to be simpler and more effective than the MOG model. Like the MOG model, KDE provides consistent results, even with some background motion and small illumination changes. It can be computationally costly though, allowing little time for other information to be extracted in a real-time surveillance system. Recently Elgammal *et al.* [26] has utilised fast Gauss transforms to minimise this problem, however it is still significant for current computer hardware.

More recently the codebook model has been proposed [57] to overcome some of the limitations in the complexity of computation and memory that are associated with the MOG and KDE models, whilst allowing for multimodal backgrounds and varying illumination conditions. This method utilises a training period to create multiple codebooks on a pixel level to create a compact representation of possible background modes. These are modelled with an $\bar{R}_i, \bar{G}_i, \bar{B}_i$ vector to represent the midpoint of the codebook colour cluster, $\check{I}, \hat{I}$ to represent the minimum and maximum brightness, f to represent the frequency of codebook occurrence, λ as the maximum length the codebook is not accessed for, and p, q representing the first and last access times. Multiple codebooks can be used to represent multiple background modalities, allowing for complex backgrounds, however unlike the KDE methods it does not require the storage of past pixel values in memory. Unlike MOG it does not assume the number of modes in the background, it can include brightness components, and also does not require the memory of the KDE method. This model has been recently been extended to

include modelling for the suppression of highlights and shadows [22]. This is achieved through the application of the codebook within the HSV space to apply shadow suppression [17].

The literature has presented a number of techniques for creating complex background models. These models aim to address many of the factors that make a background complex. Toyama *et al.* [112] performed a quantitative analysis of background models presenting figures for false positives and false negatives under the various problems they identified. This analysis does not look at the real-time nature of the algorithms compared, and was also conducted before the KDE background model was proposed. The analysis also raised the issue of whether some of these factors, such as shadows, should be addressed in a separate module, which has not been adequately researched. Finally the question of quality versus speed has not really been addressed in any significant way. Higher resolution images can now provide clearer object information, but require more time to analyse. Combining this increased resolution with the increased computational power available has not yet been addressed, especially for the complex models, which are yet to be proven at real-time speeds.

2.1.2 Removal of Shadow and other Segmentation Noise

This section describes the current techniques that are applied as a post processing step after the segmentation process. It is not aimed at being a full review of the topic, but is provides a consideration of the types of errors that can be mitigated with these techniques. Much of the noise in segmented images is created due to numerous factors ranging from imperfect background models as mentioned in Section 2.1.1 above; however other aspects like vibrations in supposedly static cameras can also be significant. This creates a range of noise types from speckled images to noise around the edge of objects, or the segmentation of objects into numerous parts.

The main methods used in the surveillance field to remove or reduce this noise is morphological erosion and dilation [46]. These processes are generally applied to the boundary of objects in a binary image to dilate (enlarge) or erode (shrink) the object. This removes small regions of speckled noise, or to join together components of an object that have been incorrectly segmented into separate parts. Although these processes are generally applied to binary images describing background and foreground regions, it is also possible to apply them based upon the colour image [122]. In this process, dilation only occurs on pixels of very similar colour to those on the boundary, and erosion only occurs where the edge pixels

are of differing colour to those on the inside of the object. Whilst such colour morphology techniques may provide improved morphology, they have not been widely applied in the surveillance literature, perhaps due to their computational intensity. These morphological techniques can smooth the edges of objects, and remove a significant amount of image noise; however they can lead to segmentation errors where objects consist of narrow protrusions or have detailed outline contours. The impact of such errors are reduced for higher resolution object views, but can become significant when objects are small or far from the camera.

Shadow removal has been recognised as an important problem for the accurate segmentation of objects [95] as object shadows are true changes in the image,but are not actually part of the object itself. Prati *et al.* [95] provide a useful overview of the techniques used to identify regions of shadow that might be segmented as part of an object within an image. They separate the literature into the following four approaches: statistical parametric, statistical non-parametric, deterministic model-based, and deterministic non-model techniques. They suggest that the complexity of generating a deterministic model that can handle many of the factors within cluttered general surveillance scenes with multiple time varying light sources has limited the usefulness of current techniques in this approach. Their results suggest that a general-purpose shadow detection system that uses minimal assumptions should be based upon a deterministic non-model approach; however each of the methods have their own advantages in certain scenarios. Therefore whilst this approach would be widely deployable, improved results could be obtained by utilising a model to make specific assumptions about the shadows that occur within a given camera view. Statistical methods should prove more reliable in indoor environments where the scenes are more stable, making the statistical learning more effective. This approach of choosing the most appropriate method would require camera calibration to improve shadow detection in a similar manner to that suggested both geometrically, or spectrally, to improve colour information or shape information. Indeed [95] suggests that further improvements could come from such specific task/scene domain knowledge.

More recently approaches from other areas have been applied to improve shadow detection. Although some of these techniques provide enhanced shadow identification, most have not been used until recently due to real-time constraints for surveillance images. One area has been the use of intrinsic images [108, 74, 115]. This approach aims to create a model of the illumination that can be removed from an image to obtain a more accurate reflectance model. These techniques have not been widely used in video surveillance because they can be time consuming to optimise for an individual image. These techniques for extracting

intrinsic images are more widely used for colour appearance and are more fully described in section 2.2 below. Graph theory has also been advanced to enhance shadow detection [120]; however the Expectation Maximisation (EM) approach used to maximise the graph probabilities are likely to be too computationally intensive for current surveillance applications.

Other recent work has looked at applying illumination and sometimes camera models in order to identify regions of an image that are affected by shadow. [76] has obtained a patent in the area of shadow edge identification through related changes in all the colour channels within an image. Where the level of change in the three colour channels at a given edge is roughly proportional, then that edge is likely to be caused by shadow rather than a change in the underlying object. [32] extend this idea using a camera model and directed lighting sources to identify shadow regions in greyscale, chromatic and even colour images. These areas can then be backlit using an estimation of the level of shadow to produce images that have used a reduced shadows. This technique seems to work well for outdoor scenes where the sun is a single strong illumination source is present; however may not be as applicable for scenes with multiple time-varying illumination sources as occurs within a building, especially with exterior windows. Instantaneous estimations where illumination sources are only changing slowly may be an easier problem to solve, though little work seems to have been directed at this problem.

Essentially many of these techniques are widely used to minimise a range of segmentation noise factors. Whilst many can improve the object segmentation, mainly through the combination of object parts into a single region and the suppression of shadows, errors still occur. The frequency of such errors is dependent upon the complexity of the scene being observed, including the variation of colour in the background environment, and the stability of the illumination conditions. Significant errors can occur with reasonable frequency, especially in surveillance systems where camera views are designed for human operators rather than automatic computer analysis.

2.2 Colour Space Research

Colour features have long been considered as a significant feature in computer vision, with many different representations and transformations proposed for different tasks. This review investigates some of the most common colour space transformations, and their uses. For convenience the colour spaces are grouped in this section similar to [110] into the RGB colour space, opponent colour spaces,

phenomenal colour spaces, CIE colour spaces, and robust colour approaches. This grouping allows for a discussion of the broad impacts of these approaches to colour spaces, including their advantages and limitations.

The typical RGB image model forms the basis of the standard bitmap image is modelled upon the three channels of the human visual systems [110]. This model is widely used as the basis of the digital imaging with an array of photoreceptors recording light over particular frequency ranges between 300nm and 830nm [86, 117], making it device dependent. Other colour spaces have been developed for their perceptual ability and other particular tasks. These colour spaces involve a mathematical transformation of the RGB values. Once converted, these colour spaces provide some advantages over the RGB space; however they often generate limitations in other areas [34, 110, 86, 117]. Transformations within a colour space has also been proposed to reduce effects such as illumination [33].

Although RGB is good for displaying images, it is not perceptually uniform, as Euclidean distances in the space do not correspond to human perceptual differences in colours. Investigations into normalised distances [70] have increased the correlation between distances and perceptual differences; however research has generally looked into other more perceptual colour spaces. Early work developed the rgs colour space as a minor change to reduce the influence of the overall lighting level upon the spaces; however this still does not significantly reduce the correlation of the three channels, nor the device dependence. Note that this space is sometimes used as a normalised rgb space using a similar manner [77]. The rgs space can be calculated from the RGB space according to:

$$\begin{aligned} r &= \frac{R}{R+G+B} \\ g &= \frac{G}{R+G+B} \\ s &= R+G+B \end{aligned} \tag{1}$$

The German physiologist Ewald Hering proposed the opponent colours theory in the late 19th century [86]. This is based upon the observation that certain hues are never described together, such as reddish-green or yellowish-blue. Whilst this contradicts the theory of trichromaticity, Cotton [14] suggests that this component of the human visual system occurs in the post-receptor retina cells called ganglion cells. Many models of this opponent colour system have been proposed, which are generally suggested to better model the human colour perception [34]. One of the simpler transformations [34] can be calculated as:

$$\begin{aligned} RG &= R - G \\ YeB &= 2B - R - G \\ WhBl &= R + G + B \end{aligned} \tag{2}$$

A log form of this model has also been proposed. Ford and Roberts [34] also describe another colour space proposed by Ohta that approximates a decorrelation of the RGB channels. This would make it more suitable than RGB for many image processing applications. It can be calculated as:

$$\begin{aligned} I1 &= \tfrac{R+G+B}{3} \\ I2 &= \tfrac{R-B}{2} \\ I3 &= \tfrac{2G-R-B}{4} \end{aligned} \tag{3}$$

Y, C_r, C_b is another popular colour space that is proposed for looking at shadow suppression [62]. They compared it with the RGB, normalised rgb, HSV, and XYZ spaces finding it to provide the highest level of foreground object detections with minimal levels of included shadows. The Y, C_r, C_b colour space can be calculated from RGB using:

$$\begin{bmatrix} Y \\ C_r \\ C_b \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ 0.439 & -0.386 & -0.071 \\ -0.148 & -0.291 & 0.439 \end{bmatrix} \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \tag{4}$$

Isaac Newton arranged colour in a circle called Newton's colour circle [14], which forms the basis of a group sometimes referred to as the phenomenal colour spaces. This circle neglects the brightness of a colour and uses the hue and saturation to describe the colours. The Hue attribute refers to the whether the colour is red, green, blue, etc. The Saturation attribute refers to how vivid the colour is, sometimes referred to as the level of non-whiteness [110]. The brightness is also calculated for these spaces as the intensity of the light. This brightness property is often ignored to purposely reduce the impact of shadows in areas such as foreground detection [95]. The Munsell colour space is an example of an attempt to generate a perceptually uniform phenomenal colour space, with 1500 systematically ordered samples [14, 30]. The most common phenomenal space is known as HSV and can be calculated from RGB values as [40]:

$$H = \begin{cases} \theta & if B \leq G \\ 360 - \theta & if B > G \end{cases} \tag{5}$$

where $\theta = cos^{-1} \frac{\frac{1}{9}[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-G)(B-G)}}$

$$S = 1 - \frac{3}{R+G+B} min(R, G, B) \tag{6}$$

$$V = \frac{1}{3}(R+G+B) \tag{7}$$

These phenomenal colour spaces are intuitive to use, and have been included in many commercial applications such as Photoshop; however they also have their limitations [110]. They are mostly linear transformations of the RGB space, and thus are device dependent. This occurs as there is no usage of information about chromaticity or white point used. There is also a problem with the hue discontinuity occurring at 360° making arithmetic calculations more difficult. Indeed [110] suggest that a chromatic *CIE Lab* or *CIE Luv* colour space in polar coordinates may be easier to work with.

The Commission Internationale de l'Eclairge (CIE) is an organisation devoted to international cooperation and exchange of information on all matters relating to the science and art of lighting [110].In 1976 they proposed two colour spaces to provide a perceptually uniform colour space. These were designated *CIELuv* and *CIELab*. This perceptual uniformity aimed to create a high correlation between the Euclidean distance in *CIELuv/CIELab* and the human perception of colour distances using chromatic adaptation and non-linear visual responses [110]. The main difference between these two colour spaces is that *CIE Lab* normalises its values by division with the white point reference, whilst *CIE Luv* normalises its value by subtraction of the white point. Thus the transform from *CIE XYZ* to *CIE Luv* is calculated using:

$$\begin{aligned} L^* &= 116(\tfrac{Y}{Y_n})^{\frac{1}{3}} - 16 \\ u^* &= 13L^*(u' - u'_n) \\ v^* &= 13L^*(v' - v'_n) \end{aligned} \tag{8}$$

for $\frac{Y}{Y_n} > 0.01$, otherwise $L^* = 903.3\frac{Y}{Y_n}$ where u', v', and u'_n, v'_n are calculated from:

$$\begin{aligned} u' &= \tfrac{4X}{X+15Y+3Z} \\ u'_n &= \tfrac{4X_n}{X_n+15Y_n+3Z_n} \\ v' &= \tfrac{9Y}{X+15Y+3Z} \\ v'_n &= \tfrac{9Y_n}{X_n+15Y_n+3Z_n} \end{aligned} \tag{9}$$

where the tristmulus values X_n, Y_n, Z_n are those of the nominally white object colour.

The transformation from *CIE XYZ* to *CIE Lab* is performed as:

$$\begin{aligned} L^* &= 116(\tfrac{Y}{Y_n})^{\frac{1}{3}} - 16 \\ a^* &= 500[(\tfrac{X}{X_n})^{\frac{1}{3}} - (\tfrac{Y}{Y_n})^{\frac{1}{3}})] \\ b^* &= 500[(\tfrac{Y}{Y_n})^{\frac{1}{3}} - (\tfrac{Z}{Z_n})^{\frac{1}{3}})] \end{aligned} \tag{10}$$

The perceptual colour difference can then be calculated as:

$$\Delta E^*_{uv} = \sqrt{(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2} \tag{11}$$

$$\Delta E^*_{ab} = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2} \tag{12}$$

The correlation between the colour difference and human perception is often cited as the main reason for using the CIE based colour spaces. This allows for a reduction in the effects of shadowing and highlights, similar to the human visual system. Whilst such effects may be beneficial for colour image processing, it is does not follow that it provides the best ability to discriminate between objects of differing colours.

Whilst transformations into other colour spaces are often useful for particular applications, especially where colour perception is involved, transformations within a colour space are also possible. Numerous transformations have been developed to improve particular qualities of images, without transferring to a different colour space. These are often based upon adapting the image level colour statistics so that they fit criteria such as histogram spread, minimisation of low frequency components, or other ideas on generating robust colours. The techniques described here are intrinsic images, grey world, white patch, and rank preserving histogram adaptations. Other techniques such as colour clustering to extract colour descriptions are not explored here as they tend to be dependent upon application they are required for and the distance measures that are used to generate the clusters.

Intrinsic images are a good example of using image statistics to adjust an images colours. They aim to generate the reflectance image by discounting sources of light and shadows [108, 74, 115]. This is based upon the definition of an image as a combination of the intrinsic object reflectance, $R(i)$, and the incident illumination, $I(i)$ in the scene. If one considers a small part of the image, denoted here as $L(i)$, then its value is determined by the combination of the illumination in the scene and its interplay with the reflectance of the materials that make up the objects in the scene. This is a complicated combination as the light can bounce

of many surfaces, and the interaction of the light with the object depends not only upon the materials inherent reflectance, but also the angle of the material, and illumination from other sources. This can be written in a simplified form as:

$$L(i) = R(i) * I(i) \tag{13}$$

A full description of how to derive an illumination image from an image sequence is provided by Weiss [115]. This is essentially based upon applying a Maximum Likelihood approach to sets of filters on the image sequence that are treated as independent. This approach thus aims to use the statistics to separate the reflectance from the illumination model. The results provided seem promising, although the filter process does seem to lead to the reflectance images appearing washed out. More recently Tappen *et al.* [108] have proposed an approach for generating intrinsic images from a single image. Again the reflectance image results seem to be washed out with problems in lighter regions, especially where white regions appear close to shades of grey. Sometimes even the distinction between black regions an their white surrounds are poorly extracted. This is because intrinsic images suffer from the same limitation as chromatic spaces in that specific colours, such as white and black, can be very similar chromatically; however differ significantly in their intensity. Such colours are intrinsically different and can be seen as such in colour spaces such as RGB, but appear similar when their intrinsic difference in intensity is reduced or ignored.

Other transforms aim to extract robust device independent colours by analysing the histogram of image colours. The white patch assumption aims to create standardised images by transforming the colour values such that the lightest region in the image is assumed to be white. This then transforms all of the other colour to more fully spread across the histogram, although it can be problematic where an image is quite dark or does not actually include a white region. A somewhat similar approach has been proposed by Finlayson *et al.* [33] to maintain the rank order of the colours in an image, but to fully utilise the full image spectrum through the application of histogram equalisation. This approach forces the image to transform the whitest regions to white, as well as the darkest regions to black, using a non-linear spreading of the pixel colours. The grey world assumption [4] looks to transform the colour at each pixel based upon its difference from the image average for each colour channel. This assumes that although colours may vary significantly, the average of the image should be in the centre of the histogram unless device properties have influenced the image. The grey world transformation can be derived as:

$$\begin{aligned} R' &= \frac{R}{R_{ave}} \\ G' &= \frac{G}{G_{ave}} \\ B' &= \frac{B}{B_{ave}} \end{aligned} \tag{14}$$

where R_{ave}, G_{ave}, and B_{ave} denote the means of the R, G, B colour channel values respectively across an entire image.

Many other approaches, such as Retinex and ACE [99], have also been proposed to transform colours into a more useful form. These are often aimed at a specific task such as removing device based artefacts, making the images appear more natural, or restoring old or damaged images. Thus such transformations may not necessarily make the colours of an object closer to their intrinsic properties, especially where the scene does not include a broad range of colours across the spectrum. It is also worth noting that such conditions are most likely to occur within confined artificial environments, such as the building environments.

This section has described many of the more common colour spaces used in image processing. It is by no means exhaustive as there are many colour spaces or colour transformations which are often derived for specific tasks or applications. It has described many of the advantages, limitations and applications that are appropriate to categories of colour spaces. When determining which colour space to use it is worth considering the specific application use and the impact of shadows. Where shadows need to be minimised, then chromatic spaces such as HSV or Y, C_r, C_b can be applied. Where human perception of the colours are important then the CIE based colour spaces, or the Munsell space should be considered. Where the intrinsic colour is important for comparison, then chromatic spaces, as well as techniques like intrinsic images are not generally going to improve the results over using the standard RGB colour space.

When using any of these colour spaces careful consideration needs to be used to evaluate the impact of the colour space upon the system and its computational complexity. Consideration of whether illumination changes need to be mitigated and how such mitigation might influence the ability of a distance measure to determine the difference between colours. It is also important to consider the distance measures used to evaluate the colour similarity, and how approaches such as normalised distances might reduce the effect of colour saturation to create a better level of colour discrimination. Finally, where automated processes are concerned, the usage of a colour spaces that mimic the human visual perceptual ability may not translate into improved overall results.

2.3 Statistical Similarity Measurements

Colour features have been widely proposed in the surveillance literature as a useful feature for a variety of tasks from tracking to identification [47]. Whilst the particular feature and its representation are important, the determination of the level of similarity between those features are also important. Such similarity measures, also sometimes called distance measures, have been developed in the numerous fields ranging from pattern recognition [21, 23] to information theory [15], and communication applications [54]. These measures are often applicable across wider domains to determine the level of similarity between a variety of features, including colour based features, that are commonly represented as histograms or probability density functions over a spatial region. Zhou and Chellapa [127] have recently surveyed a wide range of these measures. A brief description of the most common probabilistic measures are provided here, outlining their similarities and indicating any features particular to those methods. In order to provide comparable notation, these cases consider a two class problem, where $p_1(x)$ refers to class 1, and class 2 is noted $p_2(x)$ defined for the space $\Re^d$. Is should also be noted that $0 < \alpha_1, \alpha_2 < 1$, $\alpha_1 + \alpha_2 = 1$, and π_1 and π_2 are prior probabilities of classes 1 and 2 respectively.

In 1952, Chernoff derived a method of determining the distance between two probabilistic functions [11]. This is found by investigating the overlap of the two probability functions across the entire space. It can be formally written as:

$$J_C(p_1, p_2) = -log(\int_X p_1^{\alpha_2}(x) p_2^{\alpha_1}(x) dx) \tag{15}$$

Where the special case occurs that $\alpha_1 = \alpha_2 = 1/2$, then the Chernoff distance changes slightly and is known as the Bhattacharyya distance [127]. The Bhattacharyya distance was originally derived as a geometric measure of the angle between two vectors in a high dimensional space, which represent the two distributions [8]. Thus where the distributions are similar, the measure approaches 1. It can be written for the two class case as:

$$J_B(p_1, p_2) = -log(\int_X \sqrt{[p_1(x)p_2(x)]} dx) \tag{16}$$

Another measure known as the Matusita distance [75] can be defined as:

$$J_T(p_1, p_2) = (\int_X [\sqrt{p_1(x)} - \sqrt{p_2(x)}]^2 dx)^{1/2} \tag{17}$$

Matusita [75] noted that his distance was related to the Bhattacharyya distance, though minimising the Matusita distance is equivalent to maximising the Bhattacharyya distance. This relationship can be shown as:

$$J_T = \sqrt{2 - 2exp(-J_B)} \tag{18}$$

The KullbackLeibler (KL) divergence [61] is commonly used in information and probability theory as a measure of the difference between two probability distributions. It is labelled as a divergence as it is not symmetric and therefore gives an indication of the divergence of one distribution from the other. It can be written as:

$$J_R(p_1||p_2) = \int_X p_1(x) log(\frac{p_1(x)}{p_2(x)}) dx \tag{19}$$

The original usage of the KL divergence [61] actually added equation 19 with the divergence $J_R(p_2||p_1)$ in order to generate a symmetric measure. This symmetry is useful as it provides a stable measure for the divergence which is not dependent upon whether p_1 is compared to p_2, or p_2 is compared to p_1. It can be written as:

$$J_D(p_1, p_2) = \int_X [p_1(x) - p_2(x)] log(\frac{p_1(x)}{p_2(x)}) dx \tag{20}$$

Patrick and Fisher proposed a nonparametric method to determine the distance between two probability density functions [89]. This measure has been used primarily in information theory and does not rely upon fitting a particular distribution to the data. As such it allows distances to be calculated between a very wide range of probability distributions. It can be written as:

$$J_P(p_1, p_2) = (\int_X [p_1(x)\pi_1 - p_2(x)\pi_2]^2 dx)^{1/2} \tag{21}$$

The Kolmogorov distance, when applied to the two class problem, relates to the amount of non-overlapping components of the probability density functions. Thus the distance is 0 when the functions are overlapping, but will equal 2 where the functions do not overlap at all. This can be formally written as:

$$J_K(p_1, p_2) = \int_X |p_1(x)\pi_1 - p_2(x)\pi_2| dx \tag{22}$$

Lissack and Fu explored the existing distance measures through a thorough analysis of the probability of error P_e [67]. They determined the relative probability of errors for existing measures for an M-class problem. For the case where $M = 2$ they found that the upper bounds for the functions described here relative P_e's are:

$$P_e = U_K < U_B = U_T \tag{23}$$

where U_K denotes Kolmogorov's distance, U_B denotes Bhattacharyyas distance, U_T denotes Matusitas distance.

It is also important to note that where $M > 2$, U_K is no longer equal to P_e, though it does tend to remain closer than the distance measures [67]. Given these results, they decided to extend the distance measure to also include a proportion relating to degree of overlap between the functions. For the two class case this can be written as:

$$J_L(p_1, p_2) = \int_X |p_1(x)\pi_1 - p_2(x)\pi_2|^{\alpha_1} |p_1(x)\pi_1 - p_2(x)\pi_2|^{\alpha_2} dx \tag{24}$$

Thus the Kolmogorov distance is a special case of the Lissack-Fu distance where $\alpha_1 = 1$ [127].

This section has provided a brief outline of a popular range of distance measures as they relate to the two class distance measures. This distinction is important as many other distance measures have been developed or enhanced to specifically cater for cases where there may be more than two classes. It is important to note that distance measures determined by each of these criteria are not necessarily directly comparable to each other, the probability of error P_e in class distinction is a reasonable measure with which to compare these measures [67]. A further consideration is that computing such probabilitisic distance measures is non-trivial, and only determined for a certain range of parametric families, such as Gaussian densities [127].

Further details including other interesting properties of these and other possible distance measures can be found in a variety of sources, including [21, 54].

2.4 Object Tracking

Once objects have been detected and segmented, they can be tracked through the application of probabilistic data association. This section recognises that much of

the traditional tracking literature focuses upon techniques for the modelling and prediction of object location, but only gives a brief overview of the definition of the tracking problem before focusing upon the data association component which forms the larger part of more recent surveillance based tracking techniques.

Tracking within the video surveillance scenario can be described as the updating of an object model or feature set along a frame sequence where that object is viewed [121]. This aims to associate the data about individual objects together into sets, also called a track, as they are viewed throughout a camera. This track data can provide information about that particular object, such as its motion, where it is located in each frame of the sequence, as well as its shape and appearance features. Within single camera views many approaches have been proposed, though the most successful have been based upon object motion information. These often use particle filters or variants upon these techniques, such as those presented by Arulampalam *et al.* [3]. Arulampalam *et al.* presents the tracking problem as a state based system that changes over time using a series of often noisy measurements which are made upon that system. The state based system model can be given by:

$$x_k = f_k\left(x_{k-1}, v_{k-1}\right) \tag{25}$$

where f_k is a possibly non-linear function x_{k-1} and v_{k-1} models the process noise. Tracking aims to recursively estimate x_k from the sensor measurements, or observations, which can be modelled as:

$$z_k = h_k\left(x_k, n_k\right) \tag{26}$$

where h_k is also a possibly non-linear function x_{k-1}, with n_k modelling the measurement noise. Tracking is thus based upon the filtered estimates of x_k based upon the set of all available measurements $z_{1:k} = \{z_i, i = 1, \ldots k\}$ up to the time k.

Arulampalam *et al.* [3] also describes this tracking problem from a Bayesian perspective, which adds a level of belief to the state x_k at time k based upon the given data $z_{1:k}$ also up to the time k. This requires the construction of the pdf $p\left(x_k|z_{1:k}\right)$ assuming that the initial state vector, or prior, is known. The pdf may then be obtained recursively through the two stages of prediction and update. The prediction stage uses information from the pdf at time $k-1$ and the system model to obtain the prior pdf of the state at the time k via the Chapman-Kolmogorov equation:

$$p\left(x_k|z_{1:k-1}\right) = \int p\left(x_k|x_{k-1}\right) p\left(x_{k-1}|z_{1:k-1}\right) dx_{k-1} \tag{27}$$

At the time step k, a measurement z_k becomes available, which can be used to update the prior using Bayes' rule:

$$p\left(x_k|z_{1:k}\right) = \frac{p\left(z_k|x_k\right) p\left(x_k|z_{1:k-1}\right)}{p\left(z_k|z_{1:k-1}\right)} \tag{28}$$

where this measurement is normalised using the constant

$$p\left(z_k|z_{1:k-1}\right) = \int p\left(z_k|x_k\right) p\left(x_k|z_{1:k-1}\right) dx_k \tag{29}$$

which in turn relies upon the likelihood function $p\left(z_k|x_k\right)$ that is defined in the measurement model.

It is worth noting that it is the recurrence relations between equations **(27)** and **(28)** that form the basis of the optimal Bayesian solution. This theory has lead to research suggesting a number of algorithms ranging from the popular Kalman Filter and its extended version to grid-based methods and several other variants on the particle filters.

The theoretical description of the tracking problem outlines the problem; however the surveillance community is actually more interested in the data association that is associated with tracking. This area is focussed more upon the models and features that are used for tracking purposes, rather than methods to solve the tracking problem itself; however it is also important to note that a good predictive model can help with the data association problem. These models are important as tracking methods based purely upon the location of an object blob have many limitations, and work best at high frame rates where the differences between the predicted and observed positions are likely to be very small. Where there is a significant time difference between prediction and measurement, the difference between prediction and measurement can cause significant errors in utilising these techniques. These complexity factors can cause significant errors within single or overlapping cameras; however the problems become so large when cameras are significantly disjoint that this area of research is therefore considered to be of a different nature to the traditional image tracking techniques as described in detail in section 2.6.

The rest of this chapter focuses first upon exploring the factors that add complexity to the tracking problem, Section 2.4.1, as well as the feature based tracking, Section 2.4.2, and model based tracking techniques, Section 2.4.3, that are

used in the surveillance literature aimed many at single or overlapping camera views. Finally the chapter ends with a summary of the ability of the tracking techniques to overcome the complexity factors and how they trade off between algorithm complexity and accuracy.

2.4.1 Complexity Factors within Object Tracking

Although much research has been conducted into effectively solving this tacking problem, a number of factors still cause significant complexity and errors in even the most sophisticated methods. These can be grouped into the six main complexity factors:

1. Number of detectable objects, because tracking requires an algorithm to match each object to a track to follow its motion in each frame.

2. Segmenting or tracking individual objects within a group as the group may split up, or other objects may join the group for a period of time.

3. Partial object occlusion, where one object covers a portion of another object in the cameras view

4. Total object occlusion, where an object is completely blocked from sight by another object.

5. Poor object segmentation, because it can alter their appearance. This could be due to a number of problems ranging from occlusions, to a lack of contrast between the object and the background.

6. Mutual similarity between objects, which could lead to incorrect data association.

Many of these factors are analysed in the literature when tracking methods are compared [24, 36, 44, 47, 79, 107]. The focus on errors from occlusions and object segmentation has lead to some quantitative analysis of methods [64]. Many of the existing tracking algorithms lead to good results in simple video sequences, and often reasonable results with limited errors for the authors chosen operational environment; however little work has currently been done to fully compare many of the common techniques quantitatively across this list of complexity factors, making it difficult to evaluate their usage across a wide surveillance system. The increasing amounts of available computational power and camera resolution have

also lead many researchers to move on from evaluating the tracking problem towards analysing higher level information from the tracks. The six main complexity factors still cause error rates that may impact upon further analysis, yet there has been little research into how increased resolution affects track analysis, or the accuracy and speed of the existing tracking algorithms.

2.4.2 Feature-Based Tracking

Feature-based tracking algorithms extract geometric elements, such as corners and vertices and cluster them into higher level features to perform object recognition and tracking. These features can be matched between images to perform recognition and tracking, and are often fused together to create stronger feature groups that provide higher accuracy under partial matching . Three sub categories of the feature based algorithms can be considered based upon the level of the feature used. These are:

- Global feature-based algorithms use centroids, areas, colours, perimeters, and other features as the basis of their features. An example is where a person is tracked using their centroid and bounding box [92]. As long as the centroid velocity of two people can be distinguished, then tracking is successful even during occlusions.
- Local feature based algorithms use features such as line segments, curves, and corner vertices.
- Dependence-graph-based algorithms include a variety of distances and focus on the geometric relations between features.

Hu *et al.* [47] indicates that feature-based algorithms, with the exception of dependence-graphs, can adapt quickly and successfully as they operate within 2D image planes. This makes them suitable for real-time tracking of multiple objects in areas with high object traffic. For example Jang and Choi [51] have combined features into an active template that dynamically characterises the structural and regional features of an object based upon shape, colour, edge, and texture features of the region. Minimising a feature energy function of a Kalman filter based motion estimator during the feature matching process allows their system to track non-rigid moving objects. Dependence-graph-based algorithms may provide more accurate results, but they do require computationally expensive feature graph matching, making them only suitable for tracking where there are minimal

objects, or offline analysis. These algorithms may handle a limited amount of partial occlusions by utilising dependence graphs and local features; however these algorithms have a low recognition rate due to image distortions in the camera projection, are generally unable to recover the 3D pose of objects. They also tend to suffer significant degradation with poorly segmented foreground object regions.

2.4.3 Model-Based Tracking

Model-based tracking is perhaps the most common method in the tracking literature. The models are usually generated offline by combining features inherent to the structure of the object using prior information about likely object types. These models form the basis of the standard tracking equations **(25)** and **(26)**. Appropriately chosen models will allow for a system to become more robust to the expected noise in the system, and can become robust to typical minor error sources such as small segmentation errors. Tracking is usually performed by recursively predicting the location of the model for the next frame in the sequence using equation **(27)**, and updating the measurement of the model based upon the actual view of the object within that frame of the image sequence using equation **(28)**. These steps aim to produce the best 'track', or proposed position of the model location in each frame of an image sequence. The models will sometimes incorporate an object appearance component which is also updated to improve the model, as this additional information assists with overcoming factors, such as occlusions. As rigid bodies, such as vehicles, are often much simpler than their non-rigid bodies, the literature tends to use simpler model assumptions. For this reason the literature tends to be separated into human body model tracking and vehicle model tracking, with the majority of research in surveillance focusing on human body movements. This thesis is based within a building surveillance environment, so this section will focus upon the human model-based tracking methods.

Human body model-based tracking is often done in an analysis-by-synthesis method [47], or through the application of generalised appearance models. In analysis-by-synthesis the next pose for the human body is estimated and projected into the image plane for comparison with the image data according to a similarity function. This may be recursive, or use sampling techniques, but once a match is found, then the model is updated. Hu *et al.* [47] highlights the main issues with this method as constructing the human body models and their representation, constraints, and the prediction and search strategies used to match the models with the actual image data.

Four main structures are used throughout the literature to construct and rep-

resent the human body models. These models use different assumptions about the nature of the human body projected into the image plane, and the amount of information required to model a human body accurately enough to effectively track it. Again this is an example of the trade-off between model accuracy and the algorithm speed required to reach real-time processing speeds. The four structures are: stick figures, 2D contours, volumetric models, and hierarchical models. These human body models are combined with human motion models to provide more body pose or behaviour information as well as more accurate tracking. Other search strategies can also be used in combination with the human body model to provide faster or more accurate model matching for the tracking process.

Stick models essentially model the body as a combination of rigid sticks, to represent the bone structure of a human. These are allowed to move relative to each other. Zhao and Nevatia [125] use such an articulated human stick model to provide some robustness to occlusion and to perform gait analysis using leg motion templates for walking and running. These motion templates are generated offline for matching with humans in other video sequences.

2D contours, or edges, are developed from the projection of the human body into the image plane, and consist of human body segments modelled by 2D ribbons or blobs. Elgammal *et al.* [28, 24, 27] constructs a three blob model consisting of a head region positioned above the torso region, which is above the leg region, where the region locations are determined from prior offline analysis of upright humans. This colour model, bounded by the contours can then be used to handle occlusions as the human model can be matched to the visible component of the appearance model. This model is only useful where a human is in the upright position; however this is common in video surveillance footage. Further abstraction of this using human pose classification or a stronger model could ensure the appearance is not corrupted by errors, or could adapt the model to be more relevant for other positions such as crawling or sitting.

Active contour-based tracking methods also use 2D bounding contours, or edges, to represent objects rather than extracting objects using background models. These active contours are updated dynamically to better fit the model throughout successive frames of an image sequence [79]. Koller *et al.* [60] applied active contour-based tracking to vehicle recognition, achieving real-time tracking of vehicles through a road segment. This provided tracking of vehicle shapes through a simple environment with a high degree of structure. Isard and Blake [50] use stochastic differential equations as the basis of their condensation motion models, and combine them with deformable templates to address the challenges of people tracking in near real-time speeds. Paragios and Deriche [87] use a geodesic

active contour and a level setting scheme to track various moving objects. The tracking results for both human and vehicular objects using these techniqueslook promising; however the computational cost of the algorithm seems quite high.

Active contour based algorithms generally provide a more simple and effective representation than region-based tracking with lower computational complexity [47]. They seem to provide more robust tracking under occlusions and background disturbances, but are still limited to contour based information. This can make determining further information such as 3D object pose difficult. Initialisation sensitivity is also a large problem with some of these methods [87] making automatic tracking initialisation difficult.

Volumetric models are constructed based upon the fact that humans are 3D objects that are projected into the image plane. Thus the representation of the object can be 3D and it can also be projected into the image plane to provide more accurate representations under object rotations. Zhao and Nevatia [125] use an ellipsoidal 3D model to represent people in their system and utilise a ground plane constraint to minimise the effect of shadows, as these predictable occur along the ground plane in open environments. The simplistic 3D modelling of the environment and illumination sources also provides more information that can be analysed in the scene.

Hierarchical models are seen as a way to more accurately model the complex structure of the human body. Haritaoglu *et al.* [43, 44] use a hierarchical model of human body parts on the body silhouette boundary. The position of body parts is calculated using a likelihood function, and is further used to estimate the body posture of a human in the scene. Plankers and Fua [91] present a model of the human body where the hierarchies are the skeleton, ellipsoid balls simulating tissues and fats, polygonal skin surfaces, and shaded rendering.

Motion models of human limbs and joints are widely used in the literature because they provide strong and realistic model constraints. These use prior knowledge of human motion to help recognise human behaviours [125]. These often use Hidden Markov Models (HMM's) and their variations generated from offline image analysis for comparison with image data to determine human motion behaviour such as walking, running, or standing. Zhao *et al.* [126] use the minimum description length (MDL) paradigm to develop their structured motion model for ballet dancing. Ong and Gong [84] use a hierarchical Principal Component Analysis (PCA) to develop their motion model based upon the matrices describing the transition probabilities within, and between global eigenspaces. Ning *et al.* [82] develop their motion model from training examples and represent it using Gaussian distributions.

Other search strategies aim to reduce the computational time of pose estimation and object matching for tracking by developing appropriate object constraints. The four most common areas of research are modelling the human body dynamics, applying Taylor models or Kalman filters for estimating the position or appearance, or using stochastic sampling. The most common stochastic sampling technique is the CONDENSATION algorithm developed by Isard and Blake [49, 50]; however Hu *et al.* [47] also point out that other strategies such as Markov Chain Monte Carlo and Genetic Algorithms have also been applied.

Model-based tracking makes use of prior knowledge of the 3D shape of objects, and also online knowledge of the basic appearance of individuals. This attempts to make them more robust under occlusion, self occlusion, and to distinguish objects moving within a close group. The structure and constraint of human and vehicular motion can be used as prior knowledge which can be fused with other methods. 3D models also naturally acquire the 3D pose of an object from the calibrated 2D image scene. The models are also robust to orientation changes that can generate significant changes in object appearance. Unfortunately these methods do require fairly accurate 3D models, high computational cost, and they can be affected by segmentation errors.

2.4.4 Mean Shift-based Tracking

Another commonly used tracking algorithm is based upon the mean shift algorithm. This approach applies mean shift as a robust statistical method that looks to find the local maxima in a given probability distribution. These methods use a search window over a section of the possible distribution to find the maximum in that window. The window can then be adjusted to this location with the search recomputed. This process is repeated until the solution converges to a local maximum.

In video surveillance, many approaches have have proposed to use this method for tracking individuals, with [13] being one of the more commonly referenced approaches. These methods use a probability value at each pixel to represent how likely it is to be the location of the object being tracked, usually based upon its colour. This creates a 2D probability distribution upon which the mean shift process can be applied. The models used to generate the probability at each pixel are becoming increasingly more complex and adaptable. This is increasing the robustness of the tracker; however they are relying upon increases in computational power over time to allow these methods to run at real-time speeds.

A recent paper by Artner [2] has compared some of these mean shift trackers

and found that they can provide reliable and robust results. It also outlines the following six factors that are required to provide the best results:

- The target is mainly composed of a single colour.
- The target does not change colour.
- Illumination does not change dramatically.
- There are no other objects in the scene similar to the target.
- The colour of the background differs from that of the target.
- There is no full occlusion of the object.

Many of these factors do not occur very often for some camera views, suggesting that the mean shift tracker may be appropriate in some cases. Unfortunately in real surveillance systems many people may be observed at the same time with similar colours, illumination can change significantly, and there are times where the background may be similar to the target object. These often cause some degree of error with other tracking techniques; however they tend to completely break mean shift based tracks.

2.4.5 Summary of Tracking Literature

Many tracking methods are currently proposed for tracking objects in a variety of scenes. These methods are essentially dependent upon the model that they use, and so vary greatly in their complexity and accuracy. To date none of these techniques have been proposed to work perfectly for use throughout an entire surveillance, even though many have produced good results for tracking within their specific assumptions. This is perhaps due to non-generalisation of models between object classes, such as humans or vehicles, as different features remaining invariant or discriminative within the different classes; however it is also likely due to the focus of research groups upon particular solvable pieces of the very difficult surveillance task.

Many appearance based models can overcome some of the complexity factors such as occlusions and poor segmentation using features such as object colours; however sustained total occlusions as well as the combining and splitting of groups of individuals still commonly cause errors. Post-tracking integration of tracks which have become separated has not yet been widely addressed, even though the

object models could potentially be matched to perform these tasks. The main difficulty is that the motion information can be unreliable and many models do not provide enough information to correctly reconcile these tracks through feature matching of the models.

2.5 Object Classification

Once moving regions have been identified and segmented, it is possible to try and classify the object, or objects, within that region. The image sequences being analysed may contain many different object types such as humans, vehicles, and other moving objects, like moving clouds. By classifying objects into broad groups, more accurate assumptions can be made about the way they are likely to move and behave for improved modelling, tracking, and other purposes. For example cars are most likely to stay on paved roads with a relatively constant velocity, whilst humans may have more erratic movement patterns.

Object classification is a pattern recognition problem with the two main approaches being shape-based classification, and motion-based classification. These methods are both being improved by increased object resolution in the footage. Shape-based classification is performed using a combination of a variety of object information from silhouettes, to blobs, and the bounding boxes of a moving object. The VSAM system [12] uses a combination of the image blob dispersedness, image blob area, and apparent aspect ratios as key features to classify the moving objects in their system into four categories: single human, vehicles, human groups, and clutter. This system uses a viewpoint-specific three-layer neural network classifier, but has also been extended to further classify vehicle types such as a van, or sedan. Lipton *et al.* [66] uses the area and dispersedness of image blobs to classify their moving objects into humans, vehicles, and clutter. These classification results are improved by including temporal consistency constraints within the classification process. McKenna *et al.* [77] uses human silhouette patterns to separate individual humans within a group based upon head location.

Motion-based classification tends to be based upon rigidity and periodicity of moving objects. Lipton *et al.* [65] uses residual flow to analyse both the periodicity, and rigidity of a moving object. Rigid objects are expected to produce little residual flow, whilst non-rigid objects such as humans produce a higher average residual flow, and also exhibit periodicity. Analysis of these features can then be used to distinguish between human motion, and the motion of other rigid objects, such as vehicles. Cutler and Davis [18] detect and analyse periodic object motion using a similarity-based technique. A moving object is both tracked over time,

and has a self-similarity measure computed over that track. For periodic motion, the self-similarity measure should also show periodicity. Time-frequency analysis can be used to analyse the self-similarity measure for periodicity, which can then be used to classify the object.

These classification techniques utilise the similarity of object appearance, to a model, or its change over time. Because the object appearance is reliant upon the accuracy of the object segmentation and the illumination conditions, the type of object segmentation is important to the classification accuracy. The literature has not analysed the impact of the segmentation technique upon the accuracy of classification and tends to present the object classification step as merely an extension of the object segmentation, rather than as an independent step. The impact of illumination upon appearance, though widely recognised as a problem, is often not mentioned unless the work directly relates to the mitigation of the illumination effects.

2.6 Current Disjoint Camera Tracking Methods

Accurately identifying the movements and behaviours of individuals around a surveillance system has been a focus of security systems long before the invention of computing systems. Such systems have traditionally focussed upon securing important objects, information or people within secure areas with limited access for their protection; however an increasing amount of investigation, especially in the area of policing, goes into understanding the movements or interactions of individuals within areas of interest which may not be secure. With the development and widespread instalment of video cameras in the form of Closed Circuit Television (CCTV) systems, information from video cameras about events in the vicinity of the investigation are increasingly being used to provide additional information. This is becoming important as groups such as terrorists increasingly target publicly accessible locations to spread terror, rather than attacking secure military installations, where the chances of success are more limited.

Locations of interest often consist of a set of buildings and their grounds, although more widespread sets of connected infrastructure such as train systems could also be considered. These areas often have surveillance systems installed with many camera views displayed and monitored by security guards to attempt to identify individuals or crowd behaviour that is out of the ordinary. Much research has been conducted into understanding and identifying such behaviour within a camera view; however many examples arise where the motion of individuals throughout a surveillance system could provide extra information for forensic

tracking, or for identifying the present location of key individuals to assist security guards. If this could be automated, or even just parts of it automated with a manual review process, then it could be used to process large numbers of cameras that security operators find difficult to analyse, or could at least reduce the manpower required to perform the task. Such systems could assist police or security officers to monitor and follow suspects of interest throughout a system, or for forensic analysis of events, such as identifying the movements of suspects throughout the train network in the 2005 London bombings.

Working within surveillance systems that were designed primarily to assist human operators raises many problems when trying to automate various tasks. The two primary considerations in existing surveillance systems are the installation cost and operational requirements. The camera and system installation costs leads to large areas which have limited or no camera coverage; however differing camera properties and viewpoints, imperfect object segmentation, occlusions, and variable and unpredictable illumination conditions all increase the difficulty. The operation of current surveillance systems are also reliant upon human monitoring rather than automated systems, as human inference does not rely upon full coverage to track people throughout the system.

These difficulties have lead to a limited amount of literature aimed at addressing large scale surveillance within real systems. Recent advances in camera technologies and underlying research areas such as object segmentation and tracking within a single camera have lead to increased interest in this area. Each of the key approaches in the literature will be discussed in detail in this section along with their assumptions and limitations. The approaches described in detail are given in chronological order of publication and include the following papers:

1. Cai and Aggarwal [10] - Automatic Tracking of Human Motion in Indoor Scenes Across Multiple Synchronized Video Streams

2. Huang and Russell [48] - Object Identification: A Bayesian Analysis with Application to Traffic Surveillance

3. Orwell *et al.* [85] - Multi-camera Colour Tracking

4. Darrel *et al.* [19] - Integrated Person Tracking Using Stereo, Colour, and Pattern Detection

5. Collins *et al.* [12] - Algorithms for Cooperative Multisensor Surveillance

6. BenAbdekader *et al.* [7] - Person Identification using Automatic Height and Stride Estimation

7. Tan and Ranganath [106] - Multi-Camera People Tracking Using Bayesian Networks

8. Javed *et al.* [52, 53] - Tracking across multiple cameras with disjoint views and Appearance Modeling for Tracking in Multiple Non-Overlapping Cameras

9. Hampapur *et al.* [42] - Smart Video Surveillance: Exploring the Concept of Multiscale Spatiotemporal Tracking

10. Zajdel and Krose [123] - A Sequential Algorithm for Surveillance with Non-overlapping Cameras

11. Gandhi and Trivedi [38, 39] - Panoramic Appearance Map (PAM) for Multi-Camera Based Person Re-Identification

12. Yang *et al.* [119] - Human Appearance Modeling for Matching Across Video Sequences

Cai and Aggarwal [10] present what is possibly the earliest work in people tracking across multiple cameras that are not necessarily overlapping. This work looks at determining the components of their single view tracking system that can remain stable across multiple cameras. They mainly utilise a location feature within a scene model using camera calibration between the camera views, which is only useful for overlapping or almost overlapping cameras. They propose using an average between the image intensities in both cameras to model the difference in illumination levels in the camera views so they can use intensity values of a set of points along the medial axis of the upper body. Such camera dependent transformations have formed a component of much subsequent work; however a simple average difference between the illumination intensity in one camera view and a second camera view is unlikely to effectively model the complex interplay of multiple time-varying illumination sources.

The second approach by Huang and Russell [48] also performs disjoint camera matching using a probabilistic approach to track the motion of vehicles that are observed in two distant cameras along a motorway. They use a fully observable model of the appearance to perform an exhaustive matching between the between the car observed in one camera and then observed in a second camera. This model

includes a number of parameters for each vehicle including colour statistics from the HSV space as well as length and width. They also propose to use an online recursive model to update the appearance parameters for changing conditions over time, such as changing levels of sunlight. This system is also very dependent upon the link travel time between the observing cameras. An association matrix is used to select the most appropriate matches between the two observations, and is only two-dimensional for this simple two camera case. Pasula *et al.* [88] expand this work by exploring the scalability of this traffic tracking system to cover multiple cameras. Their findings suggest that the original model does not scale well for larger cameras systems, as the method requires a propagation of correct matchings through the whole observation chain rather than being able to handle the decomposition of a global model for many sensors. For instance, if an object is correctly observed in cameras A and C, but in the intermediate camera B, then the correct object associations become more difficult than the two camera case as some features such as the speed of travel and lane position are likely to become less reliable. Other intrinsic features, such as colour, length and width, become conditionally independent across the wider surveillance system and provide similar levels of accuracy independently of which camera views they are being matched from. This is a very important finding as it clearly demonstrates that the invariance of the features over time are crucially important, especially when there are large gaps between coverage leading to greater variability in motion based features.

The third early approach in tracking people across multiple cameras was conducted by Orwell *et al.* [85] utilising the power of colour appearance for matching individuals. This work is based upon extracting a model of an individual's colour by performing colour clustering using a mixture of Gaussians, which are optimised by an Expectation-Maximisation algorithm. The changing illumination level is mitigated by adding an expected level of noise to the system to model the expected changes. Where there are multiple known observations of an individual within the system they propose to even estimate this illumination noise. As with all illumination mitigation approaches this added level of noise allows for the same colour under different illumination to be more accurately matched; however it also allows for differing colours to have a higher level of similarity. This paper also notes that other object features such as shape and position could be fused with the colour features in order to improve the accuracy to a level where it might be useful in an automated system.

The fourth approach by Darrel *et al.* [19] presents a system that could be used for a true disjoint camera tracking system. It is based upon the fusion of height, colour, and facial features from what seems to be a single stereo camera set which

users interacted with over a period of time. The height estimates were obtained using stereo depth perception to identify the best possible head point. This is likely to produce very high levels of accuracy, which could potentially overcome to some degree the similarity in height of many people [101]. The colour features used in this work can be separated into skin colour, which is used to identify the face and also seems to be compared for identification, the hair colour, consisting of the candidate head region which is not identified as the facial skin colour, and the colour of the rest of the person. The facial features used in this system are of less interest because this system tends to have had users approach it to obtain close facial shots. Such detailed facial images tend to be obtained sparsely in real systems. The medium and long term results of this system are promising; however the object resolution seems to be considerably higher on average than what would be typical for surveillance and the testing is performed on a single camera, reducing scene illumination changes. Stereo cameras are often not available in current surveillance systems.

The fifth approach is by Collins *et al.* [12] who created the VSAM system. VSAM is aimed at tracking the ground position of vehicles, people and groups of people throughout an entire site. This makes no assumptions about indoor or outdoor environments, though the work presented was from outdoor scenes with cameras that can move their field of view. This can be described as an active system as it aims to utilize multiple redundant cameras in order to track a specified object through the site to provide real-time information, rather than potential post-event analysis. It uses multiple sensors to try to ensure that at least one sensor is tracking the targeted object through a previously created three-dimensional model of the site. Such a system is based largely upon the assumption that there are very few regions where the object is out of view such that there is little uncertainty about their group plane position or geolocation. This geolocation is the dominant feature, with Multiple Hypotheses Tracking (MHT) performed if there is uncertainty. The object classification of vehicle, person or group of people is used with object colour to verify the most likely hypothesis. This work seems to be aimed at continual active tracking of an object through the system, and does not provide any provision for an object's track being lost and later discovered. In such a case the two object tracks seem to be considered as if they are separate objects, which could perhaps be manually reconciled by a human operator. This system does not consider reidentification or matching of objects other than reconciling their location between two views, minimising its usefulness in real systems with large gaps in coverage.

The sixth approach is by BenAbdekader *et al.* [7] who developed a system

based upon height estimation and stride and cadence based gait measures for identification. The height estimates are obtained by converting the image based height of the object's bounding box into a real world height through camera calibration with a reported accuracy of within 3.5 centimetres. This calibration is also used with temporal information with respect to frame rates in order to provide the stride and cadence-based gait information, although it is reliant upon a frame rate higher than twice the gait frequency. The results presented for this system are promising; however it appears to run on a single camera, in a single area. This reduces the effect of errors in camera calibration and differing frame rates, as well as eliminating the variations in gait that might occur with different ground surfaces. These factors make a full evaluation of the system to wider surveillance difficult.

The seventh approach by Tan and Ranganath [106] utilises facial features as well as the colour and texture of clothing to determine the identity of an individual from a database of possibilities. Their results indicate that individually clothing texture is their most accurate feature, with clothing colour also providing high accuracy. Facial features were not found to be as accurate with accuracy of only 58% recorded in the database of only 11 people. The fusion of these features was found to provide matching accuracy of individuals over 95%, although this also uses transitional dependencies, which may not be so useful for a wide area surveillance system. Although the number of people investigated within this system is low, the clothing colour and texture features would seem to be promising techniques for future systems.

The eight approach by Javed *et al.* [52, 53], from the University of Central Florida investigated tracking across disjoint cameras. Their two primary contributions extended a probabilistic model to automatically determine transitions between non-overlapping cameras[52], and determining a method to generate inter-camera Brightness Transfer Functions(BTFs) [53]. Initially the work investigated automatically modelling path probabilities and the transition times between cameras, as opposed to using the manually generated ones developed by Kettnaker and Zabih [55]. They also allowed this camera transition model to automatically update with changing traffic flow. Whilst the path probabilities in some situations can correlate with actual movement, this is not necessarily generalisable to all areas under surveillance. For example, in a corridor where cameras are situated at either end without any exits, a person moving into the corridor from the camera view at one end could be expected to either enter the corridor view at the other end after a period of time, or possibly turn around within the corridor and re-enter the view of the first camera. Whilst most traffic may take approximately the same time to move along the corridor, using such transition times may have a significant

impact upon the accuracy where the individual does not conform to the general model. This could be important as cases where the transition time is significantly different could occur due to activities of interest between those views, such as stopping to steal an art work. The second component of Javed *et al.* 's work [53] involves calculating BTFs. This is significant as it allows for people observed under differing cameras to be transferred to a similar level of brightness in order to compensate for illumination changes. These transfer functions require a number of assumptions, such as the background indicating the scene level illumination changes, and appropriate BTFs can be determined quickly where there are time varying illumination levels, such as natural sunlight. It also assumes people within the scene are flat to limit the complexity of the interplay between the illumination and the observed individual. Even with the low dimensionality of the range of BTF's between cameras, the effect of illumination sources on the background are often not necessary a good indication of the complicated interplay of time-varying illumination sources on the 3-D surface of an articulated moving object.

The ninth approach is the People Vision project conducted at the IBM T.J. Watson Research by Hampapur *et al.* [42]. This is possibly one of the most complete video surveillance projects designed around improving a surveillance system to be more effective. It combines and improves upon a variety of baseline technologies whilst keeping in mind the implications of large surveillance installations. The system aims to improve baseline object segmentation techniques through the application of optical flow to determine salient motion. The tracking system is based upon the development of a multi-blob tracking system that uses shape and appearance measures to overcome occlusions [104]. Such shape and appearance models can also be used to perform tracking between disjoint cameras around the surveillance system, although this is not explicitly stated within their research. The suggestion of the usage of wide-baseline stereo to obtain accurate localisation of an individual is also very useful where they are available; however the prominence of stereo cameras would suggest its increased usage when considering automated surveillance system. Object classification is also discussed as a technique to extract extra information about the objects within the system at real time speeds. They also propose to identify the head location of an individual through an analysis of the silhouette to identify the extremities through their distance from the centroid of the object. Once the head is located, then a PTZ camera could be used to obtained a zoomed facial shot that could be useful for either surveillance personnel, or an automated facial identification system. When these technologies are applied with long term monitoring and motion analysis, this research promises to provide a powerful tool for surveillance.

This system [42] provides many possibilities for the identification of abnormal behaviour to alert security staff; however the most powerful concept proposed by this group is the semantic storage of the data into the Viewable Video Index (VVI). This searchable database promises to provide queryable semantic information in the form of object, temporal, or spatially based information. Obviously the most powerful queries are likely to cross these boundaries, such as finding all the red cars that travel over 20 kph in the viewable region that relates to the car park. The combination of all of these technologies and techniques provides probably the most sophisticated surveillance research system to date and also indicates that future surveillance installations should seriously consider their technologies to provide useful automated or improved semi-automated surveillance. These considerations include spending more money upon infrastructure such as wide-baseline stereo cameras as well as PTZ cameras for large common areas, as well as serious consideration for the placement of cameras to cover other areas such as building corridors.

Much work has been conducted into the determining path probabilities, culmination in the tenth approach in a recent article by Zajdel and Krose [123]. This work looks at using Dynamic Bayes Networks to automatically determine a set of interconnected graphs that model the transitions of individuals between cameras in a system. The transition probabilities between cameras can be determined automatically using the observed appearance of individual objects through the system. The appearance of an object is modelled by three strips relating to the upper, middle and lower appearance colours, with the very top and bottom of the object discarded. These three appearance strips, each consisting of 25% of the object height, can be compared to identify individuals; however this work is based more significantly around the geolocation and camera view transition information throughout a set of corridors. This information can be very useful for corridor areas where transitions between camera views can be limited to a small number, but are not so useful in broader areas where transitions between views may not be so limited.

Gandhi and Trivedi [38, 39] present a colour representation method which stores a representation of an individuals colours based upon their spatial position within a cylindrical representation of the individual. This cylindrical surface consists of a number of regions which are rectangular on their surface, with the colour of that region being represented as the average of the pixels within that region that are observed by each camera view. Thus the representation includes information from each camera view to create a full model of an observed person; however the size of the rectangular pieces is likely to determine the ability to represent varia-

tions in clothing or appearance which may be of multiple hues. Thus significant colour information loss may occur with strongly patterned clothing. The represent is also likely to be very sensitive to the alignment of the cylindrical representation. The main concern with such an approach would be the automatic registration of an individual into the system to begin the matching process where most of the regions in the surveillance system consists of only single camera views.

The final approach considered here is by Yang *et al.* [119] who present a new human appearance model aimed at tracking across video sequences. This model is based upon the usage of a brightness colour feature, an RGB rank feature, and a path length of those features from the top of the head of the individual person. This aims to incorporate spatial information with the illumination invariant colour features to become more invariant to pose without requiring a three-dimensional model with articulated joints. They also propose to limit the track comparisons required by selecting key frames for matching between tracks, and match the appearance models using the Kullback-Leibler distance. The results presented in this paper clearly demonstrate that the RGB colours are so affected by illumination that illumination mitigation or illumination invariance is required for colour appearance features. Using RGB rank is clearly shown to be more effective than their method of separating colour and brightness, with both methods being tolerant of downscaling the original image. These results present a very recently developed promising appearance model that is tolerant to many factors including pose and illumination. The results presented do not necessarily present the full picture of this method though, as only the true matching rates are reported and analysed across only two different cameras. These matching results are based upon minimising the overall error rate; however this total error rate is not reported. This article also does not report upon whether their spatial colour representation is able to perform accurate partial matches, such as where a portion of the object is occluded. Such accuracy of partial matching is important for cases such as their subway scenario, where people are often only partially viewable.

These twelve systems have put together many components to build models which can be used to attempt to identify, or match people across disjoint cameras. These techniques use a variety of shape, appearance, camera transition time estimation, and localisation techniques. The combination of these matched tracks can allow for individuals to be tracked across a whole surveillance system. Some of these techniques rely heavily upon path probabilities, or camera transition time estimates between camera views to increase the accuracy of the system by limiting the possible transitions between cameras, though the limitations of these techniques are often not widely discussed due to the simulated surveillance systems.

Matching tracks can improve the information available if applied with realistic assumptions that work for abnormal cases in real systems. It has to be carefully applied to ensure that all possible transitions are allowed because often the cases of abnormal movement occur for the individuals of real interest, such as where they stop to steal an object. The free nature of human motion makes the transition times between largely spaced cameras unreliable at best, and possibly even misleading. The appearance of an individual is also widely used because although appearance is not biometric in nature, it is very visible and tends to remain invariant within a surveillance session, as people do not often change their clothing within most surveillance environments. Where such changes do occur manual reconciliation of the tracks by a human operator may be required, although humans also find such appearance changes difficult. The major difficulties are how to effectively incorporate the spatial component of the appearance features, and how to mitigate or make the appearance invariant to illumination. Shape features have also been proposed with height and gait both being used for humans, as well as width and length for cars. Shape features for humans are often applied with very limiting assumptions or conditions, such as utilising single cameras in a single location to observe people over long periods of time; however they still promise to add significant information to that provided by appearance.

Of the small number of surveillance companies developing commercial software in this area, most address the area by analysing individual cameras for information that is sent for combination in a central location. Their focus is largely upon providing accurate information to the operators with a minimum of false alarms. Thus their software tends to lag the cutting edge research until their results have been proven to be very reliable under a wide range of conditions. Thus even the major companies with commercial products in this area, such as Object Video and IBM are yet to begin addressing tracking across mutiple cameras, especially where there may be gaps between their views.

2.7 Literature Summary

The literature has shown that there is a wide range of research into video surveillance. This research has focused upon many of the underlying technologies, as well as describing some of the current approaches to wide area video surveillance. Though there is still opportunities to improve many of the underlying technologies, they are adequate enough to build upon for further research into subsequent tasks such as tracking, object analysis, or content based image retrieval. This advance in computer vision techniques has also been helped by advances in camera

resolutions and reductions in cameras prices to expand upon the traditional focus on single or multiple overlapping cameras. This area of disjoint camera tracking or track matching is a recent area of study that has few publications yet, as the motion features that form the basis of most of the current tracking literature are often unreliable. This makes research into other possible shape and appearance features necessary to provide an adequate solution to the problem to overcome some of the limitations of path probabilities for camera transitions. With the increasing focus upon terrorist activities it seems that intelligent building surveillance systems developed for realistic environments would be beneficial for increased security. Commercial products to date have focussed upon extracting accurate useful information from individual cameras, but are yet to adequately combine much of this information across cameras due to the limited accuracy of research results at this stage. Tracking across cameras is still important, even where real-time applications are unachievable, as the most significant usage of surveillance systems is probably for post event analysis. Such analysis involves video reviewing after the 2005 London bombings, where even if the results are only accurate under a semi-automated approach, they can still save considerable time and manpower for video analysis.

The literature has presented many features to model the shape and appearance of humans and other possible objects of interest. The most common features used for humans have been colour appearance, height, gait and facial features. Each of these features has its limitations in identification accuracy and long term stability, with the best results being obtained when multiple features are fused together. The current limitations in appearance features occur in their limited invariance to illumination changes, their often limited spatial information, and the inherent lack of discrimination when people are wearing similar colours. Height features are often not available with a high degree of accuracy throughout the majority of surveillance cameras, where there are only a single camera views available. Gait features have not been reliably tested for their accuracy across wide surveillance systems with multiple floor surfaces and often limited resolution. Facial features have provided high levels of accuracy for small facial databases; however results for large databases are not so promising, especially when obtained at low resolutions. Addressing these limitations in object features and investigating combining them would therefore provide a significant contribution to literature.

3 Colour-based Robust Appearance Features

This chapter explores the usage of colour-based appearance features for tracking, matching, and identification of individual objects throughout a surveillance system. Appearance features have often been utilised in surveillance to distinguish differing objects of interest [47]. Colour features have been the main focus of appearance in video surveillance because of the low level of object resolution to distinguish other possible features, such as textures. This set of colour features within the object has to be differentiated from the use of colour to segment an object from the background, as it is the colours themselves that are important, not their level of contrast from a background model. An important consideration for the usage and storage of colour features is the representation that is used to store that information. This is not trivial as in the case of a single value for representing a height estimate, as a range of presentation possibilities are available. These range from directly using the R, G, and B histograms, to the histograms from other colour spaces such as HSV, as described in section 2.2, to representative colour clusters. These can be obtained from either key frames, or from the entire track of the object within a given camera view. The choice of colour representation will also affect how they can be compared to determine the appearance similarity of objects. A general discussion of probabilistic distance measures that can be applied to colour features is provided in section 2.3 of the literature review.

This chapter first presents an overview of the background information to appearance features, including discussing many of the assumptions that are commonly applied to the representation of appearance and the limitations of using appearance features. Section 3.2 then describes a technique to extract a compact colour representation that stores pixel level information in the full three-dimensional colour spaces into its Major Colour Representation (MCR). Section 3.2.1 explores how the MCR features can be optimised using an online k-means algorithm. Section 3.3 explores how these MCR features can be integrated along a time sequence into MCRs to improve its robustness to gait effects and small segmentation errors. Section 3.4 shows how these features can be compared to determine a level of similarity between two MCR appearance features, with Section 3.4.1 expanding this comparison to be integrated along the tracks of the two objects through time. Although this method can be used to compare the MCR of entire objects, Section 3.5 explores how spatial regions of the object can be used to extract spatial MCR features. These can be used to represent spatial appearance components, such as clothing colour. The results of the comparison of MCR features for a set of people are explored in Section 3.6. The initial results on the

global MCRs of manually segmented individuals are given in Section 3.6.1. The results of the fully automated system are then given in Section 3.6.2, where the global MCRs are compared with the Upper and Lower MCRs. The discussion of these results is then outlined in Section 3.7. The chapter then concludes with a summary of the MCR extraction and comparison process, along with a discussion of possible future research directions.

3.1 Appearance Feature Background

The two major problems with the colour features of an object are that they can be altered under differing illumination conditions, and they are not truly biometric for particular objects, such as humans, where colour appearance often changes. This is because people often change clothes depending upon their activities, unlike other objects such as vehicles. Even features such as hair colour are relatively easily changed when compared to true biometric features such as fingerprints. In fact it is this ability to change the colour appearance of clothing that can make it good for differentiating individual people due to the large possible variations.

Although humans can change clothing colours, assumptions can be made within each surveillance environment about the likelihood that people will change their clothing. These assumptions are based upon the surveillance area because although people will often change their appearance dramatically in their home environment, often through large changes in clothing, they are unlikely to change their clothing within a work environment. This thesis is therefore based upon the assumption that an individual is unlikely to change their clothing within a surveillance session. The surveillance session is defined as the period of time between when an individual enters the surveillance system through one of the possibly entrances, until the person has finished their business within the surveillance area and they exit the system through one of the possible exits. When this surveillance session is applied to the working day, within a typical working environment, then people are unlikely to change their clothing. A discussion about the violation of this assumption occurs in section 3.7 at the end of this chapter.

Colour features are second only to location in their application within the research aimed at probabilistic data association, or tracking across disjoint cameras; however little work has currently been applied to mitigate the effects of errors, such as illumination or the impact of segmentation errors upon these features in this context. The most common approach is to model the histogram of colours within an object tracked in one camera, to compare to those observed in another camera [19, 53]. Although the intrinsic colour of an individual is assumed to be

constant, illumination factors can shift the appearance observed in the two differenct cameras. The mitigation of these illumination influences is explored in Chapter 4. The impact of segmentation errors is also a consideration as the incorrect inclusion or exclusion of parts of an individual can affect the proportional amount of colours identified as part of the object. In fact this research has found that the impact of large segmentation errors upon spatial colour features, such as clothing colours, and even global colour features can be significant enough to dramatically change these features. Chapter 5 describes in detail the technique that has been developed to use these changes to identify segmentation errors and remove them from a robust appearance representation. This chapter instead focuses upon extracting a colour representation feature that can be used to store and compare the colours of a given object or region.

Much research has been conducted into the best way to represent colour appearance features for comparison. Colour spaces and transfer functions between them are described in Section 2.2, whilst this section focuses more upon the representations. Each of colour space has its own advantages and limitations [110], with the most common space being the three channel RGB space.RGB allows for over 16 million possible colour combinations in a 3-D space, which can be very cumbersome and time consuming to work within. Traditionally the three colour channels are therefore considered separately as 256 colour spaces; however this technique completely disregards the coordination that naturally exists between the colour channels to truly represent the colours. Other colour spaces, such as HSV, have been introduced as a pixel based representation to incorporate a degree of invariance to illumination changes, whilst retaining chromatic information. Colour clustering techniques have been proposed as a method to reduce the large 3-D colour space into a small number of key representative colours. Unfortunately adjusting the size and spread of these colour clusters can be very time consuming; however with a compact representation of the colours into a few key clusters makes comparing those colours can become faster and easier. The Principal Colour Representation (PCR) is a colour clustering technique developed by Li *et al.* [63] that uses fixed cluster sizes rather than fixed cluster numbers to improve speed and accuracy. This cluster based representation reduces the search space by only storing colour values which have significance to the current image, or the most common colours. It also allows for a somewhat reduced cluster set for performing comparisons between colour feature.

A final critical aspect of appearance features are the proportions or regions of appearance that they represent. Previous research within the surveillance field has focussed upon the broad object histogram, which consists of all of the colours

observable within an object [19], or specific appearance templates that show pixel level colour occurrence [107]. These two approaches include either no spatial information, reducing the information available, or have high spatial information such that the articulation of humans creates problems with matching the templates. More recent research has started to look at the spatial relationship of colour within an object without explicitly using templates. Gandhi *et al.* [38, 39] proposed a cylindrical representation consisting of regular blocks of average colour. Although it is reliant upon multiple overlapping cameras and accurate object orientation alignment, it creates a representation that inherently includes the spatial relationship of the object's colours. Such interest in the spatial location of colours allows for a greater analysis of an individuals appearance. Spatial colours, such as the colour strips used by Zajdel and Krose [123] have also been proposed to distinguish colours within specific regions of objects. Yang *et al.* [119] propose a path length coding of pixel colours to retain the information of the distance of a colour in pixels from the top of the head. If extracted correctly, then this spatial information can be used to distinguish improved appearance features that can spatially relate to aspects such as clothing colours. Further representations like spatiograms [9, 83] have also been proposed to include spatial information within colour. Figure 1 shows an overview of these current approaches to spatial colour information.

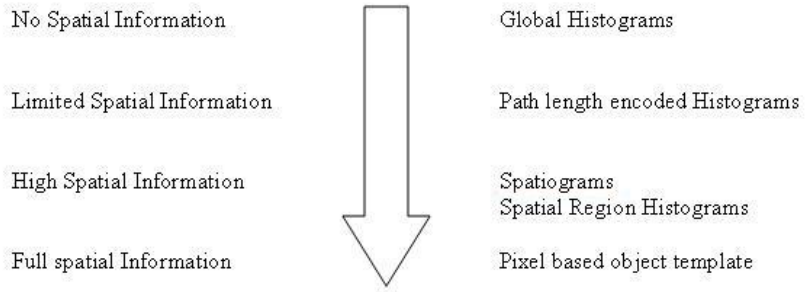


Figure 1: Approaches incorporating spatial colour information

This literature demonstrates that both the appearance representation and the spatial region which the colour represents are important to extracting useful appearance features. The representation needs to be a trade off between accuracy, compactness, the ability to describe colour variations, and the computational complexity. It also needs to have an appropriate method for comparing the similarity

between representations in order to allow for the comparison of appearance features. Other methods such as compensating for illumination can be incorporated to a degree into the colour representations through the usage of normalised colour distances; however this step is best preformed upon the object before the appearance representation is calculated. Colour appearance features also tend to be robust to a degree of segmentation errors and object pose changes through their usage of large colour regions.

3.2 MCR Colour Feature Extraction

This section describes the technique used to determine the Major Colour Representation of a given object or region within an image. This representation is based upon the same principles as the Principal Colour Representation (PCR) proposed by Li *et al.* [63] for tracking people using their colours. It aims to accurately model the most common, or major, colours of an object within a 3D colour space using a compact representation. This ensures that it retains all the colour information about the combination of the three colour channels at a pixel level together. It is also more compact than histograms as it only stores colours which represent actual pixels rather than the entire colour space. This reduced representation requires colour clustering, although the computational cost is reduced by minimising the parameters to optimise through using a fixed cluster size and allowing for an arbitrary number of clusters. This allows the representation to work similar to a Kernel Density Estimator for probability density functions, such that it can easily adapt to represent any spread of colour information, like the difference between plain colour, or patterned shirts. The representation itself does not utilise any spatial object information, but it can if it is applied on a spatial region, as described in section 3.5. This MCR feature is not inherently robust to illumination changes, but its invariance can be improved through either utilising a low number of very large clusters, although this reduces the amount of colour variation stored in the MCR, or through the application of illumination mitigation, as detailed in chapter 4, as a pre-processing step to extracting the object appearance.

The concept of clustering colours is dependent upon choosing a threshold of colour distance within which to group similar colours. From the many possible distance measures, this work uses a normalised geometric distance in the RGB colour space. This distance is normalised in the Euclidean distance between two RGB colours is divided by the sum of their magnitudes. This allows for the equal comparison of the greater perceivable differences in high illumination with the reduced perception of colour differences under low illumination levels. This choice

is similar to the colour distance developed by Li *et al.* [63], which was shown to be robust to some illumination changes and noise. This normalised geometric colour distance between any two colour pixels can be defined as:

$$d(C_1, C_2) = \frac{|C_1 - C_2|}{|C_1| + |C_2|} = \frac{\sqrt{(R_1 - R_2)^2 + (G_1 - G_2)^2 + (B_1 - B_2)^2}}{\sqrt{(R_1^2 + G_1^2 + B_1^2)} + \sqrt{(R_2^2 + G_2^2 + B_2^2)}} \quad (30)$$

where C_1 and C_2 represent the colour vectors for the two RGB pixels.

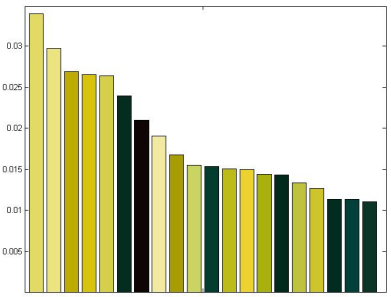
Given this colour distance, it is possible to cluster the colours of a segmented object which are close without losing any significant accuracy in representing its appearance. Several colour clustering methods are available from the literature [69, 80, 85, 100, 104, 124, 128]; however the ability to accurately represent a wide variety of possible colour and colour variations should be maximised. In [85], a method for clustering colours of moving objects was proposed based on a mixture of Gaussians. Each Gaussian component in the mixture is associated with a cluster and the number, relative weights, means and covariances of the Gaussian components are optimised with an Expectation-Maximisation algorithm. This leads to lower computational complexity to achieve an accurate and compact set of colour clusters. The proposed colour clustering process aims to minimise this computational complexity by reducing the number of parameters to be optimised. This is achieved by allowing a variable number of simple spherical clusters, which all have the same radius under the normalised distance given in equation **(30)**. Thus the parameters to be optimised are reduced to the mean location of the cluster and the weighting of the cluster, or number of pixels associated with that cluster.

This fixed cluster size reduces the parameters to be optimised and provides an accurate representation even in the frequent case of data that do not clearly separate into a fixed number of large clusters. As opposed to other methods such as [63, 85] the MCR technique does not look for a set number of clusters, but rather uses as many clusters as necessary to represent 90-95% of the pixels associated with the object or object region. This allows for a large set of clusters to represent complex objects with multiple colours of varying shades, whilst still providing a very compact notation when colours are uniform or have little variation. Experiments upon individuals observed in surveillance cameras have shown that this final 5-10% of less significant colour clusters actually consists of up to 70% of the colour clusters extracted. Thus they can be removed to leave only the major colour clusters, dramatically limiting the amount of clusters without significantly impacting upon the accuracy of the object appearance representation.

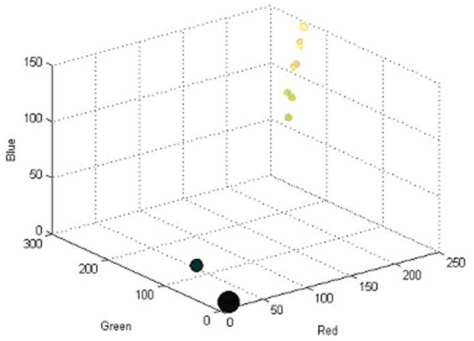
The initial colour clustering step combines pixels whose colours are within a given threshold in the normalised colour space. This is similar to Li *et al.* [63], except that the MCR process considers that all pixels equally corresponding to the appearance of the object, so this process does not use any pixel based weighting process. The process proceeds by scanning the object's pixels in row-major order. As the first pixel appears, its colour is set as the centre of the first cluster. If each following pixel is within a threshold under the normalised RGB distance from an existing clusters centre, the pixel count for that cluster is increased by one; otherwise, a new cluster is created, centred on that pixel. The size of this distance threshold is critical to size of the MCR and the amount of colour shade variations that can be captured. In the normalised colour space, this clustering is equivalent to having clusters with a common radius, spaced where the object's pixels are located. In the RGB colour space, clusters can be denser at lower magnitudes where the ability to perceive intrinsic differences in colour are reduced. The accuracy of the initial major colour clusters is also improved by using an online k-means algorithm to optimise the initial colours. This expectation maximisation technique iteratively alternates between membership calculation and centroid adjustment to improve the colour representation as described in full in section 3.2.1.

Figure 2 shows a picture of a flower containing several tones of yellow and green, as well as the MCR outcome of this first step. The original image is depicted in Figure 2a and has 115,537 different colours in a 150 x 113 pixel image. These colours can be clustered using the MCR process with a cluster threshold of 0.01 in the normalised distance measure to extract 839 clusters. Figure 2b displays the twenty most common colour clusters within the MCR as coloured bars. The height of the bar is proportional to the clusters pixel association count, given as the percentage of overall colour pixels. Figure 2c shows the ten colour clusters with highest count in the MCR, which are displayed by small coloured spheres with their size proportional to the colours count. The MCR for this image is further simplified by only storing the most common, or major clusters representing 90% of the pixels for later use. This reduces the number of clusters from 839 to 297 whilst still representing the majority of pixels in the image as can be shown by the reprojection of these colour clusters in Figure 2d. This reprojected image substitutes the cluster colour for the true image colour, with the white areas occuring for those pixels in the 542 less common clusters which are removed from the tail of the MCR. Clearly the MCR is capable of providing a compact representation that is flexible enough to to allow for accurate storage of varying shades of colours, as well as the dramatically different colours.

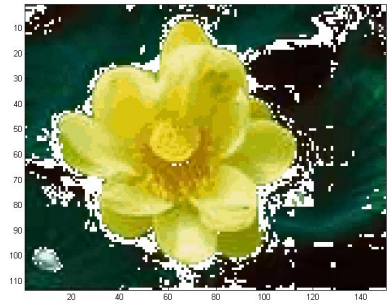
a) Original ‘tn_flower’ image



b)Initial top 20 MCR clusters



c) Top 10 MCR clusters with size proportional to cluster significance



d) Reprojection of MCR clusters to recreate the image

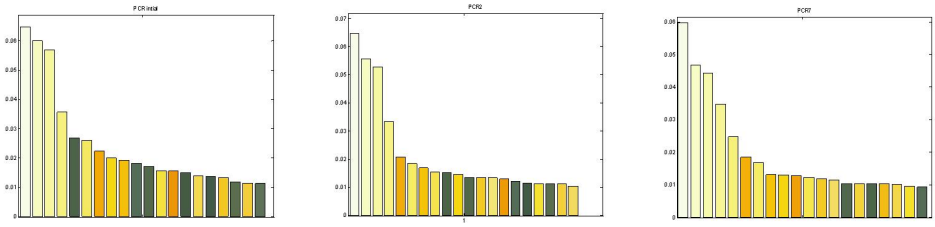
Figure 2: Major Colour Representation of ‘tn_flower’

3.2.1 Optimising MCR Using an Online k-means Algorithm

The extraction of initial colour cluster centres utilises a reasonably simple initial cluster creation procedure. Thus the cluster centres may be significantly displaced with respect to the clusters centroid i.e. the average position of its member pixels. In initial experiments it was found that this may affect the comparisons between object representations. This demonstrates that using a technique such as the *k*-means algorithm to refine the clusters of centroids, such as that proposed by Lloyd [68], can have a significant impact upon the accuracy of the final MCR. The *k*-means algorithm is an Expectation-Maximisation technique iteratively alternating membership calculation and centroid adjustment. Such algorithms are notoriously sensitive to the initial choice of parameters as they converge to local optima; however the usage of the initial clustering step allows the *k*-means algo-

rithm to start from reasonable initial values. Thus the local optima generally prove to be an adequate solution.

Figure 3: Original ore gold rose image (left) and reprojection of the 90% most frequent pixel clusters after 7 k-means iterations (right)



(a) Iter 0 (b) Iter 2 (c) Iter 7

Figure 4: MCR changes for 20 most significant colours with iterations of the k-means optimisation

The online k-means major colour clustering algorithm works by scanning the object pixels in row major order. For the current pixel, the closest cluster centre is computed and the pixel assigned to it. Then, the centre of this cluster is updated as:

$$R_c(i) = w(i)R(i) + (1 - w(i))R_c(i-1) \tag{31}$$
$$G_c(i) = w(i)G(i) + (1 - w(i))G_c(i-1) \tag{32}$$
$$B_c(i) = w(i)B(i) + (1 - w(i))B_c(i-1) \tag{33}$$

where $R(i), G(i), B(i)$ are the RGB components of the ith (current) pixel, $R_c(i)$, $G_c(i)$, $B_c(i)$ are those of the cluster's centre after the ith pixel has been processed, and $w(i) = 1/n$ the current weighting coefficient. Here n is the current number of pixels in the colour cluster.

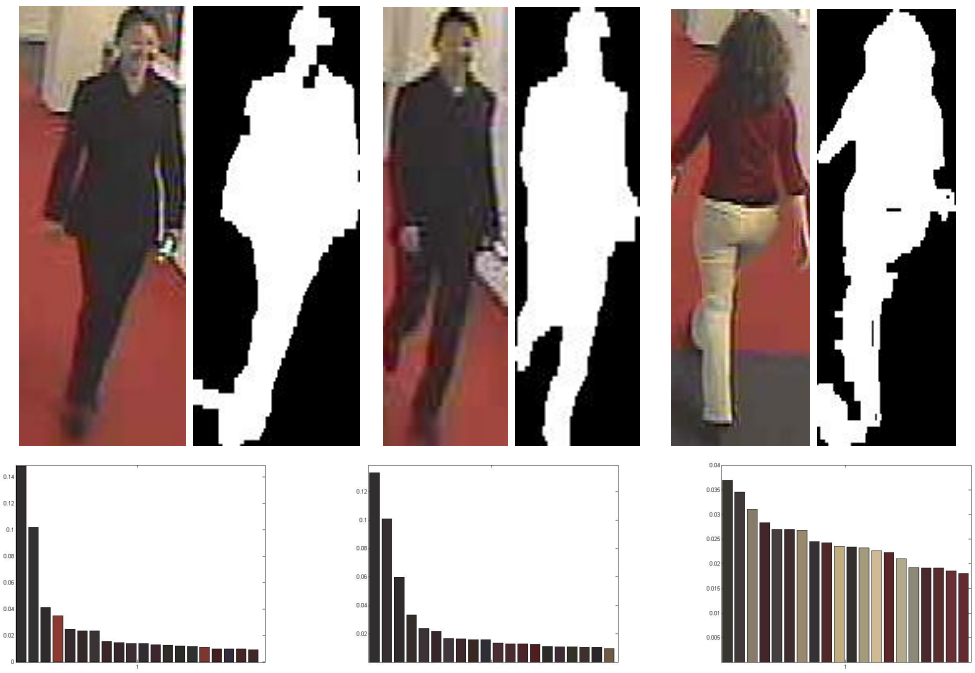
It can be seen that with the increase in the number of pixels falling into a cluster, the weighting coefficient decreases. Thus changes in the centroid position tend to gradually slow down as the scanning of the pixels progresses. Since cluster centres are moving, iterations are necessary until all pixel assignments and cluster centres stabilise. In these experiments, between 80 and 90% of pixels are usually already member of their final cluster after the first iteration.

Figure 3 shows the picture of an Ore Gold rose captured at a resolution of 480 x 322 pixels, which is very rich in tones and shades, as well as a reproject of the MCR representing the 90% most common colour clusters. This reprojection again demonstrates the ability of this clustering process to represent the major colours and even a degree of the subtle nuances in those colours. Figure 4 shows the 20 most significant colour clusters in its initial MCR, as well as those top 20 clusters of the MCR calculated at 2 iterations and 7 iterations of the described online k-means clustering algorithm. It shows that no major improvement was made by increasing the number of iterations from two (b) to seven (c), yet the increase in computation time is significant, especially for larger images. This would suggest that 2 iterations is sufficient even where there are a variety of tonal variations in the image.

Figure 5 shows the MCR from objects automatically segmented from three different frames from a single camera. The similarity between the MCR for frames 775 and 1297 demonstrates the ability of this representation to capture the dominant colours, which also appear similar in the frames. The MCR is also clearly distinct in frame 997 where a different person is observed in the same area.

3.3 Improving Robustness Using Incremental MCRs

After computing an objects MCR for each frame in its track, these frame based MCRs can be integrated over the window of the last k frames. This window of frames aims to make the MCR more robust to any errors that might occur in a given frame, including its invariance to pose changes that might occur during a gait period. This is achieved where k is chosen to be a small value of frames that is larger than half of the perceived gait period, as it allows for the full range of changes for a step within each window. Thus the optimum window size can differ for different camera speeds and gait periods; however in practise it is much



(a) Person A, frame 775 (b) Person A, frame 1297 (c) Person B, frame 997

Figure 5: MCR from three automatically detected people

more reliant upon camera speeds, as gait periods typically lie within a small range of often less than a second. This choice of k aims to keep the window short, yet provide maximum information about object's appearance under pose variation along the track. The data was obtained at approximately six frames per second, so k was chosen as 3. Indeed experimental results indicated that very marginal improvements were made with a larger window size, though the computational increase was also minor.

This augmented representation over a window of frames is denoted as the Incremental MCR (IMCR). It is based upon the merging of the frame based MCRs by combining similar colour clusters throughout the window together by merging their significance across the frames. This effectively combines the pixel associations to those clusters. Once combined, then a tail of clusters can be found that represent the last 5-10% of the minor colour clusters. This clusters generally do not occur very often across the window of frames and can be removed to reduce the IMCR size. This process of MCR integration can be formulated as follows:

Given the MCR of an object, A, at frame q represented as:

$$MCR(A_q) = C_1(A_q), C_2(A_q), \ldots, CM(Aq) \tag{34}$$

where $C_i, i = 1, 2, \ldots, M$ are the major colours centres and $p(A_q) = p_1(A_q), p_2(A_q), \ldots, p_{M_q}(A_q)$ their bin counts. The IMCR of the object at the q-th frame can be represented as:

$$IMCR(A_q) = \sum_{k=-(K-1)}^{0} MCR(A_{q+k}) \tag{35}$$

$$P_{IMCR}(A_q) = \sum_{k=-(K-1)}^{0} p(A_{q+k}) \tag{36}$$

The sign $\sum$ in equations **(35)** and **(36)** is used here to mean a special 'summation', i.e. the merging accumulation of the MCRs of frames $(q-(K-1)), \ldots, q$ based on the colour threshold. As the experiments reported in this thesis are all based upon cameras with a frame rate of approximately 6 frames per second, K was set to three in all of these experiments.

The combination of the IMCR representation with the illumination mitigation technique outlined in chapter 4, has proved robust to a number of factors that might introduce errors including minor segmentation errors, changes in shape through changes in pose, and even illumination variation that occur in typical surveillance scenarios. This technique therefore promises to provide an accurate yet compact appearance representation that is robust to a variety of the typical errors that occur.

3.4 Comparing MCR or IMCR Appearance Features

The main reason for extracting appearance features is to compare them to quantify the similarity of appearance between objects of interest. Within this section the term MCR is used to refer to either an MCR or IMCR as they both utilise the same representation, which can be of variable length, and indeed only differ in whether they were obtained from a single frame, or a small window of frames. This similarity in representation occurs as the same technique can be used to compare MCRs and IMCRs, and also allows them to be compared to each other if necessary. This allows the appearance within a single image to be compared to the appearance of a tracked object, or vice versa, allowing for a wider usefulness in forensic application as well as real-time surveillance purposes. For the purpose

of similarity, one could use a standard distribution distance such as the Kullback-Leibler divergence to compute the distance between the two MCRs and use its reciprocal as the similarity [127]. The similarity measure presented here is based upon the similarity measure derived by Li *et al.* [63], except that it does not allow the possibility of one to many cluster matching. This distinction is significant as it prevents groups of clusters from being counted multiple times in the similarity process, which could artificially boost the similarity measurement.

The MCR similarity measure presented is based on searching and comparing the most-similar colour clusters to determine the amount of similarity between the MCR of object A, MCR_A, and the MCR of object B, MCR_B. This is achieved by determining the percentage of overlap between colour cluster frequencies for matching clusters.

The process begins by assuming that there exist M major colour clusters in object A which can be represented as:

$$MCR(A) = C_{A_1}, C_{A_2}, \ldots, C_{A_i}, \ldots, C_{A_M} \tag{37}$$

with their cluster frequencies represented as:

$$p(A) = p(A_1), p(A_2), \ldots, p(A_i), \ldots, p(A_M) \tag{38}$$

Object B can be represented similarly over N major colours by the $MCR(B)$ and $p(B)$ vectors. In order to define the similarity between two objects, a subset of $MCR(B)$ is firstly defined as:

$$MCR'(B|C_{A_i}, \sigma) = \left\{C_{B'_1}, C_{B'_2}, \ldots, C_{B'_N}\right\} \tag{39}$$

where the distance between $C_{B'_j}$, $j = 1, 2, \ldots, N$ and C_{A_i} is less than a given threshold, σ.

This subset represents the colour clusters that are considered to be close enough to C_{A_i} to be potential matches. $C_{B_j|A_i}$ is defined as the most similar colour to C_{A_i} in subset $MCR(B)$ satisfying:

$$C_{B_j}|A_i : j = argmin_{k=1,\ldots,N} \left\{d(C_{B'_k}, C_{A_i}\right\} \tag{40}$$

Then the similarity of colours C_{A_i} and $C_{B_j}|A_i$ can be defined as:

$$Sim\left(C_{A_i}, C_{B_j}|A_i\right) = min\left\{p\left(A_i\right), p^{[A_i]}(B_j)\right\} \tag{41}$$

where $p^{[A_i]}\left(B_j\right)$ is the frequency of $C_{B_j|A_i}$. The min operator in equation **(43)** is used to retain the 'common part' of $p(A_i)$ and $p^{[A_i]}(B_j)$ as the similarity

between the two colours. It is possible to note that their 'different part', or absolute difference, $\left|p(A_i) - p^{[A_i]}(B_j)\right|$, is the well known Kolmogorov divergence under equal priors [127]. In this sense, the similarity measurement presented here is analogous to the complement of the Kolmogorov divergence as in:

$$Sim(C_{A_i}, C_{B_j}|A_i) = maxp(A_i, p^{[A_i]}(B_j - p(A_i) - p^{[A_i]}(B_j) \tag{42}$$

The similarity between the whole objects A and B, in the direction from A to B is then given by:

$$Sim(A, B) = \sum_{i=1}^{M} Sim\left(C_{A_i}, C_{B_j|A_i}\right) \tag{43}$$

The similarity between object A and object B in the direction from B to A, $Sim(B, A)$, is defined in a similar way. Note that $Sim(B, A)$ generally differs from $Sim(A, B)$ as for any given $C_{B_j|A_i}$ and $C_{A_k|B_j}$, $i \neq k$. That is to say that the closest colour cluster of Object B, B_j, for any given cluster of Object A, A_i, does not always ensure that the colour cluster A_i is the closest colour cluster in A to the colour cluster B_j. This property can generate asymmetric similarities depending upon the direction of calculation, which indicates that sometimes the colours in object B may in large part form a subset of object A. Such cases of asymmetric similarity measurements are not indicative of matching appearances, and hence need to be considered in determining the final similarity measurement of the MCRs. Thus deriving a symmetric similarity measurement first takes the minimum and maximum similarities between the two MCRs:

$$Simmin(A, B) = min\left\{Sim(A, B), Sim(B, A)\right\} \tag{44}$$

$$Simmax(A, B) = max\left\{Sim(A, B), Sim(B, A)\right\} \tag{45}$$

These values can be combined into a single final value, $Similarity(A, B)$ based upon their symmetry. Where $Simmin(A, B)$ is less than a given discrimination threshold, $\eta_{discrim}$, the similarity of objects A and B is defined as:

$$Similarity(A, B) = Simmin(A, B) \tag{46}$$

The rationale in this case is that $Sim(A, B)$ and $Sim(B, A)$ are either very asymmetric or both low. Hence it is appropriate to bound $Similarity(A, B)$ by their lowest value. Where $Simmin(A, B)$ is $\geq \eta_{discrim}$, the symmetry of the values can be incorporated as:

$$Similarity(A,B) = 1 - \frac{Simmax(A,B) - Simmin(A,B)}{Simmax(A,B) + Simmin(A,B)} \quad (47)$$

In this case, the two visual objects are likely to be the same physical one. As a further verification, the difference between the maximum and minimum similarities in a ratio form is used. In equation **(47)**, a large difference between the maximum and minimum similarity leads to a low similarity value. The definition of $Similarity(A,B)$ in equations **(46)** and **(47)** aims to prevent asymmetric, partial matches between two objects and allows for a final similarity threshold for matching assessment to be determined more easily. In practice, the measurements above are usually computed over IMCR values to increase the robustness of the similarity measurement to pose changes.

Whilst this similarity measurement is computed between any two appearance features, multiple appearance features are generated whilst the individuals are tracked through their camera views. Each of these features can be compared to determine the level of similarity between the frames. The availability of multiple comparisons allows for the extension of analysing these similarity values in time along the tracks of the two individuals.

3.4.1 Time Integration of Similarity

In order to evaluate the matching between the two tracks of objects A and B over a sequence of N frames, two basic alternatives are possible: (a) extending the object representation to cover whole track and performing a single, overall matching operation, or (b) repeatedly comparing pairs of IMCRs from the two tracks and integrating the results. The latter option is intrinsically more robust to segmentation errors which may occur occasionally at the frame level and could possibly pollute an overall track MCR. By using multiple matches integrated along the track, segmentation errors may pollute some of the frames and their corresponding similarity measurements; however the larger majority will remain unaffected, assuming that the majority are free from significant segmentation errors.

Two approaches are possible for this integration of similarity: fusing decisions or posteriors, as explored in Chapter 7, based upon statistics of the similarity values, or through fusion of the similarity values directly. Both of these approaches compare IMCR pairs in frame order along a sequence of N frame windows. This linearity of comparisons between frame pairs in frame order is aimed at keeping a linear computational complexity, $O(N)$, for the algorithm; however comparison of all IMCRs is possible at a higher complexity. Linear computational complexity

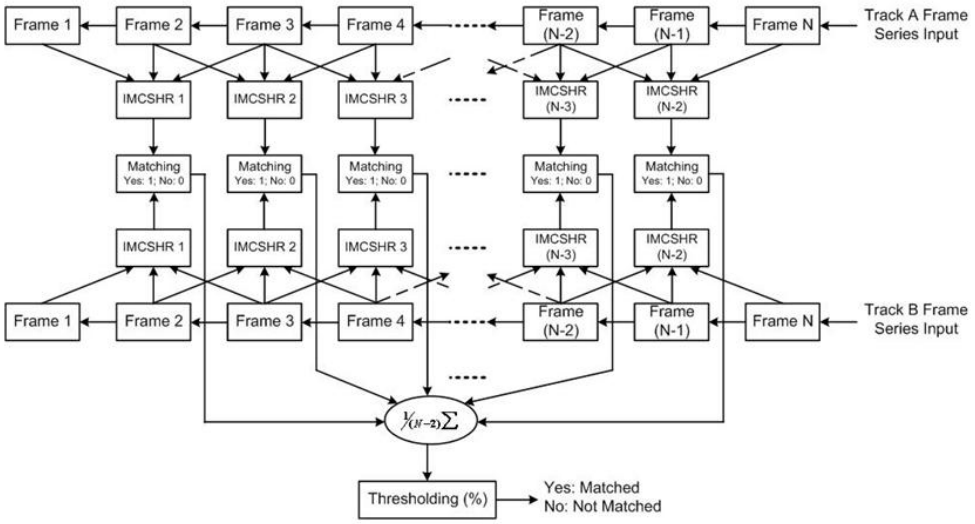


Figure 6: IMCR matching of two tracks using time integration

in the number of frames is considered the minimum reasonable complexity for matching over a frame sequence and allows the algorithm to meet real-time constraints. It also makes an *on-line* version of the post-matching integration possible as the surveillance application scenario implies that the two tracks cannot be acquired from a single individual at the same time. The main difference between the two approaches is that decision fusion assigns a value of 1 to those comparisons above a threshold, and a value of 0 when below the threshold. Two tracks can then be considered as matching when the percentage of matching IMCR pairs is above a second threshold. This process is shown in Figure 6.

A fusion of similarity values can be considered as a series of dependent classifiers. They are considered dependent as each classifier is performed upon same data at a different times, and thus no single classifier could be considered more accurate than another. Fumera and Roli [37] demonstrate that the optimum Bayesian fusion is thus obtained using a simple average of each of the classifiers. The similarity for MCR features between Track A and Track B, $S_{MCR}(A, B)$, gives a value between 0 and 1 to evaluate the level of track similarity, and is written as Equation 48.

$$S_{MCR}(A, B) = \frac{\sum_{i=1}^{N} Similarity_i(A, B)}{N} \tag{48}$$

Assuming one track has already been recorded in the system and the other is forming, matching can be stated as soon as N frame windows from the forming track become available; however subsequent frames can still be incorporated to improve the matching accuracy. The difference in output between the decision based time integration and the Bayesian based time integration is the end result being either a matching decision based upon a set thresholds, or providing a quantitative number to relay the amount of similarity. This time integration is explored further in Chapter 7, where the advantages of Bayesian fusion are explored in detail.

3.5 Extracting Spatial MCR Colour Features

Although the MCR feature is compact and flexible enough to represent all of the global colours of an object, it can also be used to capture spatial colour features of an object. Global colour appearance has been widely used in the literature [47]; however more recently colours are being incorporated with their spatial location to provide more information about an individual object of interest [39, 123]. Such methods aim to separately model components of the object's appearance to provide added information about those appearance components. When such spatial appearance features are applied appropriately, they can be used to model the appearance of object features, such as clothing colour or texture, hair colour, skin colour, or a variety of other possible appearance features. The main difficulty occurs with extracting these features adequately, as well as their usefulness in discriminating between people for matching purposes.

These spatial appearance features should be distinguished from other work aimed at illumination invariant colour features, such as CSIFT [1]. The spatial appearance features are based upon the colour information itself, rather than directing it to be colour information at identified 'key' geometric locations, like in the CSIFT approach. This is because the articulated movement of humans makes the stability of these 'key' geometric locations unreliable, leading to dramatically reduced accuracy when compared to using these same features on static or moving rigid objects [1], such as cars. Thus appropriate spatial regions need to be carefully considered for each class of objects of interest.

Although there are many possible spatial colour features that could be used for analysing individual humans, only a limited number of these can be accurately extracted and used to discriminate between individuals. The most obvious features are the upper and lower clothing colours, which are often a single intrinsic appearance, and thus can be pose independent in typical office building environ-

ments. Sometimes this clothing can be of a multi-coloured appearance, and can be pose dependent with colours that differ on the front and back; however this thesis does not address these concerns directly. A discussion of the impact of the non-uniform intrinsic clothing colour can be found at the end of the chapter in Section 3.7. Skin colour is also another widely used feature, although its main usage is for classification of object colour regions as skin regions, often for performing face identification. The usage of skin as a classification feature is based upon its limited chromatic variation, suggesting that it may not have enough variation to be useful in the identification of individuals. Hair colour has also been suggested as a feature [19] to provide an added degree of discrimination between individuals, especially those wearing similar clothing colours like uniforms. Although hair colour often remains stable for long periods of time, unless a person uses hair die to change it, the main difficulty occurs because it is observed on a small region on the upper extremities of a person, especially for frontal views. This makes it sensitive to even relatively minor segmentation errors. Footwear appearance is also a possibility, although it suffers from many of the similar segmentation problems that affect hair colour, as well as possible problems with shadows that may not be adequately removed close to the feet.

This section focuses upon two spatial MCR colour features, as well as the global colours. These two features aim to extract the upper and lower clothing colours of an individual, which are commonly used along with an estimate of height for police descriptions. These features tend to be robust to segmentation errors as they do not include the extremities, and cover large portions of the individual. They also tend to be uniform in colour within most business or office environments. The effects of illumination sources, as well as crumpled, creases, and the cut of the clothes tend to make the MCR a small number of similar colour clusters rather than a single intrinsic cluster. These features can be compared to the usage of a single global MCR, which uses all of the colour information without spatial information. In such a case the global MCR may not be able to distinguish between a person wearing a white shirt and black pants, versus a person wearing a black shirt and white pants.

The three colour features explored and compared within this section are:

1. The global MCR feature, which represents the colours of the whole segmented object without any spatial information.

2. The upper MCR feature, which represents the colour of the top portion of clothing. This corresponds to the region from $30 - 40\%$ of the person from the top of the object's bounding box as shown in Figure 7. This narrow band

was chosen to ensure that it avoids the inclusion of the head and hair of the object, as well as low necklines, but does not go so low to overlap with the leg area.

3. The lower MCR feature is aimed to represent the colour of the lower portion of clothing. This corresponds to the region from $65-80\%$ of the object from the top of the object's bounding box as shown in Figure 7. This narrow band avoids the very bottom of the object which can be prone to shadows, or artefacts where the feet touch the ground. It also tries to avoid overlapping with the belt or upper torso area of the person.

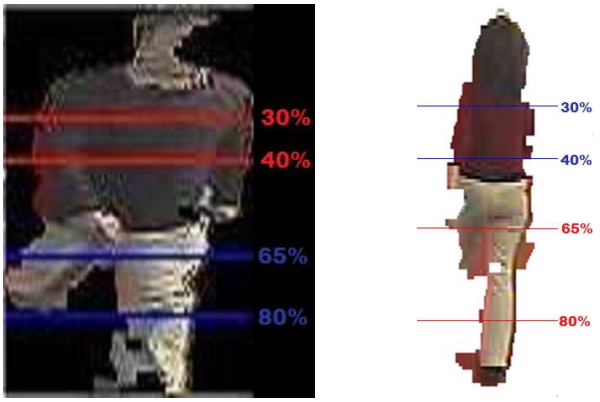


Figure 7: Examples of upper and lower regions of segmented individuals

The narrowness and positioning of both of the upper and lower MCR regions, as shown in Figure 7, also allow for them to remain constant under minor segmentation errors that will only have a minimal impact upon a person's appearance features. This choice was based upon an analysis of samples of various people observed under differing walking poses through different camera views to identify relatively stable colour regions under normal circumstances. The spatial regions identified are sensitive to large segmentation errors as they change the proportions of the object such that the bands will include regions which are not just the clothing colour. Where segmentation is known to be reliable, these bands could be expanded to include more information on the clothing colour, as it increases their sensitivity to segmentation errors, as discussed in Chapter 5. These spatial MCR features can improve the identification of differences between individuals, such as where people are wearing similar global colours due to matches between ones upper clothing with the others lower clothing. These spatial features may also be

useful to improve the MCR matching process not only across cameras, but also to correct tracking errors.

One obvious limitation is the size of the region, as small objects may not provide enough colour information for accurate MCRs. Often thresholds are used to remove foreground regions that are too small to be important. Where objects are less than 50 pixels high, they can be difficult to segment without significant errors and can only provide minimal object information. Where other frames observe the object in higher resolution, small frames should probably be discounted from the process as their probability of being erroneous is much higher.

Figure 7 shows both the upper MCR feature region (enclosed between the two top lines) and the lower MCR feature region (between the two bottom lines) for an object viewed in three different frames. These features include the main colour of the upper and lower clothing, but minimise the pollution of the feature with regions that are unrepresentative of the colours. The full removal of colour regions outside the clothing colour is difficult as it would require either the assumption of a single uniform colour, which could be problematic for even minimally patterned or multicoloured shirts, or it could be based upon estimating a full body model of the individual, which is computationally complex for highly articulated models such as people. For these reasons this research is based upon ratios defining the colour regions to statistically limit the inclusion of erroneous colours. This choice also allows for easy adaptation to the changing image size of objects viewed as they move around within the camera view.

3.6 Experimental Validation of MCR Appearance Features

This section outlines the experimental results achieved for the matching of MCR based appearance features. Early results presented in Section 3.6.1 are based upon a limited number of manually segmented tracks from individuals obtained across four cameras. These results show that the time integration of similarity using decision fusion could accurately match the same individual, whilst distinguishing between the differing individuals. These promising results have lead to larger experiments based upon automating the process, which are presented in Section 3.6.2. The results consist of an initial analysis of 26 automatically extracted tracks obtained from four individuals across two cameras with differing illumination conditions. This dataset is the same dataset used for the evaluation of results from the automatic height estimation in Chapter 6, and the fusion of features in Chapter 7. This dataset was specifically chosen to provide challenging cases where some of the upper or lower clothing were very similar in colour, although the overall

global colour set differs. The cameras used in these experiments are installed in the Faculty of Information Technology building at the University of Technology, Sydney, and were chosen for their differing illumination conditions. The cameras are operated daily for surveillance purposes by the University's security services and have not been installed or chosen to ease the performance of automated video surveillance tasks. This makes the footage obtained a good example of the technology used in existing surveillance systems.

The results of the initial manual and automated global MCR comparisons, even with the time integration of similarity are promising, but lack the accuracy to be used widely by itself. The development of the spatial MCR's to represent the upper and lower clothing colours, which were analysed across the same 26 tracks. The results obtained for these features show they are more useful than the global colours, but further work was required to increase the accuracy and robustness of these features. This has lead to many of the advances in the following chapters. These results would also suggest that higher resolution cameras, and higher frame rates would improve the overall results.

3.6.1 Colour Experiments on Manually Segmented Individuals

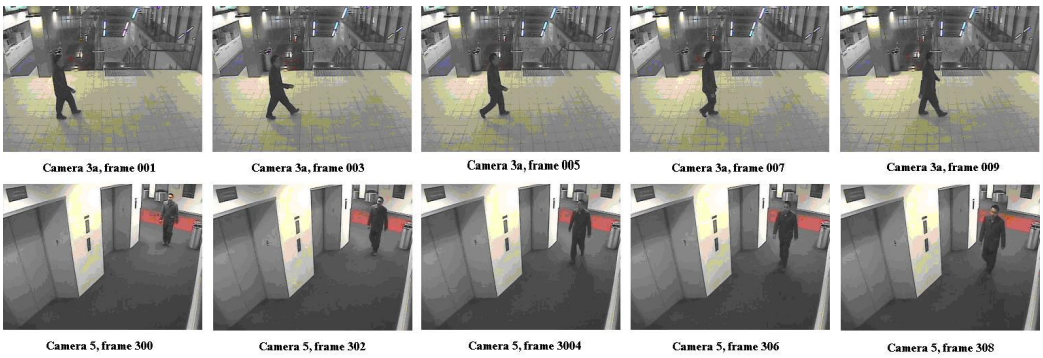


Figure 8: Same individuals observed in camera 3a and camera 5

The initial results were obtained from the analysis of manually segmented and tracked objects to determine the accuracy of MCR appearance features. These MCR results are based solely upon the extraction of the entire, or global, object histogram. They are based upon the time integration using decision fusion, though similar results are obtained when time integration is based upon Bayesian fusion. Three different sets of images and tables are provided. They first detail the case

Table 1: Results of IMCR Matching - same person

Test Case	Frame No.	Camera	Similarity	Matching Results
1	001-005	3a	0.9817	1 (Matching)
	300-304	5		
2	003-007	3a	0.9758	1 (Matching)
	302-306	5		
3	005-009	3a	0.9772	1 (Matching)
	304-308	5		
4	007-011	3a	0.9856	1 (Matching)
	306-310	5		
5	009-013	3a	0.9452	1 (Matching)
	308-312	5		
Integration	001-019	3a		100% (Match)
	300-318	5		

of a matching individual, then show the results where two non-matching individuals are observed. Finally a number of key cases are presented to outline the overall accuracy of the IMCR similarity method for matching and non-matching cases both within the same camera view, and across camera views where illumination may be expected to change significantly. The two cameras are significantly disjoint in both space and time, and the person's appearance in the two tracks could not be trivially matched. Moreover, illumination varies significantly with the object's position within each camera view. The results given in Table 1 show that the IMCR matching and post-matching integration are capable of coping with such variations in appearance, and the person is reliably matched even at very low frame rates.

Figure 8 shows sample frames from the two tracks of the same individual observed in two video surveillance cameras (camera 3a, frames 001-019, and camera 5, frames 300-318). These were analysed using five sets of overlapping five frame windows, with similarity values for each IMCR window considered matching if they exceeded the threshold of 0.8. If more than 80% of the IMCR windows are matching, then the two tracks are considered to be matching. The results given in Table 1 show that the IMCR matching and post-matching integration on manually segmented individuals is capable of coping with such variations in appearance.

Figure 9 shows sample frames from the two tracks of two differing individuals observed in the same two video surveillance cameras (camera 3a, frames 001-019, and camera 5, frames 010-018). These were also analysed using five sets of

Figure 9: Differing individuals observed in camera 3a and camera 5

overlapping five frame windows, with the same similarity threshold. In this case only 40% of the IMCR windows were determined to be matching, much less than the 80% threshold, thus the tracks are considered to be non-matching. The results given in Table 2 show that the IMCR matching and post-matching integration are capable of discriminating between individuals who have significant differences in appearances, even if portions of their clothing have similar appearances. It is also worth noting these results are successful even though no explicit illumination mitigation is used as this stage.

These two examples of the track matching process clearly indicate that the IMCR appearance comparison method can work across different cameras to match an individual and distinguish between two differing individuals. To give a broader indication of the accuracy of the system 5 manually segmented tracks were compared from the two individuals across four cameras to outline five possible test cases. The results of the integrated IMCR matching are given in Table 3 with samples of the frames used given as Figure 10. These indicate that the system can be useful for the typical cases of tracks viewed at different times in the same camera, or across differing cameras in order to match an individual, whilst discriminating between differing individuals.

The results presented in this section are based upon manually segmented objects from a limited number of tracks. They demonstrate the usefulness of this technique to perform an appearance based analysis to determine matching objects throughout a wider surveillance system. These results are limited in that they do not have to deal with the effects of poorly segmented objects, and they are subject to problems with varying illumination levels and the inability of the global histogram to reflect the spatial location of colours. These results therefore prompted

Table 2: Results of IMCR Matching - differing people

Test Case	Frames	Camera	Similarity	Matching Results
1	001-005	3a	0.3538	0 (No Match)
	010-014	5		
2	003-007	3a	0.7588	0 (No Match)
	012-016	5		
3	005-009	3a	0.7224	0 (No Match)
	014-018	5		
4	007-011	3a	0.8348	1 (Match)
	016-020	5		
5	009-013	3a	0.8075	1 (Match)
	018-022	5		
Integration	001-019	3a		40% (No match)
	010-022	5		

Figure 10: Typical frames used for test cases

a wider evaluation of MCR appearance features based upon automatically segmented and tracked objects that are analysed with spatial MCR's to compare with the global appearance information. It also suggested exploring techniques to mitigate the effect of illumination, as detailed in Chapter 4.

3.6.2 Colour Experiments on Automatically Obtained Tracks

The manually segmented MCR results demonstrated that this technique could be useful for comparing objects that were segmented and tracked perfectly. Automation of the entire process is required to evaluate the system in greater detail. This section presents two distinct sets of results for the automated system. The initial automated results were limited to just global MCR features. These results show that a global colour feature alone is useful; however more information about the

Table 3: Results of IMCR matching - differing people

Test Case	Frame No.	Cam	Typical Sim	Integrated Match
1 (Same object, time disjoint)	282-294 001-013	3_0 3a	0.9785	80% (Match)
2 (Same object, space disjoint)	001-013 300-312	3a 5	0.9817	100% (Match)
3 (Different objects, time and space disjoint)	050-062 010-022	4 5	0.3696	20% (No Match)
4 (Same object, time and space disjoint)	282-294 300-312	3_0 5	0.8410	100% (Match)
5 (Different objects, space disjoint)	050-062 010-022	4 5	0.3696	20% (No Match)

object is required to improve accuracy. The second set of results revises the same data; however it applies spatial MCR features along with the global MCR, as well using the controlled equalisation approach for mitigating illumination changes as described in Chapter 4. The used of spatial features also allows for the application of segmentation error removal to be applied as detailed in Chapter 5.

Figure 11: Typical backgrounds used for test cases

The first set of results is based upon an analysis of four individuals viewed across two cameras with differing illumination conditions. These cameras are part of the existing surveillance system at the University of Technology, Sydney, and are installed for normal security and surveillance services. The background view from these cameras is given in Figure 11, showing that the scenes are reasonably complex for automated surveillance; however are fairly typical of the

views that occur in typical surveillance systems. These views have differing background colours and varying illumination conditions both along the camera view, and between camera views, creating significant challenges for segmentation and the constancy of appearance.

Figure 12: Four people of interest (Person's A, B, C, D from left) and good automatically segmented masks (from frames 775, 1095, 1542, 2044)

The actual track data consists of four individuals of interest who are recorded across two camera views. Their clothing, shown in Figure 12 was selected to be typical to indoor environments and are not intended to be of high contrast to the background for easy segmentation. Indeed the segmentation shown is of fairly high accuracy as significant segmentation errors, including object fragmentation, does occur in some frames. The global appearance of these individuals were compared both within cameras, and across cameras to produce the results about six key cases given in Table 4.

The results given in Table 4 demonstrate that even though the original assumption of correct segmentation is broken, correct matching of individuals is high and discrimination between two individuals is largely maintained, even without ad-hoc tuning of the parameters. Particular cases, such as where Person Ds legs are not segmented correctly create false impressions of largely homogeneous dark colour. This can then be incorrectly matched with Person A, who is actually of a similar, but truly homogeneous dark colour. This case of segmentation error accounts for the majority of cases where two individuals are incorrectly matched both within the same camera, and across cameras, and suggests that the usage of separate clothing colour features could improve the final accuracy. The impact of segmentation errors is also noticeable in correctly matched objects, where a minority of windows are not matched. This shows the effectiveness of the integration of the IMCR process. It also indicates that occlusions of objects may lead

Table 4: Results of automated IMCR matching - 6 different cases

Test Case	Cases	Camera	Typical Sim	Match	NonMatch
1 (Same person, Same camera)	10	5_corridor or 5_lift	0.9436	10	0
2 (Same person, Disjoint camera)	13	5_corridor and 5_lift	0.8214	9	4
3 (Different people, Same camera)	8	5_corridor or 5_lift	0.3726	0	8
4 (Different people, Disjoint camera)	10	5_corridor and 5_lift	0.3913	4	6
5 (Person A in cluttered track)	1	5_corridor	0.9187	1	0
6 (Person B in cluttered track)	1	5_corridor	0.9408	1	0

to incorrect results, and need to be identified for removal from comparisons. The dramatically improved results within the same camera view compared to differing camera views are likely caused by the differing chromatic responses of the individual cameras, illumination changes, though these are already mitigated with controlled equalisation, segmentation errors, which tend to both more frequent and larger in the 5_lifts camera, and also possibly the pose or direction of travel.

Cluttered scenes also lead to significant segmentation errors with individuals incorrectly joined together and need to be identified as a source of possible errors. Two cases are shown in Figure 13 and reported in Table 4 as Cases 5 and 6. These cluttered scenes were correctly matched because the cluttering was transient and only polluted a small number of the frames within the track (four frames, or less than 20% within each case).

The second automated experiment expands upon these findings by applying extra upper and lower clothing colour as well as global colour to measure the object appearance features and applying major segmentation error removal. These extra features also allow for the automatic identification and removal of frames with large segmentation errors as detailed in Chapter 5. Due to the large numbers of segmentation errors in portions of the track data, 26 reasonably reliable tracks were used, which have fewer than 35% of their frames affected by segmentation errors. This demonstrates the complex nature of the scene and the difficulty of the segmentation process using the fast online-adaptive Guassian background model [116]. Although improved segmentation could be used, there are still sizable er-

Figure 13: Poor segmentation in two sample cluttered frames

rors in even the most complex techniques [96]. The segmentation removal technique, detailed in Chapter 5 removed over 80% of erroneously segmented frames, leaving significantly more robust appearance information from the tracked individuals. These carefully selected tracks were compared to each other in a pairwise fashion, giving over 300 possible comparison combinations. Of these, 60 comparison combinations are used as training data, with the remaining used for testing. Figure 12 shows that each of the four people is of a minimum 50% different colouring. The results from the testing data are presented as Reciever Operating Characteristic (ROC) curves in Figure 14 for each of the individual features.

The ROC curve is a graphical method that can show how the variation of the operation threshold affects the performance of the binary classifier [31]. In this case the curve compares the percentage of correctly matched individuals against the number of incorrectly matched individuals as the operating point of the system is adjusted. For a detailed understanding of ROC curves and their analysis see [31].

Figure 14 clearly demonstrates that the individual upper and lower MCR features are more accurate than the global MCR feature. This may be due to a number of factors including the inclusion of regions such as skin colours in the global colours, which have less variation than clothing colours, and are likely to increase the object similarity. The upper MCR provides greater accuracy than the lower MCR, likely due to the greater variation in upper clothing colours. Only two individuals are wearing similar upper clothing colours, the lower clothing colours of the four individuals are either black or white. Even with these similarities in intrinsic colour, the upper MCR produces matching at 72% with only 10% false alarms. The lower clothing colours reach 70% accuracy with 20% false alarms, but the global MCR feature only reaches 70% accuracy with 45% false alarms.

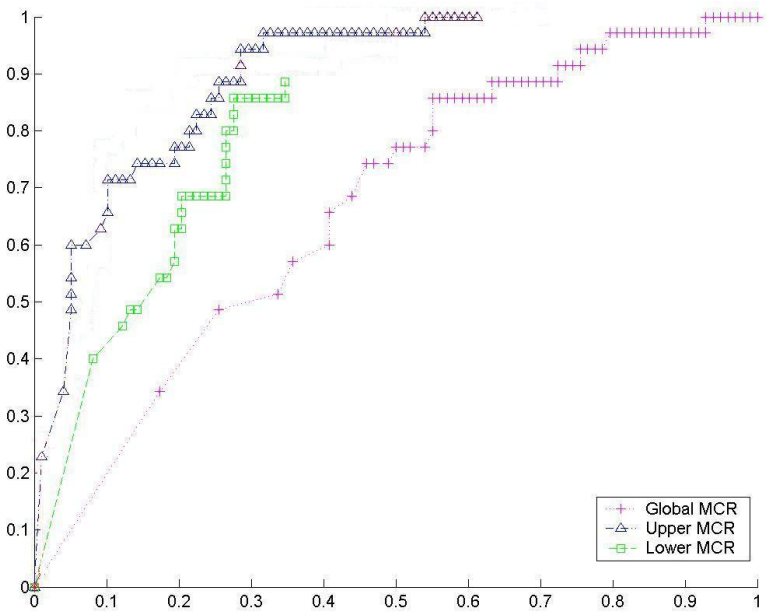


Figure 14: Accuracy of individual colour features

This shows the enhancement of colour appearance features through the incorporation of spatial information.

Each of these features show that the accuracy of matching individuals is only likely to be achieved when multiple features are fused together. The main problem with these features are the high rates of false alarms that occur for acceptable levels of matching accuracy. It is also important to note that track combinations obtained from four individuals across two cameras were chosen to represent key difficult cases, and thus provide a proof of the concept case. Larger experiments based upon more statistically significant data would be required to fully demonstrate and quantitatively analyse the effectiveness of this system.

3.7 Discussion of MCR Appearance Results

This work has investigated the implementation of appearance feature representations for the analysis of individual humans. The representation needs to be compact, allow for some colour variation and be robust to variety of influencing fac-

tors, primarily in illumination variations and segmentation errors. It also needs to provide for a method of determining the similarity between two representations. This thesis presents a method based upon the Principal Colour Representation (PCR) developed by Li *et al.* [63]. It expands this work by adding an optimisation step to improve its accuracy, and investigating the removal of non-major colour clusters, those clusters that have a low pixel association level. The representation is further improved by extracting MCRs across a window of frames. An incremental similarity process is used where objects are compared a number of times to reduce the impact of segmentation errors and pose variations. Finally the extraction of these MCR features are explored for spatial regions which are of particular interest for a given object. These relate to regions of clothing colours for the analysis of humans as objects of interest.

Numerous results are presented that show the MCR features achieve the ability to extract a compact representation of the colour appearance of an object. Experiments looking at the tone variation of flowers are given as Figures 2 and 3. They show through reprojection that MCR's can capture colour tone variations in a compact manner. This ability is likely to increase the accuracy of this representation over those which use a fixed number of clusters in a variety of circumstances where the number of fixed clusters may not match the object analysed. The retention of the major colours representing the top 90% of the pixels increases the compactness of the representation by removing up to 70% of the less significant colour clusters. This removal may also reduce the impact of minor segmenation errors as they are likely to relate to small clusters; however devising an experiment to quantify this impact is difficult. The usage of a normalised colour distance in the clustering process, equation **(30)**, is also significant as it provides a small degree of compensation for illumination changes. It allows for the increased perception at higher illumination levels which tend to amplify colour changes, which are suppressed at low illumination levels. This method also allows for explicit pixel based illumination mitigation, such as that detailed in Chapter 4, to be performed upon objects before the representation is calculated to improve the robustness to illumination changes, as it is not illumination invariant itself.

The results from the experiments conducted on automatically segmented individuals demonstrates that the global colours are a useful measure of appearance, but the spatial colours can provide improved accuracy. The accuracy of these individual features are not sufficiently high to be used alone, suggesting that feature fusion, is going to be important. This is also important as one of the major limitations on appearance features are their non-unique, non-biometric nature. Individuals can change clothes, but are unlikely to during a surveillance session;

however the likelihood of two differing individuals wearing similar clothes is non-negligible. Fortunately the manual review of falsely matched individuals is relatively easy for a human operator, when compared to performing manual correction of unmatched individuals by manually searching the video data.

Where uniforms are dominant, such appearance features will not distinguish individuals, except for those not in uniform. For most scenarios, these individuals are likely to be of higher interest. The spatial colour features are also likely to be more sensitive to minor differences in clothing than a global feature, due to their narrower focus. They could be extracted as simple strips as proposed by Zajdel and Krose [123]; however this work has looked at extracting more semantic information. The spatial colour features proposed relate directly to clothing colours upon the upper and lower body regions, as these two regions are often different in colour to each other, yet reasonably uniform within that region. The regions have also been chosen to reduce the impact of poor segmentation, minimise the inclusion of shadows, and minimise the influence of walking pose changes. Achieving such goals could become more difficult where simple arbitrary strips are used, rather than object class specific stable colour regions.

The most time consuming step of the track matching process is the extraction of the MCR features from each frame due to the complexity of colour clustering. Thus where the MCR features are already being used for track matching purposes around the surveillance system, utilising these features to enhance robustness would not add significantly to the computational complexity. These features could be compared to identify and remove segmentation errors, or possibly correct tracking errors. This suggestion of little added computational cost to utilise MCR features to improve the robustness of a number of surveillance components from segmentation to tracking provides additional usefulness.

3.8 Summary of MCR Appearance Features and Future Enhancements

This section summarises both the extraction, and comparison of MCR based appearance features. It outlines the steps involved in both the extraction of IMCR features from an image sequence, and the integrated comparison of those features between two tracks in simple terms. Possible extensions to these processes are also given in the form of future work that might be useful to improve the performance of the MCR based appearance features.

The extraction of IMCR's can be summarised as the following four step pro-

cess for an individual person tracked in a video sequence. First the individual needs to be segmented from the background. Then the initial MCR colour clustering can be performed for the global, upper and lower object regions. This clustering can then be optimised using the on-line k-means process. Finally the MCR's can be merged across a small window of frames equal to a minimum of half the gait period to add time integration and improve robustness to pose related changes.

The comparison of IMCR's can be summarised as the following five step process once the features have been extracted. It begins by finding the MCR cluster for Person B that is closest to each of the MCR clusters for Person A, within a distance threshold. Their significance can be added with that of other matching clusters in order to assess the similarity of Person B to Person A. This process is repeated to compare Person A to Person B. The overall similarity of the two individuals can then be assessed based upon the symmetry of Person A to Person B and vice versa. A threshold can then be used to generate decisions about the matching of individual MCRs or IMCRs. Such decisions can be merged to determine the overall track matchability. Alternatively the similarity can be statistically modelled from a training set as pdfs from the non-matching **(H0)** and matching **(H1)** classes. These probabilities can then be used to determine thresholds, or form the basis of Bayesian fusion to determine track level similarity.

These two processes form the basis of the extraction and comparison of a colour appearance representation. The results presented here show that the appearance features can provide significant discrimination and are likely to be very useful when fused with other features, as explored in Chapter 7. Further enhancements could be performed to identify other alternative appearance features that might be useful where clothing colour may not provide high discrimination. Such situations occur where uniforms are worn; however often regions of skin and sometimes even hair are left uncovered. An investigation of the variation of skin colour observed in surveillance quality footage would be needed to determine if such a skin feature might provide a useful level of discrimination between individuals. Skin regions also tend to change slowly over time due to sun exposure, and are likely to be semi-biometric in nature. Such skin regions have been proposed to identify faces throughout the literature, and could be used with the direction of travel to identify the pose, or walking direction, of an individual. Such information could lead to pose related appearance information, which may improve the accuracy of a wide surveillance system across areas such as University campuses where the clothing of an individual may be of differing colour or design on the front or back. The difficulty of this approach is the ability to relate together the

varying pose related appearance information of a single individual. The identification of a face region, or advanced shape analysis, could also lead to the accurate identification of a stable hair region. The appearance of hair often remains stable for long periods of time, and is likely to provide additional appearance information if a uniform is worn that does not include head wear. The main limitation of a hair feature is currently the poor reliability of the segmentation at the extremities of an object, where the hair occurs; however this feature could be used if the hair can be reliably separated from the rest of the object.

A significant extension of this method could look at the appropriate spatial appearance features for various classes of objects. Currently this project has focussed upon humans as the main object of interest as their highly articulated motion makes reliable features hard to determine. Other object classes, such as vehicles, robots, or even natural objects which might be of interest, such as trees or animals, are likely to have different stable appearance features that could possibly be determined by different spatial appearance features. These features are likely to be dependent upon the object class, where a range of such class based object appearance features could be useful in a variety of computer vision research areas ranging from video surveillance to robotics. Such an individualised class based feature set is unfortunately also likely to be rapidly expanding as further object classes are explored.

The aim of this research is to improve the accuracy of automating video surveillance using the existing surveillance systems. A number of general and clothing based appearance features have been explored, with other features like skin colour or hair colour being identified as obvious possible expansions of the appearance features. It is also important to consider that the accuracy of these features should improve through more robust scene modelling to improve the mitigation of illumination, and with accurate object segmentation techniques. It could be that such improvements may improve the system accuracy further than other advances identifying appearance regions like skin or hair.

4 Mitigating the Effects of Changes in Illumination

This chapter explores techniques that can be used to mitigate the effect of changes in illumination upon the appearance of a moving individual. Considering illumination effects is important because applications in the computer vision field that extract information about humans are often built upon the exploitation of appearance cues in videos. Colour based appearance features are increasingly being used due to the recent availability of cheaper, higher resolution cameras of good pixel quality; however, significant problems still affect the reliable use of appearance features for the analysis of humans in videos. The previous chapter explored the extraction of appearance representations, and identified variations in illumination and the articulated human geometry as the most significant challenges for appearance features. This chapter looks at practically improving the invariance of appearance features to illumination changes.

Colour invariance is needed because the colour appearance of an object in a camera view is not the intrinsic colour of the object itself, but rather a view-dependent measurement of the light reflected from an object, and the camera sensitivity to that light [33]. This problem is different from building local colour invariants such as those derived using CSIFT [1]. Such local colour descriptors describe the object's colour in a spatial neighbourhood, and are not actually colour invariant. Improvement in the colour invariance itself is actually the important component when looking at broader colour comparisons. This chapter focuses upon techniques can be applied and evaluated across objects viewed in differing illumination conditions to make similar objects more matchable, whilst still discriminating between differing objects. We refer to these techniques as 'illumination mitigation' techniques as they aim to substantially remove the effects of variable illumination on the object appearance.

Figure 15 shows examples of two such people of interest automatically segmented from the background, and how their red channel colour appearance may alter under differing illumination conditions and pose changes. This figure clearly shows the illumination problem that can occur with appearance features.

This chapter first describes the background of the illumination mitigation approaches including common approaches to colour invariance. Additional information on common colour spaces and their properties are provided earlier in the broad literature review in Section 2.2. Sections 4.2-4.4 outline the common techniques used to mitigate the effects of illumination through transformations that remove some of the effects of illumination that are compared in this chapter. These common techniques include illumination filtration, histogram stretching,

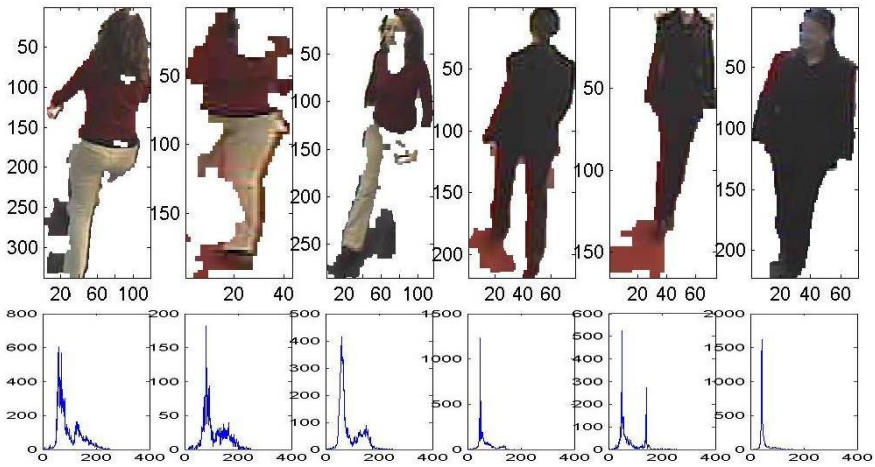


Figure 15: Sample people of interest and their red histograms under differing illumination conditions

histogram equalisation and were chosen to avoid the construction of a scene dependent illumination model. The novel techniques of controlled equalisation and a centralised version are also described and compared here to explore the enhancement of the equalisation approach. Section 4.5 outlines the process used to evaluate the mitigation of the illumination for surveillance images by comparing object similarities. Section 4.6 reports the experimental results of compared techniques, with a full discussion provided in section 4.7. The chapter then concludes with a final summary of the process and discusses future enhancements.

4.1 Illumination Mitigation Background

The colour appearance of any object observed in a camera is a view-dependent measurement of the light reflected from an object, and the camera sensitivity to that light [33]. This is given using an assumption of Lambertian surfaces. This suggests that although the colour is a useful feature for describing an individual person, the perceived colour of that person depends upon the scene illumination I, as well as the intrinsic object reflectance R. Indeed the camera sensitivity to particular wavelengths also needs to be considered where colours may be matched between different cameras. This section focuses upon techniques to mitigate illumination changes, with the literature on colour spaces provided in section 2.2, and research relating to colour representations provided earlier in Chapter 3.

In automated applications, segmentation of the individual is always affected by a certain degree of error, which can also add errors to the colour appearance, but currently can only be truly addressed through improved segmentation techniques, or the identification of seriously erroneous frames. Where large regions of colours are used, such as on the torso or legs areas of a person, the effects of segmentation errors are minimised.

Compensating for illumination changes are broadly classified by Finlayson *et al.* [33] into colour invariants and colour constancy. Colour invariants aim to apply transformations to the appearance that make the observed colours independent of the illumination of the scene. Colour constancy seeks to estimate the illumination of the scene to discount it from the appearance so as to extract the intrinsic colours of objects. Whilst accurate models of the illumination of the scene could extract the intrinsic colours of objects, the implementation of this technique is very difficult for complex scenes with multiple illumination sources that may also be time-varying. In previous work, Javed *et al.* [52, 53] propose to estimate the intensity transfer functions between camera pairs during an initial training phase. Such functions are estimated by displaying common targets to the two cameras under a significant range of illumination conditions, and modelling correspondences in the targets' colour histograms. However, the authors' assumptions in [52, 53] that objects are planar, radiance is diffuse and illumination is the same throughout the whole field of view do not hold in real life. Illumination varies at pixel-level resolution and such variations have first-order effects on appearance. In addition to this new lighting conditions that may occur over time would require identification so that a new training phase could be conducted to include these conditions. Weiss [115] proposed an effective method to estimate illumination from a sequence of frames of the same scene. Though the method works well for static objects such as the background scene, it is not designed for moving targets. Indeed these targets tended to be included into the illumination component as they are transient in the scene, as are the illumination changes. Even where they were identified as part of the image and not illumination, this technique is not designed to estimate the effects of the illumination changes over 3D moving targets, especially highly articulated ones such as people.

Approaches to colour invariance have had greater success in mitigating the effects of illumination, which Finlayson *et al.* [33] suggest occur because although the RGB values change, the rank ordering of the responses of each sensor is preserved. This implies that the values for a particular colour channel, such as R, will change from illumination source A to source B; however the ordering of those values will remain invariant, as shown in Figure 15. Recent work by Yang *et al.*

[119] also shows that the accuracy of matching the appearance of individuals can be improved by using a ranked RGB rather than simple RGB values. This observation of rank preservation has been demonstrated for what can be assumed as typical lighting in human environments, which largely consists of natural sunlight, fluorescent lighting, or incandescent lighting. Although other lighting sources are sometimes used, they are rarely used in open common spaces where surveillance occurs, so consideration of these sources is outside the scope of this investigation.

A range of techniques are used to provide colours that are invariant to illumination, or at least less dependent upon illumination, with the most common being chromaticity spaces. Chromaticity can be simply derived from the RGB space using the following transformation:

$$r = \frac{R}{R+G+B}, g = \frac{G}{R+G+B}, b = \frac{B}{R+G+B} \tag{49}$$

This chromaticity vector (*r,g,b*) has only two independent co-ordinates and is defined such that it is invariant to the intensity of an illumination source. Changes to the illumination will scale the RGB values by a factor *s* as (*sR,sG,sB*), leaving *r,g,b* invariant. If the illumination source changes in spectral output, say from a white fluorescent source to a yellow incandescent source, then a single scale factor is not sufficient to compensate for such a change. A second diagonal space has also been proposed where each sensor response in the R, G, or B spaces can be independently derived. This model allows for a shift in illumination intensity as well as a possible shift in the frequency spectrum for that illumination. The response could be modelled using the grey-world representation [4], which was described in Section 2.2.

These common techniques are useful for providing measurements that are somewhat invariant to illumination; however they have a degree of difficulty in adequately compensating for the multiple illumination sources that could also be time varying, in the case of natural sunlight. These multiple illumination sources also have complicated interplay with the complex 3-D surfaces of moving objects, such as humans, where the effect of illumination in the background, or portions of the background may vary significantly from its effect upon foreground objects. Chromaticity techniques also have difficulty in identifying the difference in intrinsic black and white surfaces, or differing shades of grey which may have similar chromatic values, but be distinct colours. For these reasons this work investigates various illumination mitigation techniques that transform the RGB data of the object to make the same object more similar under varying illumination conditions, whilst still allowing for the discrimination of differing colours without requiring

either training or other assumed scene knowledge. With the exception of illumination filtration, these techniques aim to maintain the rank ordering of the colours, but look to spread the object information across the entire channel bandwidth to reduce the impact of illumination. Filtration looks to adapt the colours in a manner that is dependent upon the brightness information of a pixel in a spatial manner.

Many methods have been proposed to mitigate the effects of illumination on colour appearance, or to extract colour related features that are invariant to illumination. Less research has looked to how to compare these techniques to identify which are the most useful for a given scenario. The research area of content based image retrieval has developed many techniques for quantitatively comparing histograms, such as the Kolmogorov divergence with equal priors [127]; however there is little work that has investigated the comparison of techniques for the mitigation of illumination effects on colours for surveillance specific tasks. This requires methods that can deal with low resolution images from poor quality sensors at frame rate speeds. This is perhaps due to the focus upon other features such as facial or shape information that are less affected by illumination. Essentially the basis of the problem is similar to that of comparing two histograms that are supposed to be either from the same object or differing ones. Thus an appropriate technique for quantitatively comparing the similarity of object appearance could be used after the illumination mitigation to compare the effects of the technique or techniques to an unmodified object. The nature of the observed object could also be chosen depending upon the scenario, such as focussing upon individual humans for a surveillance scenario.

This literature suggests that although much research has focussed upon the mitigation or invariance of colour based features, there has been little application of techniques to quantitatively evaluate these techniques. This is especially true for assessing the quality of human comparison in surveillance images. This is in part due to the application dependent nature of the 'quality' of an image, as accurate object information is important for surveillance, whilst enhanced contrast is seen to improve the human perception of general images. Also a complex model of multiple time varying illumination sources would be required for each and every camera in a surveillance system to provide colour constancy for evaluation. Such a large task currently seems infeasible. The quantitative evaluation of applying other illumination mitigation techniques is of interest to a variety of areas, suggesting that a technique which could be adjusted to evaluate a variety of techniques within a variety of domains from surveillance to image retrieval and even robotics could be widely useful. A focus upon fast data dependent techniques is performed here due to the complexity of illumination in surveillance scenes, and

the desire for real-time applications.

4.2 Illumination Filtration

This section outlines a technique of homomorphic filtering of the illumination effects from the image based upon the method described by Toth *et al.* [111]. This technique assumes objects consist of Lambertian surfaces and that illumination only changes slowly over space in the image. Toth *et al.* [111] suggests that this low frequency component can be filtered out by converting values to a logarithmic scale then a applying high pass filter, leaving the mid to high frequency details which in practise relate to the reflectance component of the image.

The intensity of the illumination on the surface of the object in the τ-th frame in an image sequence can be modelled as:

$$y_\tau(k) = I_\tau(k) \cdot R_\tau(k) \tag{50}$$

where k is the pixel index in the image, I is the illumination component and R is the reflective component in the image y.

If the reflectance component R can be separated from the illumination component I, then it can be used as an illumination invariant representation of the appearance. The slow rate of change of illumination over the space of the image means that it will consist of low frequency components of the image, whilst the reflectance model will consist significantly of mid to high frequency components. Applying the logarithm to **(50)** transforms the multiplicative relationship between y, I, and R into an additive one:

$$log(y_\tau(k)) = log(I_\tau(k)) + log(R_\tau(k)) \tag{51}$$

A high pass filter kernel can then applied to remove the low frequency illumination component I. Toth *et al.* [111] do not describe the choice of parameters within this process. Hence this work has looked to apply a Gaussian filter kernel with a small range of adjustable parameters, and retain the complementary features in order to achieve the desired high-pass filtration effect. An exponentiation of the filtered image therefore contains the illumination invariant image consisting of the reflectance information. The parameters of the Gaussian filter applied to remove the illumination relate to the filter size, standard deviation, and a weighting parameter which controls the amount of filtration applied. These parameters were adjusted independently to provide a variety of filters which were applied independently to evaluate the effects of each parameter. The parameters for each

application of homomorphic filtering are given as filter size, standard deviation, and weighting when the results are presented in results in section 4.6.

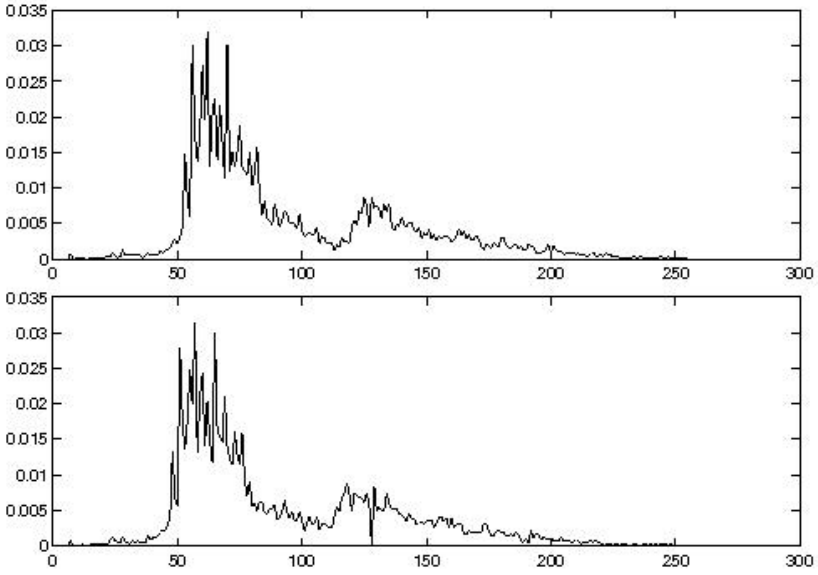


Figure 16: Individuals R values before and after illumination filtration

An example of the results of this filtration process is given as Figure 16. Although the output histogram does not appear to have changed much, small portions of the histogram have been adjusted to remove the low frequency spatially based co-occurance in colour channels. It is these components that Toth *et al.* [111] suggest relate to the slowly changing illumination levels.

4.3 Histogram Stretching

This section outlines the use of histogram stretching to perform the illumination transformation. This method proposes to stretch the object's histogram separately for each of the RGB components to allow for changes in the illumination spectrum. Stretching the histogram can make it appear more similar across a range of illuminations conditions without explicitly modelling those illumination sources. It also preserves the rank ordering of the histogram, which Finlayson *et al.* [33] suggest adds to the success of many of the colour invariance techniques.

The key points for histogram stretching are the selection of the upper and lower limits of the output histogram, and the upper and lower limits of the input

histogram for each colour channel. Histogram stretching then performs a linear mapping of the input to output values. We maximise the spread of the histogram by choosing the upper and lower limits of the stretched output to be 255 and 0 respectively. We choose the upper and lower limits of the object histogram based upon a single parameter a which denotes the percentage amount of histogram tails to be ignored. The removal of these tail components of the histogram aims to reduce the amount of noise in the input histogram. It is calculated by cumulating the count in each histogram bin from either end until the percentage a is reached.

If one denotes the lower input limit as b and the upper input limit as c, then the output of the stretching r' for any given input value in that channel can be calculated as:

$$r' = (r - b)\left(\frac{255}{c - b}\right) \tag{52}$$

This stretching transformation is performed upon each object pixel to generate a new object image which should have a higher tolerance to illumination changes without requiring either training or other assumed scene knowledge. This stretching provides a linear transformation of values so they lie across the entire histogram, whilst still retaining a similar shape to the original object component. The results of the stretching for global colour histograms is presented in Table 5 in Section 4.6 for a range of a values to explore the effect of changing the amount of the histogram that is ignored. The effect upon narrow spatial histograms for upper and lower clothing colour are presented in Tables 6 and 7 respectively. This technique is demonstrated in Figure 17.

Although simple linear stretching has been proposed to maximised the usage of the colour channel, a second form of centralised stretching has also been explored. This is based upon ideas similar to the Greyworld theory [4] in its intent to shift the colours based upon their mean value. For the R channel this is calculated by determining the mean value of the R histogram, $\mu(R)$, which is shifted to the centre of the histogram. The other values are then stretched in a similar manner between the centre and edges of the histogram space. Where the lower input limit as b and the upper input limit as c, this can be calculated as:

$$r' = (r - \mu(R))(\frac{123}{c - \mu(R)}) \tag{53}$$

for $r > \mu(R)$

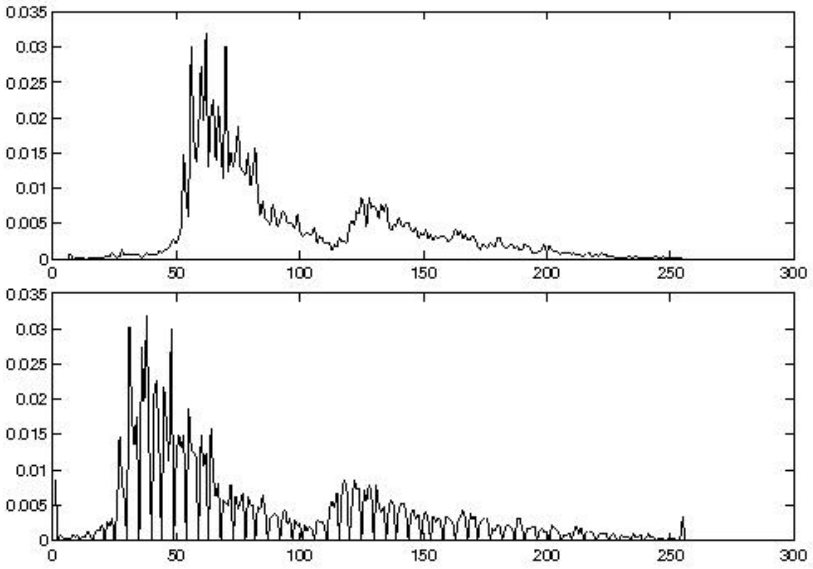


Figure 17: Histogram stretching of the individual's pixels

$$r' = (\mu(R) - r)(\frac{123}{\mu(R) - b}) \tag{54}$$

for $r \leq \mu(R)$

This centralised linear stretching transformation is performed in the same manner as the normal stretching transformation. It also requires no training or other scene knowledge; however the mean of the input histogram will be centred in the output histogram. The results for this centralised stretching technique are also presented in Table 5 in Section 4.6 for a range of a values. The effect upon narrow spatial histograms for upper and lower clothing colour are also presented in Tables 6 and 7 respectively.

4.4 Histogram Equalisation

This section outlines the use of equalisation to perform a data-dependent transformation of an individual's histogram. This method differs from histogram stretching as it can provide a non-linear transformation. First this section explains the application of histogram equalisation for the compensation of illumination effects

as proposed by Finlayson *et al.* [33], before defining a novel 'controlled equalisation' method.

Histogram equalisation, called here as full equalisation, aims to spread a given histogram across the entire bandwidth in order to equalise as far as possible the histogram values in the frequency domain. This operation is data-dependent and inherently non-linear as shown in Figure 18, but it retains the rank order of the colours within the histogram. The equalisation process is applied separately in each of the R, G and B colour components to remap their values according to the following transformation functions:

$$T_r(i) = \frac{255}{N} \sum_{j=0}^{i} p_r(j) \tag{55}$$

$$T_g(i) = \frac{255}{N} \sum_{j=0}^{i} p_g(j) \tag{56}$$

$$T_b(i) = \frac{255}{N} \sum_{j=0}^{i} p_b(j) \tag{57}$$

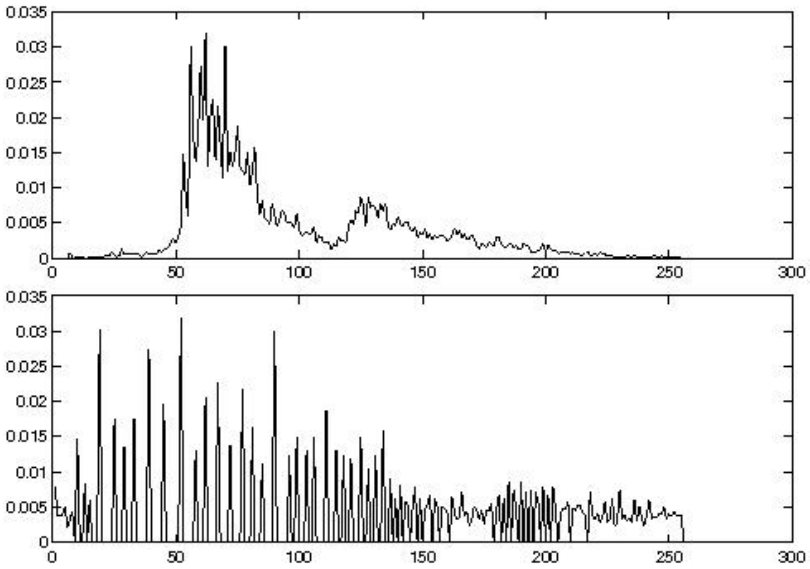


Figure 18: Full equalisation of the individual's pixels

This thesis also introduces here a form of 'controlled equalisation'. This process is based upon equalising a combination of the object pixels and an amount of pre-equalised pixels that is a proportion k of the object size. These pre-equalised pixels effectively 'control' the amount of equalisation such that the pixels are spread to a limited degree within the spectrum instead of being spread fully. Thus although an object should become more matchable under a range of illumination conditions, it is still likely to retain a higher degree of discrimination from objects of differing intrinsic colour. This controlled equalisation is shown in Figure 19.

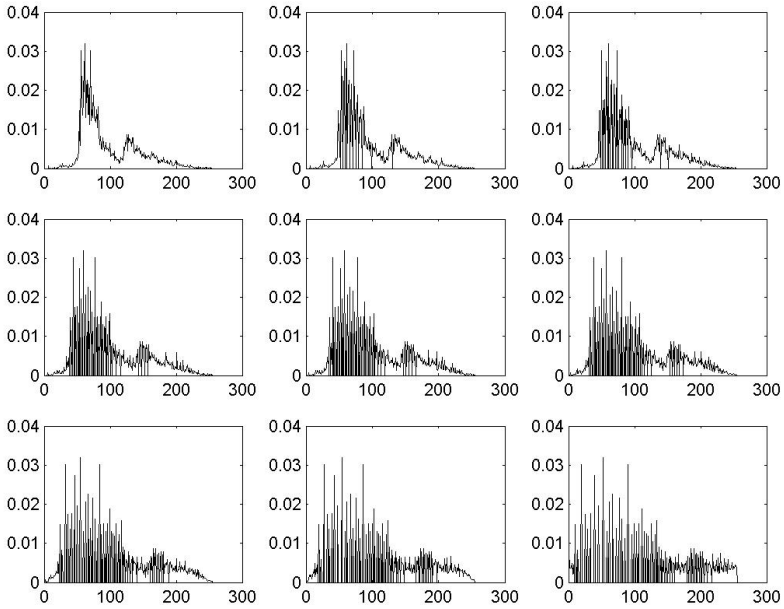


Figure 19: Controlled equalisation of the individual's pixels with varying k values

This equalisation can be formally described by designating the set of N pixels in a generic object as A, and calling B a second set of kN pixels which are perfectly equalised in their R, G, and B components. Note that the parameter k designates the proportionality of the amount of equalised pixels to the amount of pixels in A. From their union $A \cup B$, the cumulative histograms of the R, G, and B components, $p_r(i), p_g(i)$, and $p_r(i)$ for $i = 0 \ldots 255$ are computed. A histogram equalisation of the individual colour channels is then derived as shown in 58-60:

$$T_r(i) = \frac{255}{(1+k)N} \sum_{j=0}^{i} p_r(j) \tag{58}$$

$$T_g(i) = \frac{255}{(1+k)\,N} \sum_{j=0}^{i} p_g(j) \tag{59}$$

$$T_b(i) = \frac{255}{(1+k)\,N} \sum_{j=0}^{i} p_b(j) \tag{60}$$

These intensity transforms can be applied to re-map the R, G, and B components in the object's pixels providing the 'controlled equalisation'. The parameter k controls the amount of pre-equalised pixels used, which controls the spread of the object histogram. Higher values of k restrict the equalisation to the point that it approaches no equalisation, and as k approaches 0, the results approximate full equalisation.

A second variation on controlled equalisation is also possible by considering a technique that will shift the mean of the object's histogram to the centre of histogram space. This can be conducted similar to the centralised stretching technique presented in Section 4.3, by first finding the mean of the object histogram and performing the equalisation independently on each side of this mean. This is achieved simply by applying an equalised set of pixels equal to $\frac{(1+k)N}{2}$ to the left side of the mean value, and a similar number of equalised pixels to the right of the mean value. This positioning of equalised pixels ensures that the mean value of the object's histogram is shifted to the centre of the histogram space, whilst still allowing for controlled equalisation of the other histogram values, if at a differing non-linear rate on either side of that mean.

4.5 Comparing Illumination Mitigation Techniques

This section outlines the process that can be used to compare techniques for the mitigation of the effect of illumination changes upon colour appearance. The goal is to quantitatively evaluate the effectiveness of the various techniques by measuring similarities between colour appearance computed over objects. As this thesis is aimed at the analysis of humans as the objects of interest, the objects analysed are humans extracted from surveillance videos. The process occurs in five stages relating to the segmentation, application of illumination mitigation, extraction of colour features, comparison of the colour features, and for the training phase, the statistical analysis of results from objects known to be matching or non-matching.

The first stage of the process is to automatically extract the objects from the background in each frame of the videos. This research has utilised an adaptive

mixture model based upon that derived by Wren *et al.* [116], which quickly provides reasonably segmented objects. All the objects extracted along the frame sequence from a single individual are then manually collected into a single track so as to ensure correct data association. The particular segmentation method is not as important as using the same segmentation technique for all of the illumination mitigation techniques. This reduces any impact that differing segmentation might have upon the final results.

The second stage applies each of the different techniques in turn for each object in each frame so that the values of the object's pixels are remapped. The third stage then extracts the MCR histogram of the object's appearance in both global and spatial regions for the remapped images. This utilises the process described in sections 3.2 and 3.5. This produces the MCR histogram which is a 3-D, non-parametric sparse representation of an object's colour values that have already had the illumination mitigation applied. Each bin of the MCR therefore represents the illumination mitigated colour clusters, which still utilise the same normalised colour distance as the standard MCR. The number of such bins is not bounded apriori, and the position of their centroids is optimised through a *k*-means procedure as outlined in section 3.2.1.

In the fourth stage of processing, tracks of an individual are considered in pairs. One frame from each track is taken and a similarity measurement, S_f, is computed between their two MCR's based upon the method described in section 3.4. In a similar way, S_f values are computed for all other possible frame combinations from the two tracks and averaged so as to achieve a similarity measurement at the track level, S_t. A number of track pairs are built for both the cases of two different people (non-match case, or $H0$ hypothesis) or a single person (match case, or $H1$ hypothesis) and all S_t computed and stored for the two cases.

In the fifth stage, the distributions of the S_t values for each of the two hypotheses, $H0$ and $H1$, are statistically modelled by optimally fitting a Gaussian distribution on each. This is shown as Figure 20. In this way the likelihoods of each case can be described by their statistical averages, μ_{H0} and μ_{H1}, and their standard deviations, σ_{H0} and σ_{H1}. The posterior for each case can therefore be written as:

$$P(H0|S_t) = P_{H0}P(S_t|H0)P(H1|S_t) = P_{H1}P(S_t|H1) \tag{61}$$

where the priors P_{H0} and P_{H1} are assumed to be equal to 1. Thus the results are not using prior information about number of matching or non-matching cases.

The Gaussian assumption seems to well model the data, with σ_{H0} significantly

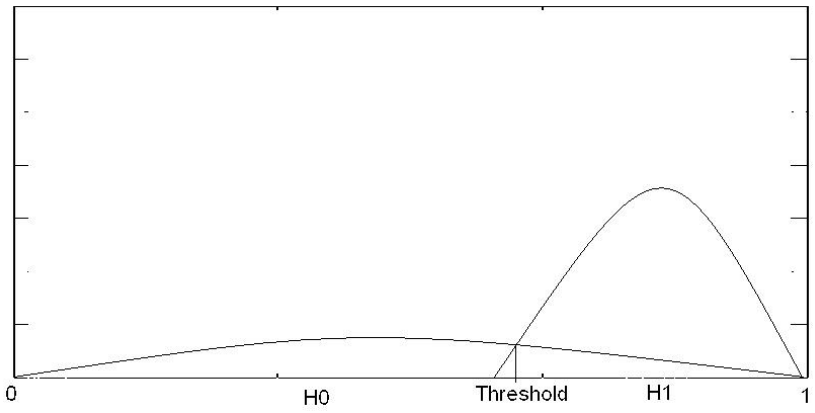


Figure 20: Intersection of the H0 and H1 curves of the height feature

larger than σ_{H1} (the dispersion of similarity values for different objects is obviously greater than that for different views of a same object). The performance evaluation for the different illumination mitigation techniques is then performed by computing the false alarm rate and the missed detection rate directly from $P(H0)$ and $P(H1)$, assuming $H0$ and $H1$ have equal priors. We derive the similarity value, $S_{t_{th}}$, for which $P(H0|S_t) = P(H1|S_t)$ as:

$$S_{t_{th}} = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \tag{62}$$

with

$a = \sigma_{H0}^2 - \sigma_{H1}^2$
$b = 2(\mu_{H0}\sigma_{H1}^2 - \mu_{H1}\sigma_{H0}^2)$
$c = \sigma_{H0}^2\mu_{H1}^2 - \sigma_{H1}^2\mu H0^2 + 2\sigma_{H0}^2\sigma_{H1}^2 \ln(\frac{\sigma_{H1}^2}{\sigma_{H0}^2})$

The false alarm rate, PFA, is then given by the tail $P(H0|S_t)$ below $P(H1|S_t)$, $(S_t \geq S_{t_{th}})$ and the missed detection rate, ***PMD***, from the tail of $P(H1|S_t)$ below $P(H0|S_t)$, $(S_t < S_{t_{th}})$. By identifying the matching errors from the estimated statistical distributions, one can calculate the most effective illumination technique. This is found by identifying the best possible trade-off between false alarm and missed detection rates using a minimum total error rate. Alternatively these statistics can also be analysed using priors that might reflect relative costs of missed detections and false alarms, or the expected rates of matching and non-matching classes.

4.6 Experimental Comparison of Mitigation Techniques

This section details the results from the comparison of these data-dependent, rank-preserving illumination mitigation techniques for a range of their parameters. These techniques are compared to the case of no attempt at illumination mitigation (i.e. leaving the colour values unaltered). These results are based upon an analysis of 15 tracks obtained from 4 individuals across 2 cameras with illumination from both natural sunlight and artificial sources. These tracks are a subset of the tracks experimented with elsewhere in this thesis and were chosen as the more reliably segmented tracks. This decision was made as the illumination mitigation experiment is not aiming at investigate the impact of segmentation errors upon the system, which might reduce the effectiveness of the system irrespective of the mitigation techniques. The similarity values are obtained from the 50 matching and 70 non-matching track pairs. The results investigate the effectiveness of the compared illumination mitigation techniques by comparing their estimated PFA, PMD and total error rate.

The results analyse three colour features: the global MCR, the upper clothing MCR, and the lower clothing MCR. The results for the global MCR features are presented first in Table 5 to show the effect that illumination mitigation can have across the entire object. The results presented include the mean and standard deviation of the both the non-matching and matching classes that are used to generate the $P(H0)$ and $P(H1)$ functions. These functions are then used to determine the theoretical errors relating to the probability of missed detection (PMD), the probability of false alarms (PFA), and the total error rate.

The results in Table 5 clearly demonstrate that applying histogram stretching reduces the similarity compared to the case of no mitigation attempt (first row). Even when the mean of the histogram is shifted to the centre of the colour channel, the linear stretching method actually increases the overall error rate. This occurs through an undesirable reduction in the similarity of matching objects, which is more significant than the decrease in the similarity of differing objects. The application of illumination filtration is suggested to remove highlights upon objects, as well as compensating for general illumination. The results show that varying the filter parameters to increase the size and variation of the Gaussian filter produces some improvement in both the matching of similar colours and that of differing colours with respect to no mitigation attempt; however these improvements are small. This would be consistent with a small reduction in the highlight

Table 5: Global similarity measurements for matching and non-matching tracks

Method	Param	Matching		Non-Matching		Theoretical Err%		
		mean	var	mean	var	PMD	PFA	total
None		0.8548	0.0965	0.2051	0.3421	1.94	9.39	11.34
Equal	Full	0.9104	0.0468	0.2313	0.3727	0.57	6.63	7.19
Equal	0.5	0.9118	0.0405	0.2410	0.3856	0.49	7.10	7.59
Equal	1	0.9147	0.0329	0.2456	0.3923	0.37	6.93	7.30
Equal	2	0.9155	0.0370	0.2487	0.3968	0.45	7.54	7.99
Cent Equal	0.5	0.9159	0.0282	0.2451	0.3912	0.30	6.47	6.77
Cent Equal	1	0.9169	0.0266	0.2435	0.3887	0.27	6.15	6.42
Cent Equal	2	0.9150	0.0237	0.2451	0.3912	0.24	6.17	6.40
Stretch	0.1%	0.7899	0.1067	0.1910	0.3092	2.66	10.21	12.87
Stretch	1%	0.7712	0.1128	0.1797	0.2926	2.93	9.81	12.74
Stretch	5%	0.7464	0.1176	0.1617	0.2659	3.05	8.52	11.57
Cent Str	0.1%	0.7040	0.0956	0.1806	0.2912	2.88	12.02	14.90
Cent Str	1%	0.6869	0.1109	0.1806	0.2904	3.97	14.14	18.11
Cent Str	5%	0.6376	0.1236	0.1665	0.2672	5.40	15.39	20.79
Filter	5 1 0.5	0.8559	0.0860	0.2108	0.3479	1.64	9.23	10.87
Filter	7 2 0.4	0.8536	0.0832	0.2089	0.3452	1.54	8.90	10.44
Filter	7 3 0.5	0.8657	0.0785	0.2127	0.3507	1.37	8.56	9.93
Filter	7 2 0.6	0.8719	0.0759	0.2190	0.3598	1.32	8.90	10.23

and shadowing effects which slightly vary appearance. The results for the colour equalisation techniques show the best improvement in matching scores, indicating that as Finlayson *et al.* suggest [33], they are capable of providing colours that are more illumination invariant. Equalisation reallocates the histogram bins to dynamically compress the bins with low counts and expand non-linearly the bins with high counts. This non-linear reallocation seems to better compensate for the appearance changes occurring under the limited illumination variations observed both within and between the two cameras. Whilst full equalisation produces a reduced overall error rate that is lower than that of the controlled equalisation, the controlled version seems to provide increased similarity in the histograms of matching individuals. The improved error rate also occurs because of the reduced level of variation in the matching results as they are consistently more matchable. These rates are even further improved when the mean of the histogram is centralised as a part of the equalisation approach. Picking the 'best' technique for

global features between these full, controlled and centralised controlled equalisation approaches would require one to define costs for both a false alarm and a missed detection. This is required because although the centralised controlled equalisation approach minimises the overall error rate, other techniques provide increased accuracy in matching. Deriving appropriate costs depends significantly on the actual application, as some will require high matching accuracy, and other application may be able to minimise the cost of false matches using semi automated techniques.

Table 6: Upper MCR similarity for matching and non-matching tracks

Method	Param	Matching		Non-Matching		Theoretical Err%		
		mean	var	mean	var	PMD	PFA	total
None		0.7256	0.1549	0.1680	0.2865	5.91	13.54	19.45
Equal	Full	0.7329	0.1220	0.1562	0.2637	3.38	8.99	12.37
Equal	0.5	0.7647	0.1130	0.1597	0.2679	2.57	7.53	10.11
Equal	1	0.7903	0.0938	0.1724	0.2878	1.81	7.16	8.97
Equal	2	0.7989	0.0724	0.1731	0.2923	1.09	5.79	6.88
Cent Equal	0.5	0.7953	0.0885	0.1754	0.2882	1.61	6.77	8.38
Cent Equal	1	0.8064	0.0906	0.1836	0.3006	1.74	7.56	9.29
Cent Equal	2	0.8007	0.0849	0.1889	0.3105	1.67	8.25	9.92
Stretch	0.1%	0.5876	0.1535	0.1343	0.2270	8.34	14.41	22.75
Stretch	1%	0.5446	0.1531	0.1185	0.2019	9.18	13.52	22.70
Stretch	5%	0.4739	0.1656	0.0875	0.1489	11.75	10.13	21.87
Cent Str	0.1%	0.6757	0.1242	0.1670	0.2690	4.72	13.16	17.88
Cent Str	1%	0.6471	0.1114	0.1608	0.2579	4.22	12.71	16.93
Cent Str	5%	0.6356	0.1293	0.1598	0.2586	5.73	14.66	20.39
Filter	5 1 0.5	0.7367	0.1443	0.1607	0.2781	4.80	11.36	16.16
Filter	7 2 0.4	0.7348	0.1537	0.1683	0.2886	5.65	13.15	18.80
Filter	7 3 0.5	0.7402	0.1459	0.1655	0.2874	4.98	12.22	17.21
Filter	7 2 0.6	0.7468	0.1286	0.1700	0.2952	3.97	11.70	15.67

Table 6 demonstrates the effect of applying these techniques upon a narrow spatial object histogram. These results are from the upper clothing colours which consist of a much smaller set of colour clusters than the global histogram. It is also important to note that although these results are based upon four individuals of differing clothing, some of the upper clothing are similar in colour. Whilst one individual is wearing a white top, and one a blue shirt, the other two are wearing

black tops, one a suit jacket, and the other a black jumper. For the purposes of these results, each individual is considered as a separate entity, and thus whilst the two black tops may be of similar colour, if they are matched it is considered a false alarm. The results clearly demonstrate that the overall error rate in this individual feature are generally higher than those for the global histogram as the differences between the histograms is not as distinct. The table also shows that the mean similarity of such narrow histograms is considerably lower, although the mean similarity of non-matching is also lower. The effects of the techniques are very similar to the global results, with histogram stretching variations worsening error rates, and homomorphic filtering providing marginal improvements. Again the equalisation based approaches provide significant improvements, with overall error rates reduced to almost a third of that achieved with the natural image. Unlike the global results, all of the controlled equalisation approaches improve upon full equalisation, with the controlled equalisation at $k = 2$ providing the best results. It is also worth noting that these improvements are the most significant in lowering missed detections, where less than 20% of these errors occur using the best mitigation when compared to the unmitigated approach.

Table 7 also shows the effect of applying these techniques upon a narrow spatial object histogram; however this time based upon the lower clothing colours. Of the four individuals, two of the individuals are wearing white pants, whilst the other two individuals are wearing dark coloured pants. As with the results for the upper clothing similarities, each individual is considered a separate entity, and whilst they may have similar lower clothing colours, it is considered a false match where they are deemed matching. Although one might expect this similarity in clothing colour to lead to greater error rates, the results actually show higher accuracy than the global MCR features. The difference is evident from Table 6 in that whilst the mean similarity of non-matching lower MCR's is much lower than for the other features, so are the mean similarity of the matching cases. This lower level of matching similarity is likely be the main factor that leads to the greater probability of missed detection (PMD) for this case. Table 7 demonstrates different results to those shown in the previous results in this section. Firstly the most promising equalisation based approaches sometimes provide decreased overall accuracy, although the centralised controlled equalisation does show the lowest error rate for missed detections. Histogram stretching, contrary to previous results, show generally promising results, even though they don't tend to improve the detection rate; however the centralised stretching method shows the worst error rates. The homomorphic filtration method shows the same slight improvement in error rates, consistent with the previous results. This would tend to show its

Table 7: Lower MCR similarity for matching and non-matching tracks

Method	Param	Matching		Non-Matching		Theoretical Err%		
		mean	var	mean	var	PMD	PFA	total
None		0.7256	0.1598	0.1126	0.2181	4.17	6.16	10.33
Equal	Full	0.7275	0.1634	0.1132	0.2077	4.13	5.57	9.70
Equal	0.5	0.7339	0.1535	0.1286	0.2293	4.17	6.93	11.10
Equal	1	0.7396	0.1578	0.1339	0.2385	4.57	7.74	12.31
Equal	2	0.7508	0.1508	0.1331	0.2397	3.94	7.06	11.00
Cent Equal	0.5	0.7655	0.1466	0.1376	0.2413	3.54	6.62	10.16
Cent Equal	1	0.7791	0.1276	0.1540	0.2637	2.94	7.27	10.21
Cent Equal	2	0.7961	0.0988	0.1712	0.2878	1.94	7.19	9.13
Stretch	0.1%	0.6942	0.1579	0.1099	0.2048	4.44	6.16	10.61
Stretch	1%	0.6518	0.1569	0.0951	0.1694	4.18	4.60	8.78
Stretch	5%	0.5818	0.1866	0.0702	0.1228	6.00	3.56	9.56
Cent Str	0.1%	0.7148	0.1352	0.1514	0.2536	4.27	9.60	13.87
Cent Str	1%	0.7127	0.1595	0.1381	0.2319	5.30	8.60	13.90
Cent Str	5%	0.6952	0.1367	0.1356	0.2240	4.05	7.58	11.63
Filter	5 1 0.5	0.7388	0.1679	0.1041	0.2033	3.83	4.85	8.68
Filter	7 2 0.4	0.7368	0.1640	0.1066	0.2076	3.81	5.10	8.91
Filter	7 3 0.5	0.7390	0.1586	0.1114	0.2160	3.75	5.51	9.25
Filter	7 2 0.6	0.7515	0.1606	0.1142	0.2202	3.74	5.54	9.28

usefulness in smoothing out shading and highlights, perhaps even indicating its usefulness in combination with other techniques.

4.7 Discussion of Illumination Mitigation

This chapter compared various data-dependent, rank-preserving techniques in an attempt at improving the invariance of a person's appearance across camera views without exploiting any scene knowledge. These techniques were chosen because of their lack of requirements on the knowledge of scene illumination, and their low computational complexity. This makes such techniques ideal for complex surveillance scenes where the real-time implementation is important, but illumination changes may be complex and time varying. The results presented show that some of these techniques can significantly mitigate the effects of illumination variations, with varying levels of improvement to the matching error rate, and the total error rate. Therefore, their use seems strongly beneficial for surveillance ap-

plications which may have varying illumination levels. The histogram stretching technique tends to diminish the similarity of matching objects and increase that of differing objects and its use is therefore not recommended. The illumination filtration technique alone provides a marginal improvement in the similarity of an object's appearance under illumination changes, which is likely due to its removal of illumination highlights on the object. These results were the most consistent; however are not likely to provide reasonable improvement alone. The equalisation of an individual's colour histograms provides a significant improvement in appearance similarity under differing illumination in most of the results presented. Whilst full equalisation produces a good improvement in overall error rate, controlled equalisation produces more improvement in similarity between matching objects. The centralised controlled equalisation tends to provide the best overall error rate, with often the lowest level of missed detections.

The results of these experiments are a little varied depending upon the level of colour variation and the complexity of the MCR histogram which is observed. The most promising results were obtained using the centralised version of the controlled histogram equalisation. This combination tends to produce the best overall error rate, whilst reducing the amount of missed detection errors. The results are not totally conclusive however, and suggest that similar experiments might be useful for any particular application in order to determine the most appropriate techniques. An investigation of the costs of missed detections and false alarms is also necessary for the particular application in order to select the best techniques as most applications are likely to require different interventions depending upon the error type. It is also worth noting that although these techniques have been applied independently, improved results may be obtained by combining the centralised controlled equalisation and filtration processes. Such combinations may seem beneficial; however an evaluation of the error rates for such combinations would be required along with a deeper investigation of the trade off between computational complexity and accuracy should be considered.

4.8 Summary of Illumination Mitigation and Future Enhancements

Many techniques have been suggested in the literature to compensate for the effects of variable illumination over an object's appearance. In a scenario of video surveillance, explicitly estimating the illumination over 3-D deformable moving targets such as humans simply proves impractical. This chapter has therefore

looked to develop a five step process that can be used to compare techniques for mitigating the effects of illumination changes upon colour appearance for a given scenario. This scenario can be adapted based upon the objects of interest; however the results may be generalisable for wider applications with typical illumination conditions. The process used to compare the improvements achieved using the various techniques can be summarised:

1. Segmentation of the object from the background.
2. Apply illumination mitigation technique.
3. Extract the MCR appearance feature(s).
4. Compare the MCR appearance features between tracks that are known to be matching or non-matching.
5. Analyse the theoretical error for the illumination mitigation techniques applied to determine the most accurate method

This process has been applied in the surveillance domain to find that illumination filtration provides a small degree of improvement in the identification error rates, likely due to the reduction of patches with saturated colours which show very low frequency spatial changes. Equalisation techniques seem to provide significantly greater error reduction, especially the centralised controlled equalisation technique; however there is no reason that these two techniques could not be both applied to both remove illumination highlights and equalise the image. Indeed the framework provided could be used in future investigations of this nature, or to evaluate the usage of a range of other scene independent techniques, or even to evaluate the usage of scene information for reducing the effect of illumination on colour appearance. This would be achieved using the same process as the scene information would be incorporated as a component of step 2 through the application of assumptions about the general illumination level based upon changes in background colours, and would therefore be automatically incorporated into the MCR features for quantitative evaluation.

A second area of possible future enhancements would be to create a general database of objects moving through scenes of typical illumination for general scenarios. Such a database would be especially useful where algorithms were made available for easy comparison to new techniques; however a number of difficulties could arise. Primarily the size of the database is determined based upon the

number of scenarios considered, as well the range of illumination conditions that might be captured for evaluation. Secondly where the database allows for scene information to be used, that scene information needs to be captured in enough detail to provide for future possibilities that might become available with significant increases in computer power. Such information may potentially progress as far as ray tracing from illumination sources to objects. The level of detail required for future experiments about the illumination sources is difficult to ascertain, and could be extremely difficult to extract and record effectively.

5 Identification of Segmentation Errors

This chapter looks to identify frames with significant segmentation errors in an individual's track. Such erroneous frames are a significant source of error for further processes, such as wide area tracking. These can cause changes in object features, as they lead to both the inclusion of regions that are not part of the object, or the removal of regions that are part of the object, reducing the effectiveness of subsequent processes. Whilst this can become a problem for any system, these changes in features can also be analysed along the known track of an object to identify those erroneous frames. The identification of segmentation errors through their effect on object features is explored through an analysis of the changes in colour appearance and size features along the frame sequence. This is possible as such features have been designed to be invariant to illumination and viewpoint variation, and thus should remain intrinsically the same for correctly tracked objects within a single camera view. The features used and compared include the global MCR, upper and lower clothing MCR's and the relative changes in bounding box size. Note that the extraction and comparison of such MCR features are described earlier in Chapter 3. The identified errors can then be compared to those identified by human expert identified major segmentation errors, which are defined here as errors which affect more than 15% of the pixels that are, or should be associated with an object. Errors due to the incorrect removal of pixels from an object are considered just as important as errors associated with the addition of incorrect pixels to an object. Rather than a comparison of a manually selected pixel based object model, this method looks at a less time consuming and more qualitative evaluation of the human expert on the overall segmentation at the object level.

As this technique is aimed at automatically identifying segmentation errors, rather than analysing the quality of the segmentation, no comparison of segmentation techniques is performed. The segmentation used is based upon an adaptive Gaussian model, similar to that used in the Pfinder project [116], because of its speed and reasonable accuracy. Major segmentation errors may occur more frequently than they do with other more complex background modelling techniques; however, this is less of a problem where such errors can be identified. This technique is likely to be most useful where it allows for significant error removal with minimal impact to the reliable portions of the data, especially where frame rates are low and track lengths are short. This identification of segmentation errors could be useful to improve a range of applications including, but not limited to: a) matching single objects from disjoint camera views, where matching is enabled by accurate extraction of features such as shape and appearance in each

view [53, 70, 123]; b) creating a synthetic and faithful *pictorial summary* of a tracked object using one or a small number of frames where the object is not affected by major segmentation errors; or c) accurate searching for the object in image or video archives.

This chapter begins by providing the background and literature aimed at the identification of segmentation errors. This is distinct from the object segmentation literature outlined in Section 2.1, as it aims to identify those frames in a video sequence where the errors are significant, not to evaluate the segmentation quality. Two methods are then proposed to do this, firstly Section 5.2 looks at investigating the changes in bounding box size over the time to identify errors, whilst Section 5.3 looks at analysing the change in appearance features along the track of an object to identify errors. The experimental techniques to verify and compare these techniques are presented in Section 5.4 along with the results of those experiments. The implication of these results are then discussed broadly in Section 5.5. The chapter concludes with a summary of the techniques and possible expansions of the technique into other application areas.

5.1 Segmentation Error Identification Background

Segmentation of moving objects has been widely used through the computer vision literature to extract objects which are moving compared to the background. This is based upon the assumption that objects which are not currently moving, and have not been moving are likely to be of little interest. A full review of the difficulties associated directly with segmentation is given within the literature review chapter as section 2.1. This background looks more specifically at the difficulties associated with identifying major errors in this process. Essentially this review of the literature suggests that errors in segmentation of both minor and major proportions occur when using all of the common segmentation techniques [96]. Some segmentation techniques perform better in specific circumstances; however errors still occur, with minor errors around the edges of objects being frequent in the complex scenes that dominate the real surveillance environment.

Identifying segmentation errors is made more difficult for articulated objects which can change their shape within constraints, as this can lead to appearance changes through self occlusions. A number of hypotheses do hold in a statistical sense for tracked humans in a surveillance environment: they tend to walk upright, wear clothing that may differ for the vertical layers relating to the torso and legs, and have an appearance that is often similar for different equatorial views. Illumination provides another challenge as it can vary over time, and in different patterns

depending upon camera location and whether it is indoor or outdoor. This change perceived appearance features and the contrast of the object and the background. The literature on identifying segmentation errors in a track seems to be relatively limited. For instance, Erdem *et al.* [29] have tried to reduce segmentation errors for a 3D television application, to improve the temporal stability of object segmentation, rather than identify and remove errors. They achieved their aim by minimising changes in the global colour histogram and turning angle function of the boundary pixels of the segmented object in each frame to maximise temporal stability. Little other research has really looked at the idea of identifying frames that might have erroneously segmented objects from a video sequence without either manual assessment, or a pixel level manual annotation of the ground truth objects. Instead the focus has remained upon directly improving the segmentation.

5.2 Identifying Segmentation Errors Through Changes in Bounding Box Height

Bounding boxes have been widely used in the literature to speed up the analysis of objects by creating a simple rectangular model of the object. This can then be used to identify object bounds and when they overlap. Hence it is a good candidate for fast identification of segmentation errors in a person's track, if that object is correctly tracked either manually or by using one of the many popular motion estimation techniques, such as [125]. Once accounting for perspective distortion, the changes in the object or bounding box size are likely caused by large segmentation errors, or possibly occlusions.

As people are articulated objects, their position as well as their size and shape can change within limits. Thus, bounding boxes for an accurately segmented person can change due to movement actions such as walking, or from the camera perspective as a person moves towards or away from the camera position. The expected limits on the size changes can be modelled for a given camera as a part of camera calibration. This could either be through a manual assessment of the changes in bounding box for a given individual, or could be automated through statical analysis of a number of individuals moving through the scene. A more complex model based upon expected changes in the direction an individual is moving is also possible. This may allow for a greater sensitivity to segmentation errors; however this hypothesis has not yet been tested.

The expected changes in the bounding box size can be simplified by assuming that the camera frame rate is not slower than a few frames per second, and people

are walking upright in the viewed area. These two assumptions generally hold for the video surveillance environment where people are traversing a space viewed by the camera in order to travel from point A to point B. For most typical frame rates, the allowable amount of low impact segmentation errors is often higher than the changes due to perspective distortion. This notes that some level of segmentation errors are to be expected, especially in complex environments [96]; however often this low level of segmentation only has a minimal impact upon object features. Although many object statistics could be used to analyse changes in the shape of an object, this study found the changes in vertical height of the bounding box tends to remain invariant whilst a person is walking, once perspective distortion is discounted. This measure still remains sensitive to actions where a person might bend over, or become partially occluded; however these actions may also cause significant change in the appearance of an individual. Typical values of the ratio in vertical size between one frame and the next vary in a small range around one, depending upon frame rate and the amount of perspective distortion present in the camera view.

Figure 21 shows an example of the ratios of bounding box height obtained between one frame and the next for a track where there is a single frame with a large segmentation error. The error is clearly indicated by the change in ratio value to below 0.6, as the subsequent frame is dramatically shorter than the current frame. The next ratio is over 1.5, indicating that the next frame is much larger than the erroneous frame as it has returned to the normal size. The 5 sample frames from the track show that the image height of the object is diminishing along the track as the object moves away from the camera. The three middle frames show the frame before the error, the frame with the error and the frame after the error. The other two frames from the start and end of the track illustrate the change in object size due to the increased distance from the camera along the track. The change in image height of the individual when incorrectly segmented is evident, as is the loss of legs in this frame. Although this method works well for errors in a single frame, or a short run of frames, analysis of gradual increases in segmentation errors remains a problem for this method.

5.3 Identifying Segmentation Errors Through Appearance Feature Analysis

Large segmentation errors have a significant impact upon colour appearance features. Thus the changes in the similarity of these features along a track are likely

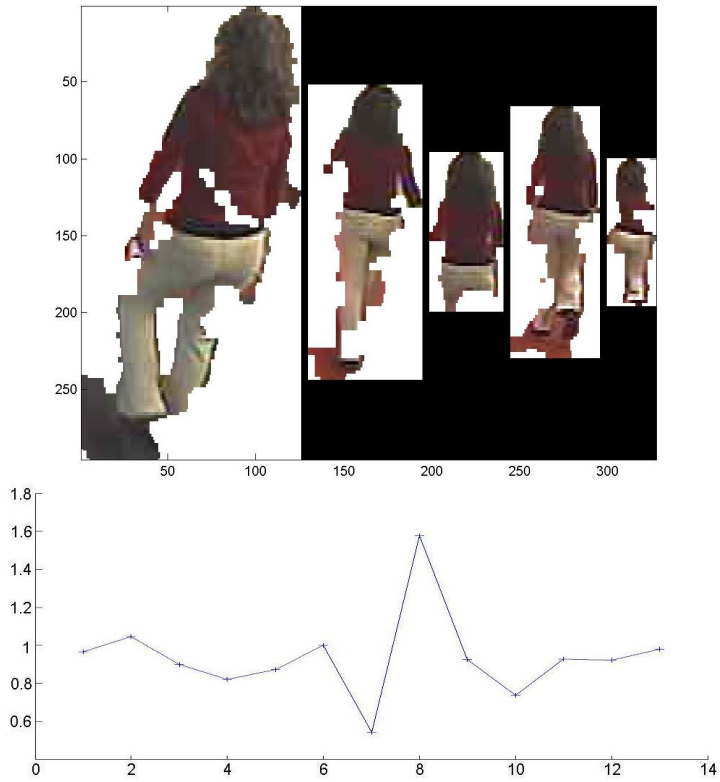


Figure 21: Example of changes in bounding box height ratios where a large segmentation errors occurs and 5 sample frames from the track including the erroneous frame

to indicate such errors. The **(MCR)** features used in this analysis as colour features are based upon the method previously described in Chapter 3. MCR are essentially colour histograms in the joint R, G, B space, built with sparse bins whose position and number is adjusted to fit the pixel distribution. Instead of just using a global colour feature, as in [29], this method also proposes to compare the use of the two extra spatial colour features relating to the upper and lower clothing colours of a person, which were detailed previously in Section 3.5. These features are chosen to represent the differing colours that often occur for the clothing on the torso, and those on the legs and are outlined below. The narrow spatial aspect of these features also allows for a more sensitive analysis of the spatial positioning of a person's colours. This ensures that changes in the position of the colours can also be detected, such as where segmentation errors remove large portions of the

object. The process for extracting the MCR's is summarised again here as:

1. The individual is segmented from the background
2. A controlled equalisation step performs a data-dependent intensity transform. This spreads the histogram to compensate for minor illumination changes that can be expected within the indoor and outdoor surveillance environments.
3. Initial MCR features are generated using colour clusters based upon the RGB values of the segmented individual
4. Online K-means clustering of pixels of similar colour within a normalised colour distance δ generates the MCR of each spatial region. The cluster centre is the average of the colour values within δ, allowing it to better represent the colour cluster. Due to the movement of colour clusters iteration of cluster improvement and cluster assignment are necessary; however, as explained in [70], three iterations provide an accurate representation with a minimum of computational cost.

The three MCR features are defined as:

1. The global MCR feature, which represents the colours of the whole segmented object without any spatial information.
2. The upper MCR feature, which represents the colour of the top portion of clothing. This corresponds to the region from 30-40% of the person from the top of the object's bounding box as shown between the lines towards the top in Figures 22 and 23. This narrow band was chosen to ensure that it avoids the inclusion of the head and hair of the object, as well as low necklines, but does not go so low that it includes the belt area, or overlaps with the lower colour, or bottom area.
3. The lower MCR feature aims to represent the colour of the lower portion of clothing. This corresponds to the region from 65-80% of the object from the top of the object's bounding box as shown between the lines towards the bottom in Figures 22 and 23. This narrow band avoids the very bottom of the object which can be prone to shadows, or artefacts where the feet touch the ground. It also tries to avoid overlapping with the belt or upper torso area of the person.

The narrowness and positioning of both of the upper and lower MCR regions allows for them to remain constant under minor segmentation errors. Such errors are common and will only have a minimal impact upon a person's features, hence they need to be allowed whilst still remaining sensitive to large segmentation errors. These features also allow for the inclusion of spatial colour features which could possibly identify the difference between people when tracking is incorrect. Increasing the width of the spatial bands of the colour regions is likely to make the features more sensitive to segmentation errors, whilst not necessarily improving the quality of the features for clothing with limited colour variation.

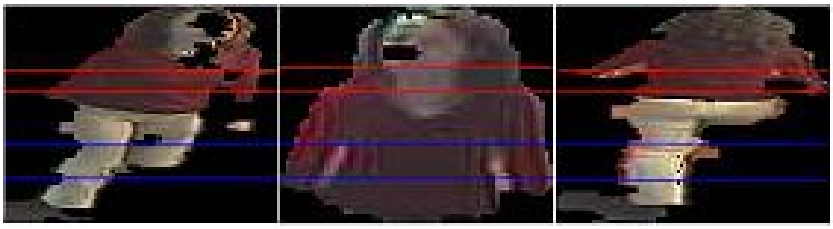


Figure 22: Example of upper and lower regions from three segmentations of one person

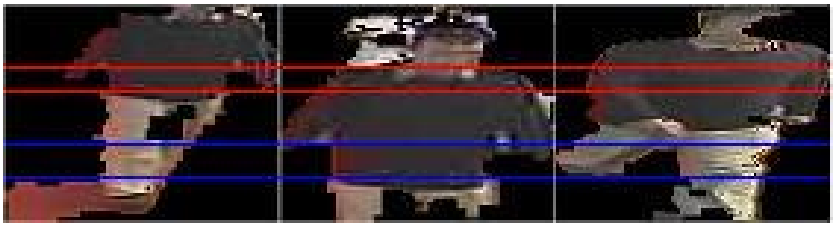


Figure 23: Example of upper and lower regions from three segmentations of a second person

Figures 22 and 23 show the upper MCR feature region between the lines toward the top of the person, and the lower MCR feature region between the lines toward the bottom of the person. Figure 22 demonstrates three frames showing frontal and rear views of a person, and a frontal view with a significant segmentation error where the lower half of the person is not found. In this frame the colours within the upper and lower regions change significantly from those in the other two frames. Figure 23 shows two views of a second person where segmentation is arguably reasonable, even if a portion of the head is not correctly segmented in the

first frame. The frame shown in the middle is poorly segmented; however a white object above the person is partially included within the object. This leads to an added amount of white in the global colours, that is not entirely dissimilar to the pants colour. Such an error leads to the frame having little discernible difference in global colours; however the upper and lower colour regions clearly indicate a change in the spatial positioning of the colours, which can be used to identify this poorly segmented frame.

5.3.1 Comparing Colour Features Between Frames

Once extracted, the MCR features can be compared to each other to determine if the features change over time along the track of the object. We begin by assuming that objects are tracked correctly, even though large and sustained changes in human object features may indicate potential data association errors. This analysis of tracking errors is not explored within this thesis, as the individual's are manually tracked. This choice has allowed the project to focus on the analysis of data from single individuals that are reliably tracked. Changes in object features along the track are therefore likely to be caused by errors in the identification of foreground pixels, or through other causes such as occlusion, cluttering, or major lighting changes.

This method performs an automatic comparison between the frames of a track to identify frames affected by major segmentation errors. It utilises the global, upper, and lower MCR colour features using the similarity measurement described in section 3.4. Given any two MCRs, A and B, for each bin of A a search is conducted for a matching bin in B to calculate their intersection. Such intersections are added up for all matching pairs, providing a similarity measurement that is equivalent to the complement of the Kolmogorov distance between two distributions with equal priors [127].

This process is used to generate pairwise similarity values for the three MCR features in each frame to every other frame in the track. We apply a statistical analysis of a known training set of the non-matching *H0* or the matching *H1* sets of features to determine Gaussian likelihood functions for given similarities being matching or non matching [37]. Classification can also be obtained by fusing together the matching and non-matching likelihoods of each of the three feature comparisons in an ensemble of classifiers. We assume the features to be conditionally independent and so apply Bayes theorem as:

$$P\left(H0|s_G, s_U, s_L\right) = B\left(P\left(s_G|H0\right)P\left(s_U|H0\right)P\left(s_L|H0\right)\right) \quad (63)$$

$$P\left(H1|s_G, s_U, s_L\right) = \left(P\left(s_G|H1\right)P\left(s_U|H1\right)P\left(s_L|H1\right)\right) \quad (64)$$

where B is a prior that can be used to bias the operating point of the system.

5.3.2 Typical MCR Patterns of Major Segmentation Errors

Appearance changes are generally caused by either major segmentation errors, where portions of the object are not extracted correctly from the background, by large changes in the illumination conditions, or by tracking errors. Segmentation errors of less than 15% are considered as minor in this process as they commonly occur [96], and they will often only cause limited changes in object appearance. Major errors produce significant changes in the appearance of a single frame; however, they often have a different impact upon the segmentation results when considering their difference across the whole track. Three different error patterns are demonstrated in Figure 24, highlighting how the different errors tend to influence the similarity of the proposed features.

The first error pattern in Figure 24a shows how large segmentation errors cause a very low similarity in the fused features between that frame and every other frame causing a characteristic 'cross' in the pairwise comparisons. Where a small number of frames have similar large segmentation errors, such as losing the lower half of the object, these frames will tend to be similar to each other, but distinct from the rest of the track.

Figure 24b shows the second error pattern where the portions of the track are self similar, but persistently different from each other. This occurs for large illumination changes, such as switching a light on. It might also be expected for tracking errors, although this research has not investigated this concept as manual tracking is currently used. In this case extracting both sets of feature representations could be useful for manual analysis of the object's track. The bounding box is not directly affected by large illumination changes; however in practise the change in the amount of contrast between the object and background often leads to major segmentation errors as well.

Figure 24c shows the third error pattern, which occurs for gradual illumination changes, such as clouds moving to cover the sun. The error pattern shows that each frame is still likely to match the majority of the rest of the track, however the initial frames could have a large difference in appearance to the frames toward the

	1885	1886	1887	1888	1889	1890	1891	1892	1893	1894
1885	1	0.72346	0.040269	0.67614	0.75905	0.82743	0.89009	0.63706	0.6335	0.51523
1886	0.72346	1	0.02513	0.6977	0.81461	0.84949	0.81277	0.89385	0.76962	0.59259
1887	0.040269	0.02513	1	0.026171	0.053725	0.03417	0.043174	0.04205	0.013697	0.020038
1888	0.67614	0.6977	0.026171	1	0.66853	0.79567	0.76798	0.69213	0.84459	0.81089
1889	0.75905	0.81461	0.053725	0.66853	1	0.82113	0.87647	0.73828	0.65974	0.83645
1890	0.82743	0.84949	0.03417	0.79567	0.82113	1	0.85833	0.91382	0.80466	0.57416
1891	0.89009	0.81277	0.043174	0.76798	0.87647	0.85833	1	0.87336	0.73598	0.65928
1892	0.63706	0.89385	0.04205	0.69213	0.73828	0.91382	0.87336	1	0.69348	0.64218
1893	0.6335	0.76962	0.013697	0.84459	0.65974	0.80466	0.73598	0.69348	1	0.91203
1894	0.51523	0.59259	0.020038	0.81089	0.83645	0.57416	0.65928	0.64218	0.91203	1

a) Track P showing a single large segmentation error

	3095	3096	3097	3098	3099	3100	3101	3102	3103	3104
3095	1	0.000938	0.000945	0.69484	0	0.046527	0.066226	0.008902	0.028491	0.048307
3096	0.000938	1	0.73378	0.003612	0.86109	0.26509	0.28603	0.061421	0.027958	0.14846
3097	0.000945	0.73378	1	0.006682	0.3299	0.26668	0.316	0.055803	0.027105	0.13583
3098	0.69484	0.003612	0.006682	1	0.005538	0.081253	0.078724	0.047608	0.26263	0.08295
3099	0	0.86109	0.3299	0.005538	1	0.34415	0.11527	0.078596	0.060614	0.19886
3100	0.046527	0.26509	0.26668	0.081253	0.34415	1	0.94116	0.65052	0.75248	0.77442
3101	0.066226	0.28603	0.316	0.078724	0.11527	0.94116	1	0.67511	0.87103	0.85858
3102	0.008902	0.061421	0.055803	0.047608	0.078596	0.65052	0.67511	1	0.59428	0.86182
3103	0.028491	0.027958	0.027105	0.26263	0.060614	0.75248	0.87103	0.59428	1	0.90352
3104	0.048307	0.14846	0.13583	0.08295	0.19886	0.77442	0.85858	0.86182	0.90352	1

b) Track Z showing large illumination changes

	1298	1299	1300	1301	1302	1303	1304	1305	1306
1298	1	0.79247	0.80477	0.74167	0.29525	0.32224	0.31006	0.093313	0.04554
1299	0.79247	1	0.9232	0.67458	0.66306	0.24841	0.26767	0.2872	0.021861
1300	0.80477	0.9232	1	0.624	0.79788	0.71125	0.72767	0.309	0.17989
1301	0.74167	0.67458	0.624	1	0.69795	0.78452	0.23059	0.28529	0.076366
1302	0.29525	0.66306	0.79788	0.69795	1	0.46039	0.51422	0.79906	0.3297
1303	0.32224	0.24841	0.71125	0.78452	0.46039	1	0.80858	0.74389	0.48604
1304	0.31006	0.26767	0.72767	0.23059	0.51422	0.80858	1	0.72726	0.48639
1305	0.093313	0.2872	0.309	0.28529	0.79906	0.74389	0.72726	1	0.72857
1306	0.04554	0.021861	0.17989	0.076366	0.3297	0.48604	0.48639	0.72857	1
1307	0.006742	0.00811	0.008823	0.030715	0.19158	0.13621	0.085893	0.5694	0.30402

c) Track W showing a gradual illumination change

Figure 24: Three typical error patterns of frame based pairwise similarity comparisons given between 0 and 1

end. This case is not caused by segmentation errors, so each frame is considered equally suitable to be included in a robust track.

A final pattern which may emerge is where there are multiple portions of the track that are self similar, but significantly differing from the other regions within the rest of the track. Where no region of reasonable length is available, this could identify a track which is not of reasonable quality to use for automatic analysis. Such patterns of inconsistent features are not currently considered in this work, but may be considered in the future for manual revision of the object segmentation or tracking.

5.4 Experimental Validation for the Identification of Major Segmentation Errors

The experiment used to validate the identification of segmentation errors is based upon the comparisons of the four object features consisting of global MCR, upper MCR, lower MCR and fused MCR results based upon a self-comparison of 26 tracks from four people across two cameras, consisting of over 300 frames. These tracks are automatically analysed to identify frames with significant segmentation errors, when compared to ground-truth analysis performed by human experts. Examples of good segmentation and the clothing worn by the four individuals studied for this experiment are given in Figure 25. Of the 26 data sets, 5 were used for training the Gaussian likelihood functions of the non-matching *H0* or the matching *H1* data sets on a frame by frame basis. The remaining 21 tracks are used as a testing set for evaluation.

Figure 25: Four people of interest and automatically segmented masks of good quality

The results are given as a ROC curve in Figure 26, similar to the results in the previous chapters, showing the detection rate compared to false detections when compared to the human expert based ground truth. These results are given to compare each individual feature, as well as the fusion of the MCR based appearance features. The ROC curve clearly demonstrates that the fused MCR features can provide a significantly higher level of accuracy, whilst also limiting the amount of false detections.

Table 8 gives the probability of detection (*PD*) and probability of false alarm (*PFA*) for each of the MCR features analysed at the selected operating point compared to the expert determined ground truth. A more detailed analysis based upon the individual people who were tracked, as well as the overall quality of the tracks analysed indicates that this method works best with individuals who are not of a

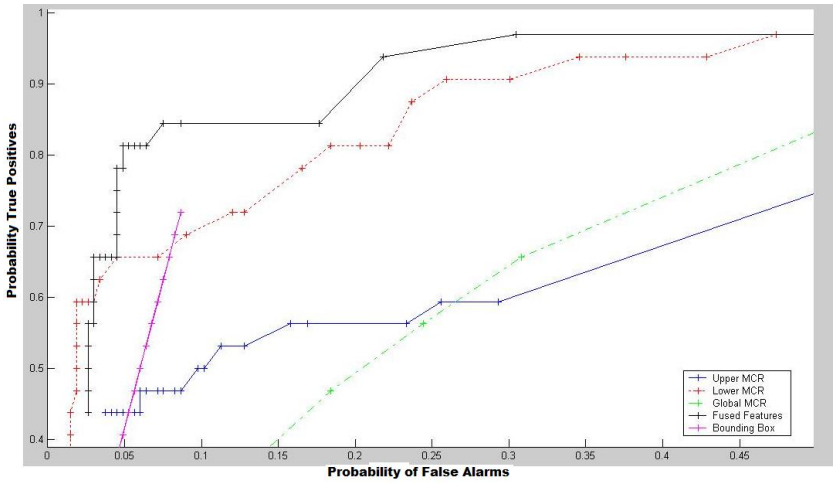


Figure 26: ROC curves of the height, colour and fused feature results

Feature	PD as %	PFA as %
Vertical Bounding Box changes	72	9
Global MCR	66	31
Upper MCR	53	11
Lower MCR	66	5
Fused MCR	84	3

Table 8: PD and PFA values of Bounding Box and MCR features for detecting segmentation errors

uniform colour, and where the overall quality of the segmentation of the tracked object is good; although the accuracy as shown is still high under less optimal conditions. The results also indicate that the use of upper and lower MCR features dramatically improves the ability of the system to detect major errors which may not be detected adequately with the global colour or bounding box analysis alone. The major limitation of the bounding box change ratio analysis is that gradually increasing or decreasing segmentation errors are hard to identify, especially where there are multiple concurrent major or minor segmentation errors. This creates a limit to the overall accuracy of such analysis that is not inherent in the fused colour features. The bounding box height also does not consider the pixels that may be incorrectly segmented through the middle of the individual. Large gaps in the individual that are not located at the top or bottom of the individual, or

do not split the individual into fragments may affect appearance or shape features without necessarily affecting the image height of the segmented blob.

Note that *PD* can be taken close to 100% if one can accept a *PFA* of approximately 30%. This operating point is of interest if a system looks to remove all frames with major errors, and the number of frames left by the selection procedure is still sufficient for later stages of processing. Such an operating point would be of interest where a system has high frame rates, and the tracks of individuals are reasonably long.

5.5 Discussion of Segmentation Error Identification

A number of factors ranging from the contrast of the person from the background, to occlusions and illumination changes, ensure that segmentation will often include a degree of errors, some of which might be very large. These errors could propagate into a number of subsequent tasks, ranging from feature extraction to tracking through matching of individuals across disjointed camera views, or accurate searching for the person in image or video archives. Other tasks could also be generated through a method similar to the identification of errors, such as creating a faithful pictorial summary of a tracked object. This pictorial summary could be made by using one frame, or possibly a few frames, where the object is not affected by large segmentation errors. If large segmentation errors can be identified and removed from the automated process, then such subsequent processes are likely to be made more robust and accurate.

This chapter has suggested how two different types of features can be analysed to identify frames in which major segmentation errors occur. The results indicate that almost all the errors can be identified for removal if enough false alarms are allowed. Even without accepting many false alarms, a significant amount of erroneous frames can be identified using either the fused appearance features, or the bounding box height changes. Although this chapter has focussed upon MCR appearance and bounding box height features, a variety of other features could be used, as long as they can be compared and remain stable along a track. For example a true height estimate could be used in the place of the bounding box as it should remain static regardless of the object size; however many such shape based features can also be dramatically affected by small segmentation errors, which are quite common. Such flexibility in the possible features is important as it can allow the system to more fully utilise the features that are already being used within a given surveillance system. For instance other appearance features, such as colours coded by path length [119], could also be compared on a frame by frame

basis, with the results reflecting the sensitivity of that feature to segmentation errors. Feature consideration is especially important where tracks are short and segmentation is noisy, as is the case with the data used here. By identifying errors this automated method can also give an indication of the overall feature quality for a given track, by indicating the rates of errors in that track. This additional information is provided at a minimum of computational cost where features that are already being extracted can be compared quickly.

The other important consideration is the acceptable amount of false detections of segmentation errors. Increasing the amount of errors that are removed is likely to lead to an increased rate of frames that are removed even though they don't have errors. This consideration is important as tracks that are of considerable length may not be affected, but shorter tracks or those where individuals are often partially obscured may be dramatically affected by high false alarm rates. Although this technique may reduce the amount of erroneous frames, efforts are still required to make features robust to the errors that occur with even the most advanced image processing techniques.

5.6 Summary of Segmentation Error Identification and Future Enhancements

This chapter has demonstrated that the effect of large segmentation errors upon features can actually be used to identify the frames where those errors occur. This chapter explored and compared two different features that could be analysed to identify segmentation errors. Although the fused MCR appearance features provided the most accurate, the exploration of the bounding box height feature also demonstrates that as long as a stable feature is chosen, then large segmentation errors can be identified. The process identifying these errors can be summarised as:

1. Extract the feature for each frame within the track, or a sizeable portion of the track.

2. Compare the features between each and every frame to evaluate their similarity.

3. Analyse the pattern of errors in the inter-frame similarities of the features to determine where frames are dramatically different, or are simply adapting slowly to conditions over time.

The results obtained from this process demonstrate that the effect of segmentation errors, like many other sources of errors, can be minimised by analysing their effects upon the features along a frame sequence. Although the fused MCR features achieved a segmentation error detection rate of 84% with only 3% of false alarms on the data analysed in this project, the results would seem to be dependent upon the features analysed. With a wide range of reasonably invariant features available, there is considerable scope to investigate the accuracy achievable with other shape or appearance features. The process could also be enhanced through the investigation of changes over portions of longer tracks in order to provide more timely identification of errors, allowing features to be utilised even before the whole track is available. The amount of frames required for analysis before errors could be identified would need to be considered carefully; however for reasonably well segmented tracks, the number of frames required could be as low as ten.

6 Height Based Robust Shape Feature

This chapter explores the use of height as a shape feature for tracking, matching, and identification of individuals through a wide area surveillance system. Height estimates, as a useful general identity descriptor, are conducted widely by the police for identifying suspects in their investigations, as well as in computer vision based identification [93]. Other shape related features, such as gait based stride length, have been proposed for human identification or matching individuals; however these often rely upon high resolution images, sometimes specific installations of cameras, and are often tested in simple laboratory environments or under a variety of assumptions. Thus, the achievable accuracy of these features in real surveillance systems is often not clear. As opposed to other features, height estimated from images has been used to a varying degree of success in various applications in complex scenes [7, 16, 19, 71, 73]. Indeed its limitations in distinguishing between individuals mostly arise from the limitations in height variance of the population [101], and the noise within the pixel based resolution of objects from which height is determined. To date little research has investigated reducing the impact of these limitations, even though it could be beneficial to other popular topics such as gait based identification.

This chapter proposes how to overcome the limitations of previous research into the estimation of height as people move within the view of monocular cameras. This is important in video surveillance as stereo or overlapping cameras are often not available, and other features such as motion information can become unreliable when there are large gaps in camera coverage. This chapter explores the use of the height feature in isolation to determine the maximum accuracy that might be achievable. When attempting to achieve accurate individual re-identification, it should be noted that the intrinsic discriminative power of height throughout the observed population is likely to be too limited. Instead this feature may be used as an accurate partial descriptor of an individual that could be combined, or fused with other features in order to obtain the overall desired accuracy. Whilst this chapter focuses upon the accuracy estimate of differing individuals, Chapter 7 investigates the usage of the height estimate feature as a component of a features based re-identification framework.

The chapter begins by providing an investigation of the background of the usage of height estimation for the identification or matching of individuals as compared to other possible shape features. Section 6.2 then describes how camera calibration can be used to extract estimates of an individual person or an object's height from a single image. Section 6.3 describes the technique developed to im-

prove the extraction of height estimates from a single camera view and how the height estimates along the track of an image sequence can be used to mitigate gait effects. Methods that can be used to statistically compare height estimates to both real estimates, or estimates from other tracks are described in section 6.4. The experimental verification of these methods are described in 6.5, where they are broken into the logical development steps based upon manual data, the initial automatically extracted data, and finally a large scale experimental verification. These results along with the identified limitations of the process are discussed in section 6.6. The chapter concludes with a summary of the height estimation procedure and future enhancements that could improve the accuracy of extracted height features.

6.1 Shape Feature Background

Shape features have been utilised successfully for object classification [6, 102] Shape has been exploited within many projects as a powerful feature for the classification of objects into groups or classes, which have the same type of shape or shape features. Indeed shape based template matching is used widely within machine vision applications in factory situations for quality assurance, with accuracy often being much higher than a human is capable of achieving. The accuracy of this process is achievable due to the objects being as identical as possible in shape. This similarity of shape within groups of objects is very powerful for classification purposes, but it limits the amount of detectable variation that could be used to identify particular instances of an object. This is exacerbated by the increased probability of errors in the segmentation around the edges of the object, making it difficult to determine if changes in the shape are part of the individual variation, or if they are just errors within the segmentation process. For example ground vehicles such as cars all tend to have the same streamlined shape with four wheels, a protruding bonnet, and doors down the side. Variations in the basic shape such as a tray back or hatch back, emblems and even the specific streamlining of the shape can be very useful to further classify the specific model of the vehicle; however they do not provide much information on the specific instance of the actual vehicle that could be used to identify it. Indeed identification of specific vehicles who belong to specific owners is so difficult from shape and appearance features that number plates are used to identify individual cars. Natural objects, as opposed to manufactured objects, tend to have a higher level of individual variation, also known as intra-class variation; however they also tend to have some degree of articulation making shape features less invariant both across camera views and

within a single camera view due to changes in the object's articulated pose.

Shape features which are used for the identification of individual objects are often used as a component of an object model, or to understand articulated object motion. Model based features are prominent in the literature for classification [47] to provide more accurate shape information for understanding human movement, which in turn may provide identification clues. General shape features that are not model based, are often dependent upon the relative orientation of the object to the camera, and the level of object articulation. For instance a person walking across a camera view could have a significantly different width to that same person walking toward the camera, even when they are observed at the same distance from the camera. Indeed low level Scale-Invariant Feature Transform (SIFT) features, and the colour based CSIFT variant [1], are widely used to obtain low level object shape descriptors for classification purposes. Few of these shape features remain invariant or at least quasi-invariant over even short periods of time and hence can not be used to identify particular individuals. Of the shape features that have been widely used for identification or matching purposes, object gait and height are the main two features that have shown promise. Stevenage *et al.* [105] suggests that people can be identified from their movements through an understanding of how their shape changes; however the complexity of the problem is very high with a limited range of accuracy using recent methods [103].

Height estimation has been used many times as a stable biometric shape feature for the identification of individual people [7, 19]. Height changes gradually as a person initially grows into adulthood, but then remains very stable throughout the rest of a person's life. Height is traditionally measured with a person standing next to a flat surface, where the measurement runs from the floor to the top of the skull. This measurement is a minimally invasive technique which requires the co-operation of the individual to stand straight and still. The two main limitations of height when observed from an image are the accuracy camera calibration, and thus of obtained measurements, and the ability of height to discriminate between two arbitrary individuals. The main factors that influence the accuracy of the height estimation obtained from a walking individual are the effect of gait on the height within any single image, the accuracy of the segmentation of the selected object, and the accuracy of calibration for the method of estimation, being either from a single or multiple overlapping cameras. A person's actions can also influence the observed height, such as whether they are walking or bending over, as well clothing factors, such as the size of soles or heels of an individuals footwear and the person's hairstyle or hat. Whilst improving camera quality and calibration techniques can improve accuracy of the measurements, the gait of an individual is

inherent to their movement. Attempts to mitigate for the change in height using the cyclical nature of this influence along a frame sequence can be used to increase the robustness of any estimates. Such estimates are also compromised when individuals are not standing or walking upright, although often these cases can be identified and removed from the estimation of the object's true height. The impact of inaccurate segmentation of an individual from the background is significant, but highly dependent upon the segmentation method used, and the complexity of the scene being observed. The impact of very high shoes and high hats are usually minimal throughout typical buildings, as they are not generally part of typical attire, and they often tend to remain stable throughout a surveillance session. Thus their impact is often negated when individuals are matched throughout a single surveillance session within a building environment.

Stereo cameras can determine the height of an individual directly from the accurate location of the point at the top of the head. Thus the height estimates are not necessarily affected even when segmentation affects significant portions of the rest of the body [19]. Single camera views, or monocular cameras, require the entire person to be extracted into a single blob as both the top of the head and the bottom of the person need to be accurately located to produce an accurate height estimate [7, 71]. Within both of these methods degradations in the accuracy of segmentation can lead to reduced accuracy which is inversely proportional to the size of the object within the image. That is to say the greater the number of pixels that comprise the object, the greater the degree of accuracy will be as each pixel error will correspond to a smaller percentage of the estimated height. Thus cameras that can obtain zoomed views of a person are likely to improve accuracy; however these are not widely available and it could be difficult to maintain camera calibration.

When considering the problem of extracting height measurements from an automated system, one first needs to consider the reality of existing surveillance systems. Although multiple overlapping cameras can mitigate segmentation problems, and possibly increase accuracy, the majority of camera coverage is by single cameras. Such views can also cover difficult areas such as stairs and ramps, which create many difficulties in defining assumptions such as the ground plane. Camera quality and typical object resolution are also important as such factors can limit the accuracy of height from quantisation errors, and also relate to how large an impact pixel level segmentation errors can cause. Consideration of many of these factors are very important to the development of a working system; however are not often considered in laboratory experiments.

The limited availability of overlapping cameras is very important, but not

widely discussed in the literature. Such single images do not provide in-depth information about how far away from the camera an object is; however this can be overcome by assuming that people will be walking touching the ground plane. This allows for a ground plane homography to be calculated, which determines the real coordinates on the ground plane of an individual. This in turn provides additional constraints on the position of the top of the head, allowing a height estimate to be obtained. This technique has been demonstrated in BenAbdekader *et al.* [7] where the top and bottom image positions of an individual are determined using the centre of the top and bottom of the bounding box. It is also important to note that although this work focuses upon humans as the object of interest, the height of other objects that are sitting on the ground plane could be estimated using a similar technique.

The usage of the bounding box of an individual as an estimate of the height of the object in the image is reasonably common and is often used as a part of other object blob statistics; however when those measurements are converted to real world coordinates, they can provide a true estimate of the height of a person [7, 16]. The height measurement within any single image may be more than a few centimetres higher or lower than their true height under a normal walking gait; however an average of these measurements has been found to remain stable [71].

Even when the accuracy of the overall estimated height is high, many people may be of a similar height, limiting the ability of this feature to discriminate between the individuals. Statistics on the height of people across a sample population of Australian people [101] indicate that men have a mean height of approximately 174.8 centimetres with a standard deviation of 7.1 centimetres, whilst women have a mean height of 161.4 centimetres with a standard deviation of 6.7 centimetres. These statistics published by the Australian Bureau of Statistics (ABS)[101] also suggest that the height statistics are roughly indicative of heights around the world, although heights through Asian regions tend to be approximately 5 centimetres lower. These height statistics indicate that the probability of two random people being of similar height is reasonably low; however not low enough to guarantee that reliable accuracy of height estimations will automatically translate into accurate discrimination between people. Thus although height may be accurate, it is only likely to be one useful component to discriminate between two differing individuals, and needs to be combined with other features to provide reasonable discrimination.

6.2 Obtaining Height Estimates Using Camera Calibration

Camera calibration forms the basis of extracting real world height measurements from either multiple or single camera images. A detailed description of camera calibration can be found in many sources, such as Hartley and Zisserman's book titled 'Multiple Camera Geometry' [46]. Often full camera calibration is suggested to provide the most accurate measurements possible; however Criminisi *et al.* [16] show that a partial calibration of an image which contains known reference lengths can be used to obtain measurements of reasonable accuracy. Such partial calibration are significant when only images are available; however the general video surveillance context does allow for full calibration either through an explicit optimisation of the real world positions and their image coordinates [113], or through one of the many recent methods that simplify the process [20, 97].

Camera calibration for a single camera view is not sufficient to obtain real world coordinates from image coordinates [113]. This is demonstrated in the camera calibration matrix given as equation **(65)** below, where the term *s* refers to the distance of an object from the focal point of the camera, also known as the object depth. The depth of object from the camera is usually determined using multiple overlapping camera views which combine multiple image coordinates, *u* and *v*, and their calibration matrices for the same real world coordinates *X, Y* and *Z*, which provides a unique solution for the equations.

$$\begin{bmatrix} su \\ sv \\ 1 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{65}$$

The depth of an object from a single image can be obtained by adding further constraints to the real world coordinates. This usually comes in the form of a ground plane homography [46], where the bottom of an object is assumed to be touching the ground, constraining equation **(65)** by using $Z = 0$. This homography is a subset of the camera calibration matrix which allows the two image coordinates to directly relate to the real world *X* and *Y* coordinates on the ground. As most people are observed walking upright and other moving objects tend to be rigid throughout typical surveillance scenarios, one can reasonably assume that the top of the object is directly over the top of the ground plane point. Other research into human movement and behaviour could also be applied to enforce this constraint for useful measurements; however this is not the focus of this thesis. This assumption can be used to constrain the real world position of the top of

the head using the *X* and *Y* values of the bottom point of the object. Expanding equation **(65)** for the image coordinates *u* and *v* gives:

$$u = \frac{p_{11}X + p_{12}Y + p_{13}Z + p_{14}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}} \tag{66}$$

$$v = \frac{p_{21}X + p_{22}Y + p_{23}Z + p_{24}}{p_{31}X + p_{32}Y + p_{33}Z + p_{34}} \tag{67}$$

These two equations relate the top of an object in image coordinates *u* and *v* to its real world coordinates, each with a single unknown, Z obtained from a single image. Rearranging equations **(65)** and **(66)** obtains the following formula which provides a single estimate of height of an object as constrained by its position on the ground plane:

$$Z = \frac{(p_{11} - up_{31})X + (p_{12} - up_{32})Y + p_{14} - up_{34}}{up_{33} - p_{13}} \tag{68}$$

Figure 27: Accurate location of the bottom point improves height estimates

This height estimation method is not new as BenAbdekader *et al.* [7] have previously used a similar procedure to try to identify people walking through a single camera based upon their height and the two gait characteristics of stride length

and cadence, or periodicity. Their height estimation was based upon automatically segmenting people from the background into a single blob. A rectangular bounding box was used to outline the object such that it encloses all of the object's pixels. BenAbdekader *et al.* [7] selected the bottom point of the object as the middle of the lower edge of the bounding box, with the top point being the middle of the top of the bounding box. This allows for fast and easy selection of the key top and bottom points; however Figure 27 shows that this might not be the best selection of top and bottom points. Significant improvement for the bottom point is achievable as the bounding box location does not account for the true positioning of the feet of a person within an image. This could distort the true ground plane position of the object as well as the image height of the object. Images where the head is not in the direct centre of the bounding box could also benefit from more accurate estimation. Such errors are likely to vary from frame to frame, and are likely to be in the order of a few pixels; however this could lead to a centimeter or more error in the estimation of an individual's height, depending upon the height of the individual in the image.

Although the image based improvement of the manually extracted key top and bottom points shown in Figure 28 is obvious, the main difficulty occurs with the automatic extraction of these points and the mitigation of gait effects. The bounding box measurements are available simply from the extraction of the object; however calculation of improved key points requires a detailed automatic analysis of the segmented object. This automatic analysis of an individual's silhouette to improve the ground plane location of a person was not found in the literature, so a detailed analysis of the approach is provided in the next section.

The major difficulty with estimating the height of an individual arises from the change in the height of the top of a person's head as they walk. Increasing the accuracy in estimating the height actually leads to an increased variation in the height of the individual, shown in Figure 30 in the following section. The periodicity of these estimates are also noticeable indicating that a statistical analysis such as the average of the height estimates along a track is likely to produce a more accurate estimate of the true height. Although the periodicity of measurements is obvious, significant errors can occur due to error in segmentation of an individual from the background. Indeed such errors are recognised as common throughout all segmentation methods [96] of an individual and would need to be further explored before reliable periodicity measurements could be used.

The key steps for the automatic estimation of a robust measurement of an individuals height begins with extracting estimates of the height from each frame, then removing errors before using statistical analysis to provide a robust estimate.

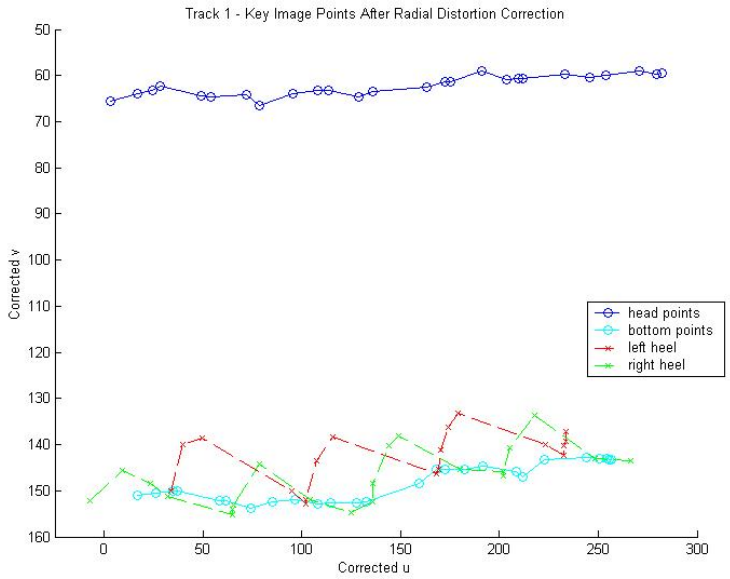


Figure 28: Manually identified key image points for the track of one individual

This can be summarised as the following steps:

1. For each frame
   (a) Segment the individual from the background
   (b) Automatically estimate the key top and bottom point locations within the image
   (c) Estimate the location of the individual on the ground plane
   (d) Use the ground plane location estimate in equation **(68)** to obtain a height estimate
2. Remove height estimates from frames with major errors, such as those from incorrect segmentation
3. Statistically analyse the set of height estimates from each frame to remove outliers
4. Statistically extract the robust estimate of the individual's height from the data without outliers

This process is likely to be reliable enough to provide a description of an individual's height with a similar or greater level of accuracy to those provided by the police when searching for a suspect. Thus it may be useful in a number of applications including non invasive biometric security and general surveillance applications.

6.3 Improved Automatic Monocular Height Estimates

Whilst the previous section outlined a method that produces an estimate of a person's height, the bounding box used does not necessary best represent the top and bottom of an object. Even with perfect segmentation of an object, the top of the head is does not necessarily lie directly at the centre of the bounding box, nor is the centre of balance at the bottom of the object always in the middle of the bottom of the bounding box. Indeed figure 27 shows that this is not the case. Many other objects might not receive much improvement in location over the simple bounding box measurements; however it is the highly articulated human motion with moving arms and legs that allow for these discrepancies from the bounding box based positions. Although many research areas from gait analysis to human behaviour modelling do perform an analysis of body part location and movement to a greater or lesser degree, most of this research is aimed at inferring more information about the motion of the person, rather than their biometric such as height. This section describes how silhouette curvature techniques can be used to more accurately identify the location of the head and feet, which in turn improves the location of the key top and bottom points, and the overall height estimate.

In order to improve the accuracy of height estimation, the position of the top of the head and the feet position needs to be determined as precisely as possible from a monocular view. Due to the typical resolution of surveillance footage, an improvement of even a small number of pixels in image height estimate can translate to a significant difference in height estimate. This improvement is shown in Figure 27 above, where the bottom centre point of the person is shown as a more accurate image position when compared to the bounding box. The feet positions are found using a k-curvature technique [35] after the object has been segmented from the background. This segmentation is currently done using a modified Pfinder [116] method, where colour based morphology [122] is used to close internal object gaps and to join body parts that have been segmented into separate blobs. When the object is segmented or joined into a single blob, then its key points can be identified from silhouette curvature. The k-curvature technique follows the chain of silhouette pixels and determines the curvature of the silhouette

at each pixel based on three pixels along the curvilinear coordinate, x_1, x_2 and x_3:

$$k = tan^{-1}\left(\frac{x_1 - x_2}{y_1 - y_2}\right) - tan^{-1}\left(\frac{x_3 - x_2}{y_3 - y_2}\right) \tag{69}$$

If $k < 0$ then one can use $k = 2\pi + k$ to ensure $0 < k < 2\pi$.

This equation allows for an analysis of the curvature around the silhouette of an object such that key points can be accurately located. These key points are found by local minima and maxima in the level of curvature as these relate to the extremities of the object. These key points can also be analysed with respect to their location within the object. For example a high curvature region at the top of the object is likely to represent the head of the object, especially where the assumption of the individual merely walking through the scene holds. This region is shown at the start and end of the curvature values in Figure 29. The second key points of interest are those areas of high curvature near the bottom of the object which correspond to the feet of the object. These values can be more difficult to analyse due to the articulated motion of the feet during walking. Thus in any frame one or two feet could be observed, and the high curvature points of these feet could relate to the toes or heels of an object. Analysis of high curvature points in the middle of the object could be used to located the hands of an object; however currently this research does not use this information.

Analysing the k-curvature has found little difference in identifying the head point location $h(u, v)$ than simply using the midpoint of the highest row of object pixels of walking individuals. Thus one can use the former as it is faster to calculate. It is important to note that this point is not necessarily same as the midpoint of the top of the bounding box as outstretched arms or legs could also influence the width of the bounding box. An optimum location of the bottom point $b(u, v)$ is more difficult because of the articulated motion of the two legs being important to the centre of gravity and hence the optimum location. One can identify a bottom point location by averaging the location of the feet positions, which are found as the high curvature regions near the bottom of the object. The method for deriving this point is outlined in the following panel:

$if u > (fract \times obj\_height)$
$if(dist(ka(u, v), kb(u, v)) < th)$
$then foot(u, v) = (ka + kb)/2$
$else foota(u, v) = ka(u, v) and, footb(u, v) = kb(u, v)$
$and b(u, v) = (foota(u, v) + footb(u, v))/2$

where obj_height refers to the person's pixel height, and $fract = 0.7$ in our

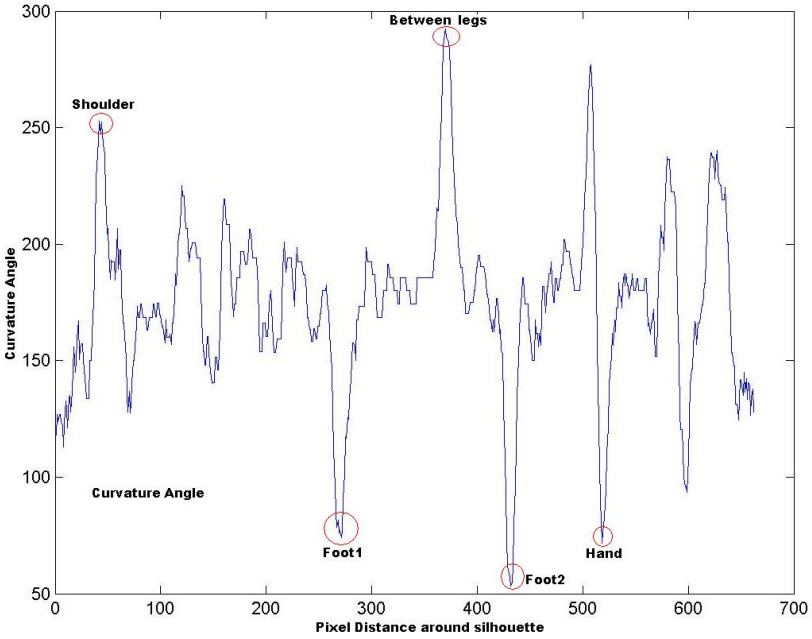


Figure 29: Curvature values for a single individual showing curvature key points beginning at the top left most object pixel

experiments. This excludes the top 70% of an object, measured vertically from the top of the object, from being considered as potential feet.

This panel shows how a bottom point can be found by assuming that areas of high k-curvature within the bottom 30% of the object are likely to correspond with an individual's feet, or more particularly their toes or heels. This assumption holds when there are no other objects near the feet of the object and the feet are reasonably well segmented as the other body extremities such as the arms or head will not protrude so low for a walking individual. Although shadow removal still retains some problem, many techniques are available to minimise this problem [17, 22, 56, 114, 125]. If the two distinct points of key curvature in the lower portion of the object are located close together around the silhouette, then they are likely to correspond to the heel and toe of one foot of the object, and can be averaged to produce a foot position estimate. Otherwise, the single significant curved area is used as the foot position estimate. Where two areas are found with high curvature in the lower portion of the object, but are relatively far apart on the silhouette, then they are assumed to represent the two separate feet, and are analysed accordingly. The two feet estimates can then be found and averaged to provide

an estimate of a midpoint at the bottom of the object $b(u, v)$ as shown in Figure 27. This image plane position can then be used to find the ground plane position $b(x, y, z)$ through the ground plane homography transformation found by setting $Z = 0$ in equation **(65)**. The usage of a bottom point tends to produce a better height estimate than simply using the middle of the bottom of the bounding box, especially as it better conforms to gait effects so they can be removed statistically. This method may also produce a better estimate of the ground plane position of the object, though this has not been extensively tested.

The automatically extracted head *h(u,v)* and bottom positions *b(u,v)* can then be converted from the top left image plane coordinate system into real world ground plane coordinates using camera calibration matrix given as equation **(65)** [71]. This produces an estimate of the height of the segmented object from that single frame. These estimates are obtained along each of the frames in the image sequence, as can be seen in Figure 30. It is important to note that these estimates are also influenced by gait along the frame sequence; however under reliable segmentation and high frame rates, relatively stable sinusoidal pattern emerges for the measurements. When the sample rate of the person walking is low due to low frame rates, such as that used in our surveillance system experiments, this periodicity can be difficult to see, especially in the presence of minor segmentation errors. The presence of this sinusoidal pattern would suggest that a simple average is likely to produce a stable height estimate.

A simple average of the height estimates along the frame sequence is adequate for manually estimated key points; however due to occasionally large segmentation errors, outlier elimination techniques such as those described by Mosteller and Tukey [81] should be applied to ensure that these errors are removed before the mean is calculated as a robust estimate of the individuals height. Such errors are also likely to be removed through the application of the segmentation error removal technique detailed in Chapter 5. The standard deviation of the height estimates after outlier removal may possibly be used as a gait estimation feature; however such a feature would also be subjected to possible variations through footwear and surface types. The effects of these variations have not been fully investigated to determine the level of this features invariance. It is also important to note that this step of outlier removal and gait mitigation is just as important for height estimation from stereo cameras as it is from monocular camera views.

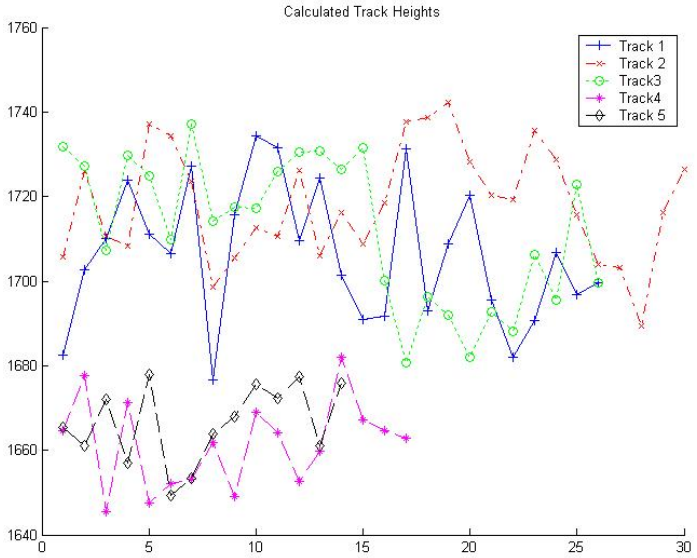


Figure 30: Height estimates for 5 tracks demonstrating the manual height estimates

6.4 Statistically Comparing Height Features

This section outlines a technique that can be used to compare height estimates found using either the method for monocular camera views as outlined in section 6.2, or the method using stereo cameras given in [46]. Common practise is to directly compare the robust estimate of one person's height to another. Indeed forms to fill out to report suspects involved in alleged criminal activity often provide height as a descriptive feature [93] and sometimes suggest a 5 centimetre range for height estimates [94]. This is perhaps due to the difficulty people have in estimating height to a greater accuracy than this. Such estimates are also given in police media reports, along with other features like hair or clothing colour, to describe individuals of interest. These height descriptions would suggest that estimates within approximately 5 centimetres of each other are possibly obtained from an individual or individuals of similar height, whilst estimates that differ by more than 10 centimetres are unlikely to be obtained from the same individual. For the case where the height is automatically estimated from video footage of an individual walking upright through the scene, such a height comparison should also hold. This comparison could be useful for police evaluating suspects seen in

video surveillance; however it does not consider the inherent accuracy limitations within the estimates obtained from video cameras, nor whether the individuals behaviour might influence the accuracy measurements, such as exaggerated gait effects, arm waving, or long periods of remaining motionless. Such accuracy limitations occur due to the relative amount of height represented by each pixel, as well as the fluctuations due to minor segmentation errors. These accuracy limitations can cause extra fluctuations in the height estimations, which can also be statistically analysed, but might bias the estimate to some degree.

An analysis of the statistical uncertainty of the height measurements can be incorporated into a measurement of the similarity in height of two objects, especially where they are obtained from two tracks obtained within the surveillance system. A simple measurement of the standard deviation of the frame level height estimates could be used to add information as to the variation of the estimate from a single track, such as where a height estimate is compared to a real world individual. When the height estimates are obtained from two tracks in the surveillance system one can define a height difference measure to use rather than simply combining the mean and standard deviation to define a region of overlap. The height difference measure can be calculated as a vector Hd containing the absolute difference in estimated height between each frame in track A with each frame in track B, calculated in a pairwise manner. This measurement could also be obtained using a linear comparison of a single frame in each sequence; however the periodic sinusoidal nature of the changes in height due to gait could lead to a significant offset being measured where the tracks are not aligned due to their gait period. The pairwise comparison can be used because although it is of quadratic complexity, it is actually of a very lightweight computational complexity. Thus this comparison is likely to overcome the problem of gait based height measurement alignment, whilst not compromising upon the real-time performance of an system.

The proposed height difference vector Hd thus contains a series of measurements representing the difference in height of the individual in track A and the individual in track B. A similarity measurement s_H can be defined by statistically analysing Hd using:

$$s_H = \frac{\sigma(Hd)}{\mu(Hd)} \tag{70}$$

This similarity estimate combines both the mean value, $\mu(Hd)$ as well as the standard deviation $\sigma(Hd)$ within height differences and is the reciprocal of the well known standardised distance. This aims to provide a measurement that not

only relies upon the mean difference in heights along the frame sequence, but also incorporates the inherent variations within the height estimates obtained through the two tracks. This measure promises to be more reliable in the presence of extra noise, which might occur in areas of poor segmentation. When this measure is considered with the limitations due to the statistical possibility of the similarity of many individuals in any observed population [101], it becomes evident that though this measure is biometric, it should be used in combination with other features to provide accurate discrimination.

6.5 Experimental Verification of Height Estimation

This section explores the experimental verification of the height estimation method. The results presented in this section are based upon the analysis of two differing experimental sets. The first dataset contains 15 tracks of a single individual obtained from image sequences with a resolution of 293 x 214 pixels in two cameras, which were extensively tested to compare the manual and automatic height estimates. A second experiment was conducted on 26 tracks obtained from four individuals across two cameras, giving over 300 possible comparison combinations. The ground truth height difference of these individuals ranged from 5 centimetres to 30 centimetres. The automatically extracted height estimates were obtained for the purposes of matching tracks of an individual throughout the system.

6.5.1 Height Experiments Comparing Manual and Automatic Height Estimates

A comparison between the manually estimated ground truth and automatic estimates of a single individual's height was conducted upon the first dataset. This contained data from 15 separate tracks in image sequences with a resolution of 293 pixels x 214 pixels. Such resolution is due to the acquisition system (a video surveillance system in operation at the University of Technology Sydney) and can be regarded as fairly low resolution. The automatically extracted height estimates are compared to a set of 3 manually extracted tracks. The physically measured ground truth was found to be 171 centimetres using traditional measurement techniques. The manually analysed tracks found the individual's height based upon manually identified head and bottom points h(u,v) and b(u,v) to be on average between 1706 millimetres and 1719 millimetres. This manual data also had an average standard deviation of 16 millimetres. This standard deviation in the data

is due to rise and fall of the head as a part of a person's gait, rather than measurement errors, and is clearly very close to the ground truth measurement.

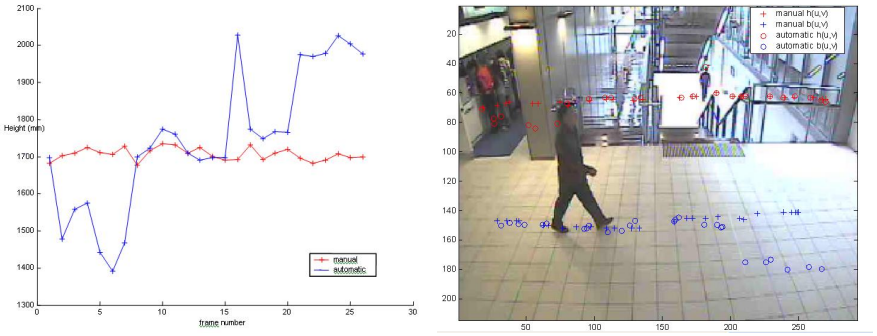


Figure 31: An individual's key points and height estimates using automatic and manual techniques on the poorly segmented track 13

The automation of finding of the head and bottom points with object segmentation and object analysis steps has introduced extra sources of error in the height estimations. Figure 31 shows the worst case scenario where a significant portion of the track is poorly segmented leading to a significant number of incorrect height estimates. It shows that the automated technique does produce more erroneous height estimates on a frame by frame basis. Significant segmentation errors occur at the start of the track where the face appears similar to the background and is lost, whilst errors at the end of the track show how shadowing effects can change the estimated height if they are not adequately removed. This analysis is important as these errors are actually typical from this camera view where portions of the track are known to produce large systematic segmentation errors for the tracks obtained. It can also be seen by the entry for this track, track number 13 in Table 9 that once these automated height estimates are analysed with outlier correction techniques over the track, they still produce very similar estimates to the expected ground truth of approximately 1710 millimetres. The results given in this table are obtained before the segmentation error removal techniques described in Chapter 5 are applied.

Table 9 demonstrates the difference in the statistics of automatic height estimations over 15 tracks of the same individual across 2 cameras under differing illumination conditions. It also compares the usage of a simple average, a robust average, a robust median, and a Tukey weighted mean [81] on producing a stable height estimate. The robust average and median iteratively remove those points

Table 9: Auto height estimates of a 1710 mm individual over 15 tracks

Track	Mean	Robust mean	Robust median	Tukey mean
2	1418	1418	1434	1434
15	1424	1468	1464	1485
3	1466	1466	1471	1489
13	1437	1513	1484	1509
4	1519	1519	1505	1520
11	1604	1683	1673	1667
1	1565	1668	1679	1660
12	1603	1679	1695	1670
14	1536	1710	1704	1708
10	1708	1708	1713	1706
9	1581	1703	1714	1689
7	1609	1674	1717	1670
5	1720	1755	1748	1759
8	1861	1917	1813	1888
6	1589	1815	1821	1779

that lie the furthest from the current median value if that distance is more than two 2 standard deviations. Because of the high number of erroneous regions in some of the data, the standard deviations of the original data can be as high as 300-400 mm. After outlier removal the standard deviation tends to approach 150mm, with many tracks being as low as 70 mm. It is important to note that these variations are still high due to the poor quality of the video data being used. With the application of segmentation error removal this level of variation can be lowered further, leading to improved height estimate.

The robust median approach seems to produce the result that is closest to the manually extracted value range if segmentation error identification is not performed; however even then only 8 of the 15 tracks are within 50 mm of the ground truth height of 1700 mm. It is important to note that these tracks include the height estimates of each and every frame that is related to an object as it is tracked within a camera view. This can include significant regions where only part of the object is in camera view, such as where an object is entering or leaving the view, or where the object height is a few pixels (in the order of 35 pixels high). These cases are likely to introduce extra errors in the data, which makes the statistical robustness of the true height more difficult to estimate. Weighting of the importance of the height estimates based upon the pixel size of the individual may also be possible

if a weighted average was used; however such concepts have not been explored in this thesis.

These results confirm the idea that this technique will have regions in the view that are less likely to be accurate, and that the overall accuracy of this process can be very poor if these errors are not addressed. Regions where a person has not entered the full view of a camera can be easily identified as the object will be touching the image edge; however other regions where the person's head or feet are occluded or are poorly segmented will be much more difficult to find. Detections of occlusions has been the subject of some work [24, 104]; though this is currently only approached from an object tracking view, rather than improving feature robustness. Regions where the person has a small number of vertical pixels are also of concern. If a person of 1700 mm height appears 100 pixels high, then the pixel-based height resolution is approximately 17 mm. At 25 pixels high, the same person has a pixel-based height resolution of 68 mm. Thus such image regions are unlikely to provide very accurate height estimations, and even minor segmentation errors will become significant.

A second experiment was performed upon the same dataset of 15 tracks from the individual of approximately 171 centimetres. This time additional steps were taken to improve the results by automatically identifying regions where the individual is not in full view, and applying segmentation error detection through analysis of appearance features as described in chapter 5. The results of removing these errors from the tracks analysed demonstrates that significant improvements can be made even for very noisy data with significant regions of error, so long as more than half the track is reasonably reliably segmented. With 13 out of the 15 tracks now being within 50 millimetres of the ground truth, and the other two tracks having a very high proportion of frames segmented poorly. In fact these tracks are identified as being very unstable and unreliable with the majority of the frames likely to have varying levels of error when the automated segmentation identification process is applied.

6.5.2 Height Experiments Using a Larger Dataset

A larger experiment was devised to analyse the height estimation results from the comparison 26 tracks from four people across two cameras, giving a total of 325 possible comparison combinations. Of these, 42 comparison combinations are used as training data and the remaining 283 for testing. Height differences between the individuals range from approximately 5 centimetres to 30 centimetres. These results were analysed based upon the similarity measure given in equation

(70). The training set was broken into known matching and non-matching track pairs and used to determine likelihood functions for the matching $H1$, $P(s_H|H_1)$, and non-matching $H0$ sets, $P(s_H|H_0)$. By modelling the matching and non-matching cases in the training set as Gaussian curves, these probabilistic likelihood functions can be calculated. These functions are similar to those used in determining the similarity of MCR features in Chapter 3.

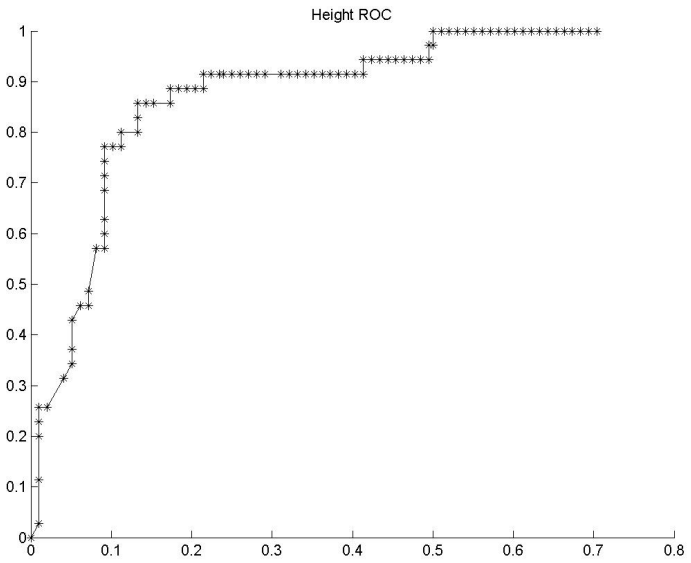


Figure 32: ROC curves of the height feature results

From these likelihood functions two methods can be used to determine if the two tracks are of a matching height. First an optimum threshold can be determined based upon the statistics of the training data to minimise the overall error. This threshold could be adjusted to allow for a higher probability of matching, or lower levels of false matches, but is ultimately used to make a decision about matchability. When this threshold was used upon the evaluation dataset it achieved a probability of detection of 91.4% with 23.5% false alarms. Alternatively the evaluation can be made based by using the observed s_H to determine the maximum of $P(s_H|H_1)$ or $P(s_H|H_0)$. For the single feature case this method will produce the same decision result as utilising the threshold. It can also be adjusted to allow for different levels missed matches or false matches by multiplying one of the probabilities by a biasing term. Unlike the threshold decision the probability

measurements can be used in a non-decision based feature fusion process later in the system. The results for the accuracy of the height feature are given as an ROC curve in Figure 32. They indicate that the height feature alone can produce some discrimination between individuals of reasonably differing height; however as with other individual features, they are likely to be best used in a general scenario when combined with other complementary features.

6.6 Discussion of Height Results

Due to the sensitivity of the height estimation to segmentation errors, especially when they are large, techniques need to be used to increase the robustness of the final estimates that are used. Improving this robustness to erroneous measurements has traditionally been performed through the usage of statistics for outlier removal [81]; however chapter 5 outlines an object feature based method to identify errors through the changes in features along an object's track. These methods aim to identify frames, or the data obtained from those frames, which provides erroneous estimations in order to remove those data points from the overall estimation. Once these errors are identified and removed, then statistical measures of the dataset, such as the mean and standard deviation, can be applied to the data to provide the robust height estimate and an indication of the errors in the measurements obtained.

The analysis of an object's shape features has been used in a variety of computer vision processes including video surveillance. These shape features have been used as a basis of many quality assurance techniques in the manufacturing industries, as well as identification of parts through shape template matching. Shape features have also been used widely in the classification literature, but have featured minimally in video surveillance. This is due in a large part to the articulated motion of humans, who are often the primary object of interest. This articulation leads to few invariant shape features from one frame to the next. The previously low resolution of images and low quality of segmentation have also lead to less reliable shape information for analysis.

The main shape features proposed in the video surveillance literature for the identification or matching of an individual are gait features, and less prominently height estimates. Gait identification is currently a hot topic, especially since the formulation of the gait identification challenge [103]. Although this feature has been much studied, current methods often use stereo cameras and constant frame rates, which are not widely available in current video surveillance, and they do not investigate the effect of surfaces and shoe types upon the gait of an individual.

The results of this feature alone also do not seem to be as high as those found with the proposed height comparison techniques, leading to the question of just how much inherent variation there is in this feature between individuals. Facial feature based identification is also a well studied topic with a variety of promising results initially published. Unfortunately the accuracy of many of these techniques are based upon specialised hardware such as multiple camera systems, or using zoomed facial views to improve facial features. Although such techniques could be very useful where they are available, current surveillance systems do not always include the hardware to make these features very useful.

Height estimation has also been proposed as a shape feature for identifying individuals [7, 16, 19]. The results have indicated it as a useful feature for identifying an individual; however these methods have not addressed how height estimation is affected by errors from segmentation, or how these estimates can be made more robust through statistical analysis along an individuals track. This chapter has outlined the work done in this project to investigate and mitigate these effects, showing that height estimation can be made more robust to errors and improving its discriminative ability to enhance a general track matching system. Unfortunately there are still likely to be many regions where height estimates are unreliable. These areas generally consist of areas where the ground plane is poorly defined, such as stairs or escalators. Other areas such as where objects are partially occluded are also problematic for monocular height estimates. Finally the use of PTZ cameras may be able to provide accurate height estimates, so long as the camera calibration through the camera movement can be maintained or reconciled. Where these errors occur, the height estimate may become very unreliable and should not be used to attempt track matching.

6.7 Summary of Height Feature and Future Enhancements

Height estimates are a common tool used in a variety of areas, including police descriptions, for identifying an individual. It is traditionally combined with other features such as clothing colour and ethnicity in order to improve the uniqueness of the description. Statistics on the height of people across a sample population of Australian people [101] indicate that men have a mean height of approximately 174.8 centimetres with a standard deviation of 7.1 centimetres, whilst women have a mean height of 161.4 centimetres with a standard deviation of 6.7 centimetres. These statistics are roughly indicative of heights around the world, although heights through Asian regions do tend to be up to 5 centimetres lower. These height statistics indicate that the probability of two random people being of sim-

ilar height is reasonably low; however not low enough to guarantee that reliable accuracy of height estimations will automatically translate into accurate discrimination between people. Also compared to colour based features, height estimates have a number of restrictions where the ground plane is ill defined, or individuals in monocular cameras are partially occluded. This confirms that height needs to be combined with other features to provide a more discriminative description.

The key steps for the automatic estimation of a robust measurement of an individuals height begin with extracting estimates of the height from each frame, then removing errors before using statistical analysis to provide a robust estimate. This can be summarised as the following steps:

1. For each frame
   (a) Segment the individual from the background
   (b) Automatically estimate the key top and bottom point locations within the image
   (c) Estimate the location of the individual on the ground plane
   (d) Use the ground plane location estimate in equation **(68)** to obtain a height estimate
2. Remove height estimates from frames with major errors, such as those from incorrect segmentation
3. Statistically analyse the set of height estimates from each frame to remove outliers
4. Statistically extract the robust estimate of the individual's height from the data without outliers

This estimate for an individuals height can then be compared to a real world estimate of the person's height, or can be statistically compared to another height estimate. This statistical comparison can be performed using equation **(70)** based upon the mean and standard deviation of the frame level differences in the height of the individual. High levels of similarity are thus found where the means are close to each other, and the standard deviations are low. This is confirmed with the results of the experiment showing a matching accuracy as high as 91.4%, although the rate of false alarms at 23.5% is too high for it to be used in isolation.

7 Fusion Methods and Results for Combining Robust Features

The features available from typical existing video surveillance systems are not necessarily biometric in nature, nor of sufficient resolution to provide significant discrimination between individuals alone. This chapter therefore investigates a framework to combine multiple features to increase the level of discrimination between individuals. There are many possibilities to combine or fuse such features together [37, 58]; however it is important to note that even the best fusion methods require complementary features to improve its results. In the building surveillance scenario presented in this thesis, improvement in overall accuracy can be achieved through improvement in the accuracy of matching an individual, or through improved discrimination between differing individuals, or through improvement in both. Thus this chapter looks at applying the main results from the literature on an ensemble of classifiers for this application. Whilst this is a well researched topic, the contributions to the field of this work include:

1. The analysis of the likelihood functions specific to this project.
2. The modelling of optimal thresholds from those functions.
3. Developing a method to determine which of the investigated surveillance based features are the most effective. This includes studying which features provide the most accurate results, as well as how to determine the complementary nature of these features to improve the overall accuracy and investigating computational speed through the identification of minimally useful features.

The chapter begins with a short background to classifier-based fusion, allowing the focus of the chapter to remain upon the improvements and results of the system performance. Section 7.2 then outlines how to apply the fusion techniques in the temporal domain to integrate features to the same temporal level if necessary. The system results are then given in Section 7.3. These results begin with a validation of the likelihood functions used for the matching and non-matching cases for the features used. The broader system based results are then presented, where they explore how each of the features adds information into the overall fusion process. The chapter concludes with a summary of the fusion method which has been found to provide the most accurate results for this surveillance scenario.

7.1 Classifier-based Fusion Background

Much research has investigated the use of Bayesian fusion methods as an ensembles of classifiers for a collection of features [37, 58]. The application of Bayes theorem shows that where features are independent, the application of a product rule provides the desired joint probability, whilst an averaging rule is to be preferred where features are dependent. It is worth noting that measuring or estimating the degree of dependence between features does not always prove obvious. [37] has also demonstrated that the highest results can be theoretically achieved by using weighted average fusion, where each classifier is assigned a weighting based upon its performance and reliability. Unfortunately the performance of the system can be substantially impacted by incorrect weights, leading to worse results than if no weighting was used at all. This is an especially important consideration in an environment where the reliability of features can change over time due to changing error patterns. Where such changing reliability can be estimated, weights can be updated and used effectively; however such a process is often difficult in practise.

Bayesian theory provides the basis for the fusion framework that is presented in this thesis to combine the features within the proposed system. The chosen features for inclusion in this framework are the appearance based upper clothing MCR (UC), lower clothing MCR (LC), the global MCR (GC), and the height estimate (H). Although these four features are the only ones currently included, it is easy to see how the method could be extended to include other features as they become available and reliable. These extra features would become another classifier in the ensemble, and should therefore be included through the addition of an extra term, or terms, in the matching or non-matching equations. As the proposed system is reliant upon features that can become less reliable as aspects like illumination change, identifying and updating the correct weights would be difficult, and thus they may detract from the overall system accuracy. It is also important to consider the level of time integration at which the features are robustly available, as some features are available at a frame level, whilst others like height similarities are only reliably available at a track level. An exploration of the integration of information at different levels is provided in section 7.2.

The features fused is this section are considered to be at an equivalent level of time integration, which is considered here to be the track level. Thus the results presented here indicate the final probability of the matching of two tracks. The features can also to be largely independent of each other as they rely upon large components of differing data, even if there is some overlap between features. This

overlap exists in the narrow band component of the spatial MCR features with the global MCR feature; however the level of overlap for each of the spatial MCRs is only approximately 20% of the global MCR. Given these considerations, [37] demonstrate that the optimal theoretical fusion can be achieved at a track level using:

$$P\left(H0|s_H, s_{UC}, s_{LC}, s_{GC}\right) =$$

$$B\left(P\left(s_H|H0\right)P\left(s_{UC}|H0\right)P\left(s_{LC}|H0\right)P\left(s_{GC}|H0\right)\right) \tag{71}$$

$$P\left(H1|s_H, s_{UC}, s_{LC}, s_{GC}\right) =$$

$$P\left(s_H|H1\right)P\left(s_{UC}|H1\right)P\left(s_{LC}|H1\right)P\left(s_{GC}|H1\right) \tag{72}$$

It is important to note that these equations use $P\left(s_x|H0\right)$ or $P\left(s_x|H1\right)$ rather than the similarity values that have been derived in the previous chapters. Also B in equation 71 is given by $P(H0)/P(H1)$ in Bayesian classification, but can be set to other values to achieve a different bias of missed detections and false alarms, depending upon the desired results. Whilst a simple threshold can be chosen for a single feature, the multiple feature space becomes more difficult to analyse in either the similarity measure space, or a probabilistic statistical space. For this analysis, it is important to note that the similarity measures are very feature dependent. Thus distances in one features similarity space may not relate well to distances in another features similarity space. When a statistical analysis of the expected similarity values for matching or non-matching individuals is performed, it can be used to apply probability models that represent the probability of a given feature being either non-matching $H1$ or matching $H0$. These statistical probabilities are important as they are directly comparable between each of the features, and their validity for any given feature can be determined by comparing the feature similarity values with the probability curve.

In order to obtain the statistical probabilities for use in this framework, a training period with a known number of matching and non-matching individuals is needed. This training period provides the samples to build the statistical models of each case. The samples can be modelled using many functions, with Gaussian curves being common in the literature, as Gaussian functions only have the mean μ and σ for each of the matching $H1$ and non-matching $H0$ cases needing to be calculated using simple statistics. These parameters can be used to recreate the $P(s_x|H0)$ and $P(s_x|H1)$ curves as shown in Figure 33 and even estimate the optimum threshold for the single feature where the curves overlap. Although this threshold is not so useful for fusing multiple features, it can be used to evaluate the

usefulness of this Gaussian statistical model as it should also model the expected changes in error patterns when the threshold is adjusted. Such an evaluation is performed in section 7.3.1 to explore the results.

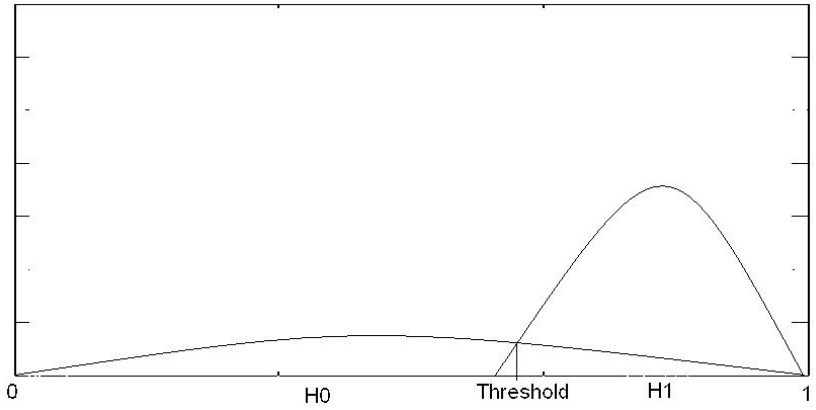


Figure 33: Example $P(s_{UC}|H0)$ and $P(s_{UC}|H1)$

A physical interpretation of $P(s_x|H0)$ and $P(s_x|H1)$ for colour-based features is as follows: $P(s_x|H1)$ could be described as the pdf of a Gaussian random variable, u, accounting for the variations of illumination and view on single objects. In turn, $P(s_x|H0)$ could be described as the pdf of the sum of two random variables, u and v, with u the same as before and v describing the variations in intrinsic colours between objects, that can still be assumed Gaussian but with much greater variance than u. Given that the sum of two Gaussian variables is still Gaussian, $P(s_x|H0)$ is Gaussian, too. However, given that its variance is large, its exact shape becomes less relevant. Similar considerations can be made for features of other nature.

The proposed fusion method is given as Equations **(71)** and **(72)** above to combine the information from all of the features. For each feature, including additional features that might become available, a probability term is required to represent it within the ensemble. Thus the modelling process is required for each feature that is to be used.

7.2 Classifier-based Fusion for Integrating Features across Differing Time Scales

The level of time integration at which the feature similarities and their equivalent likelihoods are obtained is also important to this process. As mentioned in previous chapters, the level at which a feature is obtained could be at a single frame level, across a window of frames, or generated from the entire track. Whilst some features are available at multiple levels, others such as the height similarity may only be robustly available at a track level, and thus cannot be fused with frame or window level features. One of the main considerations with the differing time scales for features is that multiple MCR feature similarities are available across a set of windows, whilst a single similarity is available from height estimation. The appropriate fusion of these features is to obtain a single similarity for each feature to be fused with the ensemble of classifier mentioned in the previous section. This idea was investigated with respect to the time integration for MCR appearance features in Section 3.4.1. This investigation demonstrated that either an average of the similarity values, or a threshold based decision fusion process could be used to determine the track level results. Whilst the decision based approach differs significantly from the track level Bayesian fusion described above, the averaging of similarity measurements is not entirely consistent with the usage of statistical likelihoods either.

Where features are to be fused to create a higher level of time integration, a new form of fusion is required. The fusion required here is distinct from both that required to fuse different features and fusing feature information to make it more robust. It differs from fusing multiple features in that the similarities are obviously dependent as they are derived from the same information from the same object. It differs from increasing robustness as it aims to combine the features to create an estimate of the feature similarity at a higher level of time integration. There is also no reason to expect one probability measurement to be more reliable than another, therefore weighting is not required for this fusion step. Thus the likelihoods of a feature similarity, s_x, for each frame can be combined into a higher level of time integration, $P^{'}$, across N instances using the Bayesian average fusion rule:

$$P^{'}(s_x|H0) = \sum_{i=1}^{N} P\left(s_{x_i}|H0\right) \tag{73}$$

$$P^{'}(s_x|H1) = \sum_{i=1}^{N} P\left(s_{x_i}|H1\right) \tag{74}$$

This provides a powerful addition to the fusion framework for combining features to a common level, such as a window of frames, or a track level. Ultimately the aim of this fusion process is to fused the features at the track level in order to generate an indication of track similarity from the given features. Once the features are all at a common level of time integration, in this case the track level, they can then be fused using the ensemble of classifiers given as equations **(71)** and **(72)** in the previous section.

7.3 Results From Fused Features

This section looks to validate the usage of the Bayesian fusion techniques and models chosen. This is achieved by providing a detailed analysis of the fused results from footage obtained in a video surveillance scenario. The experiments use data obtained from a real surveillance system in order to provide realistic views; however the individuals observed are actors to follow privacy laws, and to ensure that individual ground-truth height and clothing colour measurements can be obtained. It also allows for specific difficult cases to be constructed to test some of the limitations of the system. The first experiment looks at validating the Gaussian probability model of the two classes. It looks at validating the model by exploring the difference between theoretical and actual error values when the threshold is varied from the theoretical optimum. The model can be assumed to be accurate enough where the error behaviour of the model changes in the same manner as the real data.

With the model validated, the second experiment looks at the evaluation of the results for the individual features, and their fused results across a set of over 300 recorded track pairs obtained from two cameras with differing lighting conditions. These are presented as ROC curves showing the discriminative power of each of the individual features and the fused features. Further analysis of these results are then provided by comparing the fusion of selected features. This allows for an evaluation of combining particular features to investigate their effects upon the fused results.

The two experiments are performed on a dataset based upon a comparison of four people across two cameras from over 300 possible comparison combinations. The data was obtained from a real surveillance system situated within the University of Technology, Sydney's Information Technology building in order to provide for realistic views. Thus the data includes a degree of the compression artefacts and other errors that occur in systems which are currently used for operator based surveillance. An indication of the clothing's colour and good seg-

Figure 34: Four people of interest

mentation examples for the four individuals is given in Figure 34, where it is easy to see that the individuals are wearing clothing of approximately 50% or more differing colours. Ground-truth height differences between the individuals range from approximately 5 centimetres to 30 centimetres. The accuracy of each feature component is shown and can then be compared with the fused results. From the tracks, 60 comparison combinations are used as training dataset with the remaining used for the evaluation dataset, ensuring that there is a sizable amount of unseen data.

Person	Height(mm)	Upper Clothing	Lower Clothing
A	1600	Black	Black
B	1550	Red	White
C	1900	Light Blue	Black
D	1710	Black	White

Table 10: Ground Truth of Participants

7.3.1 Evaluation of the Statistical Models

Using the statistical probability of a feature given a similarity value allows for the features to be easily fused; however it also requires a statistical model to be applied. In order to validate the statistical model, an experiment is required which compares the theoretical error rates with the error rates obtained with real data. If the model is correct, then the real and theoretical rates should change in similar patterns when the threshold is varied from the theoretical optimum. This was tested based upon both the changes in the real and theoretical errors from the training set of known matching and non-matching data, and an evaluation set that

was not used to generate the model. Whilst the evaluation of the error changes in the training dataset is useful for evaluating the Gaussian statistical models, the evaluation data should give an indication of the wider applicability of the model for the broader unseen data.

This evaluation of the statistical models was performed on a dataset of 15 tracks across 2 cameras, with the results given in Table 11. The results are evaluated for the features investigated when controlled equalisation with $k = 2$ was applied. The threshold $s_{t_{th}}$ in equation 62, was adjusted from its theoretically determined optimum by multiplying it by an adjustment factor, with 0 indicating the error rate at the theoretical maximum. The error rates are all given here as percentages.

Table 11: How variations to the optimum threshold affect% error rates

Feature	-50%	-25%	-10%	0	+10%	+25%	+50%
Upper MCR total err	52.5	38.3	33.3	33.3	34.2	33.3	29.17
Upper MCR MD	0.00	0.03	0.07	0.08	0.09	0.13	23.3
Upper MCR FA	52.5	35.8	26.7	25.8	25.0	20.8	0.58
Lower MCR total err	36.7	29.2	29.2	29.2	28.3	27.5	33.3
Lower MCR MD	0.03	0.04	0.04	0.04	0.05	0.07	19.2
Lower MCR FA	33.3	25.0	25.0	25.0	23.3	20.8	14.2
Global MCR total err	57.5	54.17	43.3	38.3	30.8	37.5	37.5
Global MCR MD	0.00	0.01	0.08	10.8	18.3	37.5	37.5
Global MCR FA	57.5	53.3	35.8	27.5	12.5	0.00	0.00

Table 11 demonstrates that varying the threshold from the statistically determined optimum generally leads to a higher overall error rate. Although the model is not perfectly fitted to the data, it does suggest that the Gaussian model of the non-matching $H0$ and matching $H1$ cases provides an acceptable model for each feature. When the errors are separated into false detections (FD) and false alarms (FA), it becomes obvious that the theoretical point is where the errors change from being mainly false alarms to being increasing missed detection. Thus probabilities for each class based upon this model should provide results that will lead to reliable fusion of the features under the proposed framework. The evaluation results indicate that the model obtained is applicable to the wider dataset, and should provide a good model for a general system.

7.3.2 Evaluation of Fusing Features

This section investigates using the fusion of multiple features method described, and how it improves the results of matching individuals over using individual features. This experiment was performed using the small carefully crafted dataset from 4 individuals across two cameras to provide a series of results that compares the accuracy of the individual features with a combination of fused features. The fusion results presented here utilise all of the techniques presented in the previous chapters for feature extraction, as well as the techniques for improving feature robustness to errors such as those that occur with poor segmentation. The initial results provided as Figure 35 show how the fusion of all the features provides an improvement in accuracy over each of the individual features. A broader investigation is then performed by comparing the fusion of all the features to the fusion of selected features to determine how much each of the features add to final fusion results.

Although the method of fusion is important to provide the best results, improved results from fused features only occurs where those features are complementary. This occurs where the information from each of the features adds to the information provided by the other features. An example of this can be seen through the usage of height and colour information. The two features can be made reasonably accurate, but rely upon very different measurements of the individual. Thus the information obtained is likely to be affected by different error modes, such as segmentation errors affecting height, whilst illumination changes affect colour appearance. The sensitivities to different error sources is likely to make these features complementary, which in turn is likely to lead to more accurate results if they are combined effectively. Where there are large inaccuracies in features, especially where the features are not complementary, the fusion of multiple features can lead to the multiplication of errors, dramatically reducing their combined accuracy. For this reason it is important to evaluate the fusion framework using all of the features, but also to investigate whether other feature combination might produce better results.

The first results comparing the individual features with the fused results are shown in Figure 35. They demonstrate that the fusion of the chosen features can provide a probability of detection of 91% with only 5% false alarms at the chosen operating point. The fusion of the features clearly outperforms any of the individual features, indicating the complementary nature of the features. The accuracy of these results are obviously not high enough for a fully automated system in a critical area such as security and video surveillance; however they are a promising

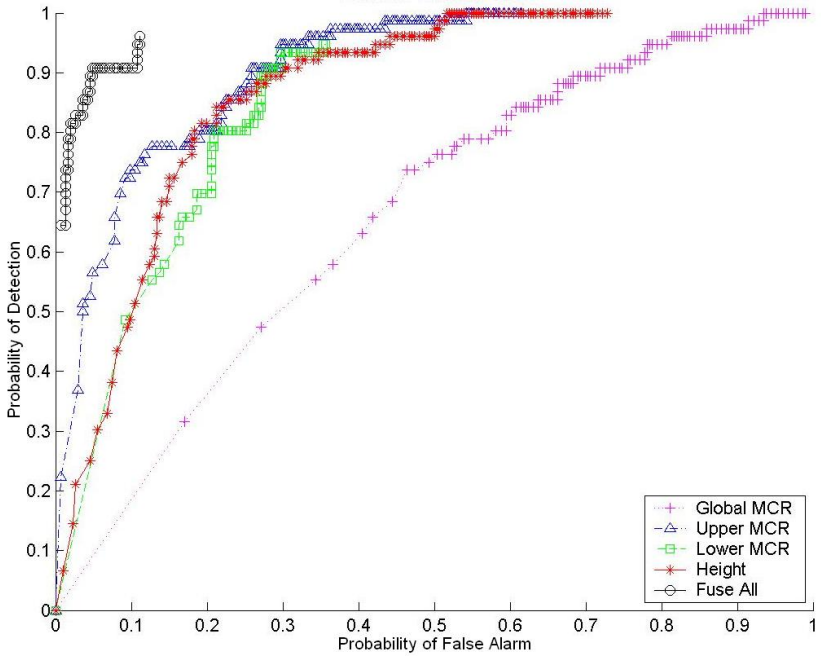


Figure 35: ROC curves of the height, MCR and fused feature results

indication that a system such as this could be useful for providing additional information to security officers in a semi-automated system. Although this points out the increased effectiveness of the fused features, it does not necessarily provide the complete picture of how each feature might contribute to the overall result. Thus further investigation of how the features compliment each other through the fusion of other feature combinations is required.

An interesting aspect of the features used is that the spatial colour MCR features individually have a higher degree of accuracy than the global colour MCR's. Thus if the number of features is limited, then the use of spatial colours would be preferred over the global colours. Where all of the colour features are available it is also interesting to investigate whether the addition of the global colour MCR to the other spatial colour features adds any extra information to improve the accuracy of the results. Figure 36 shows the results on the comparison of tracks using either the fused spatial colour MCR's, fusing both the spatial and global colour MCR's, and fusing all the features. This clearly shows that the spatial colours are individually more accurate than the global colour MCR, and that the global MCR may slightly reduce the overall fused results. This would suggest that incorporat-

ing the global MCR information about the object is largely redundant.

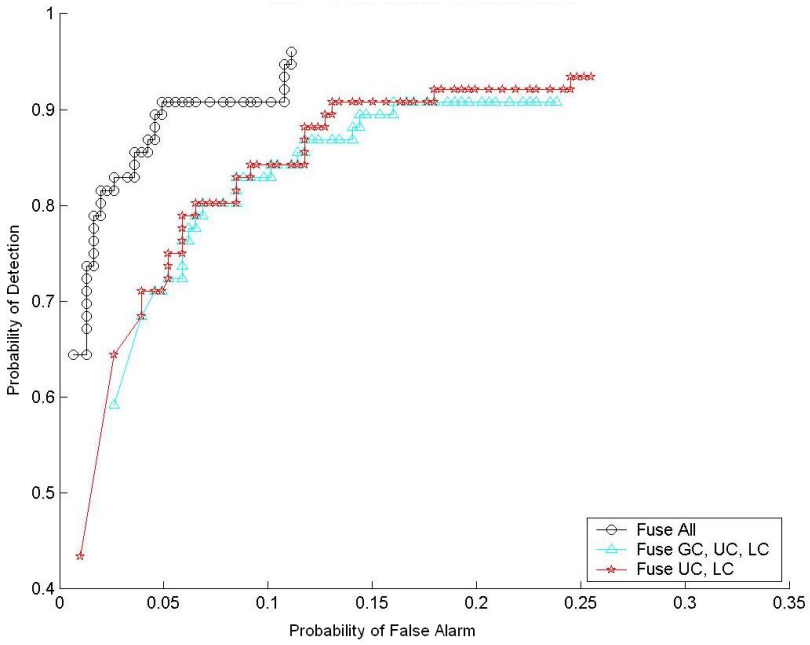


Figure 36: ROC curves for fusing spatial colour MCR's, fusing all the colour MCR's, and fusing all the features

The other significant feature proposed in this work is the height feature, which can be obtained from either a monocular, or stereo cameras. This feature is obviously derived from very different aspects of the observed individuals, and is sensitive to different error patterns than the colour appearance based features. The additional information provided by this height feature can also be determined by looking at how the results differ between the inclusion or exclusion of the height feature. These results are also shown in Figure 36, demonstrating that the addition of the height feature adds considerably to the accuracy of the fused results. This shows that the improvements that can be made through the complementary nature of height as shape based feature to other appearance based features. Such comparisons of the results obtained through the inclusion or exclusion of particular features could be a powerful tool to determine whether they are useful for improving the overall results achieved within the system. Such an experiment is also likely to be very useful especially where real time system requirements limit the computational power available for extracting the useful features.

This thesis investigated the mitigation of many sources of error to obtain the overall system results demonstrated. It is important to note that these sources of error are likely to occur at different rates in different components of any surveillance system. Some camera views are likely to have greater problems with segmentation due to the particular colouring of portions of its background, whilst other areas, especially those open to sunlight, are likely to suffer from greater changes in illumination. These changing conditions will affect how error sources impact upon the system, but may be most significant for comparisons within a given camera view, or across specific camera pairs. An analysis of results with a view to matching across specific camera views, or within a single view is also likely to identify these comparisons within the system that are likely to have either increased or decreased accuracy within a system. This is obvious when you consider that the error conditions within a single camera are likely to be more stable and lead to more accurate results than those comparisons between camera views.

7.4 Summary of Fused Features

This chapter has presented a framework based upon an ensemble of classifiers to fuse multiple features of an individual to compare them when they are tracked within a surveillance system. This framework has been based upon research into applying the Bayesian fusion framework [37, 58]. It investigated the theoretical accuracy obtained by various combinations with dependent or independent features, using statistical likelihood functions. These likelihood functions can be easily determined from a small training set of known matching and non-matching tracks with the parameters stored as the mean μ and standard deviation σ of the non-matching $H0$ and matching $H1$ cases. Given a similarity measure for each feature, the decision about whether an individual is matched is determined by the maximum likelihood of the particular comparison being matching or non-matching as calculated by the maximum of equation **(71)** or equation **(72)**. These equations are derived by fusing the likelihood values of each feature fitting the $H0$ or $H1$ cases, increasing the discrimination between individuals where some of their features are similar. This framework also allows for an investigation of the lowest level of time integration inherent to each feature. This could be the frame level, a window of frames, or an analysis from the entire track, depending upon how stable the feature is at a given level. These features can then be fused up to the track level using equations **(73)** and **(74)** for track level matching is performed using Equations **(71)** and **(72)**.

This chapter has also provided a detailed analysis of the results of fusing the

features using this framework. Matching accuracy as high as 91% with only 5% false alarms was achieved using a fusion of all features on the evaluation set. The information provided by each feature has been combined to provide this result, demonstrating the complementary nature of most of the features. An investigation into improving the usage of inter-camera and intra-camera statistics to optimise parameters may be able to improve upon these results; however investigating other areas of error source reduction, including segmentation errors and illumination changes, may also provide increased results. These results suggest that the fusion framework presented within this chapter provides a solid basis for developing a tracking by matching surveillance system. Improvements in the accuracy of the system could be improved by the addition of other strongly discriminative features; however improvement in the accuracy of the underlying features through the usage of longer tracks, higher resolution cameras, and more accurate object segmentation techniques are likely to create greater improvements in the overall accuracy.

The results of this fusion framework provide an indication of which track pairs are likely to be matching, and can be combined together to provide an indication of the movements of individuals around the system. This similarity is very dependent upon the assumption that their clothing and footwear do not change during a surveillance session; however where such changes occur, it is likely that a human operator may be capable of rectifying. Obtaining the full track of an individual from entrance to exit is also likely to be difficult to obtain in its entirety, given the 91% overall accuracy of matching. Thus if the assumption of viewing the entry or exit points of the surveilled system is broken, then only a partial track of the individuals movements can be obtained; however even this partial information may provide important information about the movements of individuals. Unfortunately not knowing when an individual has exited the system could lead to extra comparisons within the system that are unnecessary. Thus when looking in detail at the results from the combination of framework and features presented, automatically obtaining a full track of individuals within this system is unlikely, especially for long paths; however a significant amount of additional information can be provided that is beneficial for a human operator. Such a system would require significant development and experimentation outside of the scope of this thesis.

This fusion framework can provide information to security officers about the movements of individuals throughout a wide-area surveillance system. The track similarity measurements have been found to provide high level of accuracy; however there are still considerable false alarms. To overcome such limitations one

Figure 37: Pictorial storyboard summary of potentially matching tracks

could also look to minimise the cost on the overall system to the operator. This could be achieved using a 'storyboard' approach to the semi-automated tracking system, as shown in Figure 37. This involves creating a short pictorial summary of the given tracks, either by selecting the first and last frames as well as intermediate frames at regular intervals, or by using the similarity measurements to identify the key frames from the track. This pictorial summary of each track that can be compared visually by the human operator. Such a pictorial system is likely to dramatically reduce the human cost of the revision of false alarms, especially compared to the manual search that would be required to find and match the tracks from a pool of all possible tracks. Once track matching has been approved, the system can add it to other known tracks from that individual to build up the movements within the system. Although such a system may be possible at real-time, or near real-time speeds in areas that have low levels of traffic, it is likely to be most useful for forensic investigations of stored surveillance data. Here real-time computation is not essential, though fast processing is desired. The advantage however from the selection of key individuals of interest to track, where a set of the tracks that might relate to the individual could be automatically identified and manually assessed with high accuracy. This could dramatically reduce the search time required for an operator to manually identify and assess those tracks.

8 Conclusions

This project has combined a range of research from a variety of areas in order to address the challenging topic of tracking individuals across a real surveillance system. The system investigated is aimed at building surveillance in both public buildings or secure buildings where human traffic allows for the segmentation of individuals into single objects. The recent advances in the resolution and quality of video systems are capable of providing more accurate information for performing automated computer vision. However real video surveillance systems are generally upgraded only every few years with the major consideration being the trade off between camera resolution, quality, and the cost of installing significant numbers of such cameras. Automating the system is only a minor consideration in such upgrades as the current research into automated surveillance is mainly focussed upon using the latest high resolution cameras, often with overlapping fields of view. Such installations do not generally reflect the existing real surveillance systems, except for the most secure locations. Direct camera level control is also often performed under laboratory conditions, sometimes in environments that have simplified backgrounds. Instead the focus of such upgrades is upon providing a cost effective upgrade that can allow human operators to identify events, behaviours, and individuals more accurately in a distributed system where only key locations are observed.

This thesis has focussed upon automatically utilising the information available from existing wide area surveillance systems for tracking individuals. This focus included the four broad research aims of exploring appearance features, shape features, making these features more robust to their error sources, and fusing these robust features in an accurate manner. These four broad aims have been explored in the scope of a real surveillance system without requiring any knowledge of topological information from the system, such as camera locations or path transitions and transition times between cameras. Exploring these aims led to significant contributions to the video surveillance field, which are explained in the following five paragraphs that summarise the chapters of the thesis. These chapters explore the colour appearance features, the mitigation of illumination effects upon appearance, the identifying of segmentation errors, the height estimate feature, and the fusion of all the features to explore system accuracy.

The literature review has shown colour appearance features to be widely used in the area of human tracking and the matching of individuals. The exploration within this thesis has looked at both the colour representation aspect, as well as looking at how to apply those representations to measure similarity. The inves-

tigation of the colour feature representation found many colour spaces aimed at particular goals, such as illumination tolerance; however with each step information is lost, such as the co-occurrence of the colour channels if the RGB values are separated into the three channels. To overcome the size problems with the full RGB space this work has contributed the Major Colour Representation (MCR) to the literature. This colour space retains the co-occurrence of the colour channels, whilst retaining a compact nature by storing only the components of the regions of the sparse histogram that are populated by pixels. These regions are found by colour clustering, whose accuracy is improved through iterative k-means optimisation of the clusters. It is also worth noting that colour calibration has not been performed upon the cameras, and such a process may provide an improvement in the colour related features. A second contribution in improving the MCR features was found by averaging the features across a small window of frames. This window has been found to reduce the impact of shape changes and minor segmentation errors upon the MCR features. A third contribution has been to identify spatial regions upon which the MCR colour features can be extracted. An analysis of similarity results of the spatial features representing upper and lower clothing colour regions clearly shows the additional information about the positioning of colours that can be achieved using a simple global colour scheme. The final contribution of the MCR representation has been the development of a symmetric similarity measurement based upon the compliment of the Kolmogorov divergence. This measure can quantify how much of each object's colours occurs within the other object, allowing for feature fusion processes that are quantitative and not purely decision based. This combination of contributions to the effective extraction and usage of colour appearance features has also allowed for a quantitative investigation of feature based error sources as well as exploring the accuracy of a colour appearance based matching system.

A major problem with the usage of colour appearance features in a wide area surveillance system are the changes in the perceived colour across the system. These changes can be due to the camera, although colour calibration could possibly reduce this, or through changing levels and sources of illumination. As a full three dimensional model incorporating light sources as well as background and shape and properties is infeasible, this thesis has investigated techniques that can mitigate the effect of illumination upon the colour appearance of objects. Based upon the promising results of histogram equalisation for mitigating the effects of illumination on colour appearance, this thesis has contributed 'controlled' equalisation and its centralised version to allow for increased matching, whilst retaining discriminative ability. This technique equalises a combination of the object's his-

togram and a ratio of pre-equalised pixels to control the amount of stretching to prevent it from equalising too far. The centralised version adjusts the level of pre-equalised pixels to force the mean of the object's histogram to the centre of the output histogram. In developing a method for the comparison of colour appearance features, this work has indirectly contributed a method for evaluating the usefulness of a variety of techniques for mitigating the effects of changes in illumination. Using this method, an analysis of the effects of homomorphic image filtration, histogram stretching, histogram equalisation, and controlled histogram equalisation were contributed to the literature. This clearly showed that the histogram equalisation based techniques provided the greatest improvements in error reductions, whilst controlled equalisation provided the greatest similarity for matching colours. The reductions in errors from 13% without mitigating illumination effects, to 8% with full histogram equalisation, 8.5% for controlled equalisation and as low as 7% for centralised controlled equalisation demonstrates that this can be a very useful preprocessing step to the extraction of colour appearance features.

Poor segmentation is also a significant source of error in video surveillance, as well as many image analysis applications. This thesis has investigated many aspects of the analysis of object features both for the purposes of track matching, but also for increasing the robustness of those features. The quantitative analysis of the changes in the similarity of features along the known track of an individual has contributed to the area of video analysis by showing it can provide a good indication of frames where segmentation errors occur. The results show that identification of over 80% of erroneously segmented frames is achieveable by looking for changes in the MCR colour appearance features, with only 3% false detections. These results indicate that this method is useful even when tracks consisting of as few as 10 frames are analysed. Where longer tracks are available, higher detection rates can be achieved as there is a greater number of frames available, allowing for extra redundancy where higher false alarms occur. This focus upon the identification of erroneous frames still allows for improved results through advances in latest segmentation techniques; however it also counters those segmenation errors that occasionally occur with even the most accurate of the current techniques.

Height estimates are a shape feature that has been used both in the video surveillance literature, and also as a descriptive feature for the police to identify suspects. This feature has been extracted in both multi-camera, as well as monocular camera views; however this thesis contributes to the accuracy of the height estimate from any single camera frame by improving the location of the key top and bottom points from simple bounding box measurements. This increases the

accuracy of the estimate of an individual's height in any given frame, even though this also increases the influence of gait on the height estimates. This gait effect although more pronounced, is periodic in nature and has been found to be averaged out when considering the entire track of an object, or at a minimum a number of frames greater than the period of an individual's gait. Increased accuracy can be achieved where statistical outliers are removed, and also where frames with large segmentation errors are removed. Thus an automatic estimate of an individual's height is possible for comparison to real world estimates of a person's height. A second contribution of this thesis' investigation of height estimates has been to provide a similarity measurement based upon the statistical analysis of the frame based height difference obtained from two tracks. This similarity measurement not only considers the difference in height between the individual's in the two tracks, but also the noise within those measurements. The results obtained using this height similarity measurement show that this feature provides similar accuracy to the spatial MCR features on the data set analysed, with accuracy as high as 86% with only 15% false alarms where error rates are minimised.

This thesis has investigated a number of features that can be fused together to provide greater overall system accuracy. As each of the features provide a range of similarity measurements, rather than just matching or non-matching decisions, Bayesian fusion methods were investigated. The existing theoretical literature, as well as system tests, demonstrated that multiplicative Bayesian fusion tended to provide the best results where the fusion was between the independent features. Where the fusion is within a single feature to raise it to a higher feature level, such as from a collection of window similarities to a track level similarity, the feature is dependent upon the same data and therefore the literature indicates fusion using the average rule. The contributions of this chapter begin with the analysis of the results of fusing the features through the changes of accuracy when each feature is included. This analysis clearly showed that the height feature is complementary to the appearance features, as it provides a significant improvement when included in the fusion process. This point is also significant as the height feature is sensitive to segmentation errors, whilst the appearance features are more sensitive to illumination changes. This combination may help to minimise the impact of either form of error on the overall results. Additionally the spatial colour features were also found to add significantly to the fused results, whilst the global colour feature, with its dispersed range of colours, provides little additional information. The final system results achieved a matching accuracy as high as 91% with only 5% false matches on the evaluation data. These results are not accurate enough for a fully automated surveillance system, yet they are significant enough to provide

additional information for a human operator to review. The proposed framework for feature integration exploring the time and feature levels can also be used as a general framework for integration of features of various natures. The feature set used in this work can be extended with other features as they become available without any significant modifications to the approach.

This thesis has proposed a fusion framework aimed at combining shape and appearance features to determine if two observed tracks are from the same individual or differing individuals. Where tracks have been matched to the same individual, they can be combined over time to describe the observed movements of an individual throughout the surveillance area. This framework and the features used within it have been based upon a number of assumptions to simplify the system to a manageable size. Unfortunately in any real system the developers assumptions may be violated at times. Thus the seven main assumptions listed in the introduction are analysed here to examine the impact of breaking them:

1. All entry and exit points of the surveillance area are in view of a surveillance camera - Breaking this assumption may lead to difficulty in obtaining the entire track of an individual. More importantly it will make it difficult to determine when an individual enters or leaves the system. Although entry is not so important as tracks could be generated from any starting point, missing an individual leaving the system could lead to significant extra comparisons between old tracks and newly observed individuals. These are unnecessary because the previously observed individual is no longer present within the area. Although an individual's status could be set to expire after a period of time, this could lead to hours of footage undergoing unnecessary comparisons, or the separation of an individual's track where they remain for long times within the system.

2. Individuals are unlikely to change their clothing or footwear; hence, many of their intrinsic shape and appearance features will remain relatively constant for the duration of the surveillance session - Significant changes in footwear is likely to impact upon the height estimates of an individual, though often by only a few centimetres unless they include large heels. This will increase the possibility of errors but may not be large enough to have a major impact. Changes in clothing are likely to cause more significant errors as they are likely to generate dramatic changes in similarity of the MCR features. This is likely to lead to an individual no longer matching his former appearance creating a single system wide disjoint track. The two

or more segments with differing appearances will be somewhat likely to be tracked effectively in the parts and could possibly be reconciled by a human operator. Sometimes, where the change is dramatic, even human operators may fail at this task.

3. Individuals are tracked accurately whilst within the view of any of the system's cameras - Where individual tracks become incorrect, the features are likely to change suddenly where tracking becomes erroneous. This information could possibly be used to automatically identify incorrect or unreliable tracks, or could be used by a human operator to reconcile the error. Where the lengths of the track components are long, the features obtained could potentially also be used to reconcile the tracking automatically; however such investigations have been outside the scope of this thesis. The framework provided can be used to evaluate similarity along the track, and where this similarity is low, the poor quality of the track can be identified.

4. Individuals are segmented from the background into a single blob, but not necessarily accurately - The extraction of features within this thesis are based upon the object being contained within a single blob, with the features changing significantly where this does not occur. Such changes in features currently form the basis of identifying frames with major segmentation errors; however incorrect segmentation is hard to identify where such errors occur frequently. Breaking this assumption is therefore likely to cause significant errors within the system, although this may be rectified through operator intervention. Areas were objects don't have strong contrast from the background is an existing problem, even for human surveillance operators.

5. Individuals are generally observed at a distance from the camera, so biometric features such as faces may not be always available - This assumption has led to a focus upon robust shape and appearance based features; however where biometric features are available and reliable, they could also be added as a feature within the system. Such features are also likely to increase the overall accuracy of the system.

6. Where cameras are significantly disjoint, motion features may vary unpredictably between those cameras as individual's are allowed free motion - This assumption may be broken in areas such as hallways or near overlapping views where there is a very predictable motion or transition between cameras. If such information was found to be reliable, then it could be used

either as an additional feature to increase the accuracy of the system, or as a method to restrict which tracks are compared to reduce the computational load. Using such information is outside the scope of this thesis in order to focus upon maximising the information obtained through feature analysis and fusion alone. Additionally this information may be included only through the expansion of the manual calibration required for operating on a real system, and may be too time consuming to be feasible.

7. Illumination varies significantly, but within the limited range typical of natural or artificial lighting - Illumination within human environments only tends to vary within a limited range as very bright or dark environments tend to be stressful to the human visual system. Changes outside of this limited range are also a major problem for object segmentation and appearance features, and are likely to severely impair any surveillance system. Alternative illumination sources, such as blue lighting, are also very problematic, but are often used only within toilets, where surveillance cameras are intrusive upon an individual's privacy, and illegal in many countries.

This indicates that where the major assumptions of the system are broken, then either inaccurate, or incomplete information is likely to be obtained. Given the 91% accuracy of the fusion framework with the current features on a small carefully constructed dataset, a fully automated system is unlikely to be implemented directly; however the information provided by a semi-automated system will be beneficial to a human operator, under ideal conditions. Even under adverse conditions, the track information provided is likely to be either reconciled by a human operator through a suitable interface, such as a track based storyboard. Where this is not possible, then the system errors may be significant and could not be guaranteed to provide anything more than an improved starting point for human investigation over having no information at all. It is important to note that further invesitgation is required into the experimental accuracy of both the fusion framework, through the analysis of larger and more diverse datasets, and through experiments to invesitgate the usefulness of the story boarding approach through such techniques like the information gain of the system operators. Whilst the analysis of larger datasets is planned as a future publication from this work, inverstigation of the system applications and information gain are likely to form a significant portion of the future user centred component of the surveillance field.

The research conducted here focuses upon the results that are achievable with currently installed technology. This focus makes the assumption that both the

camera technology and the communications bandwidth may cause limitations to the image size and quality obtained. Such restrictions are commonly found in the majority of building surveillance systems, where criminal activity and terrorism are considered to be of very low likelihood. Whilst this research does not currently make use of the very latest high resolution technology, or placements of large numbers of cameras, there is no reason that it could not be exploited where available. Higher resolution cameras with good quality colour sensors and high bandwidth are often used in laboratory conditions, and where available in a surveillance system would only improve the accuracy of this method. The accuracy of matching individuals where cameras are placed close together or even overlapping is also likely to increase as effects such as illumination are also likely to improve. These increases in technology and knowledge of the surveillance area can all be used to improve the system through the expansion of the framework to include other features, limitation in the variance of effects like illumination, or by limiting the potential match search space where absolute path transitions are physically limited, such as within corridors. Additionally such a system could also generate statistics on the movement of individuals, such as average path transition time, in order to determine the normal cases and alert an operator when anomalous cases occur. Therefore it is envisaged that improvements in technology and surveillance infrastructure will only improve a system based upon this method, and will certainly not make it redundant.

This thesis has presented a system that can provide additional information to security officers about the movements of individuals throughout a wide-area surveillance system. Such a system would be able to attract the officer's attention to those areas where motion is occurring, but more powerfully it would be able to provide information about tracks that are likely to be obtained from the same individual. The track similarity measurements have been found to provide a high level of accuracy; however there are still considerable false alarms. Such a high level of false alarms are undesirable, especially in security and safety critical systems such as surveillance. In order to overcome such a problem, one could also look to minimise the cost on the overall system to the operator. This could be achieved using a 'storyboard' approach to the semi-automated tracking system. This involves creating a short pictorial summary of the given tracks, either by selecting the first and last frames as well as intermediate frames at regular intervals, or by using the similarity measurements to identify the frames that are the most similar to the rest of the track. This pictorial summary of each track can be compared visually by the human operator. Such a pictorial system is likely to dramatically reduce the human cost of false alarms, especially compared to the manual search that would

be required to manually find and match the tracks from a pool of all the possible ones. Once track matching has been approved, then the system can add it to other known tracks from that individual to build up the movements within the system. Although such a system may be possible at real-time, or near real-time speeds in areas that have a relatively low level of traffic, it is likely to be most useful for forensic investigations of surveillance data. Here real-time computation is not essential, though fast processing is desired. The real advantage however occurs from the selection of key individuals of interest, where a set of tracks that might relate to the individual could be automatically identified and manually assessed with high accuracy. This could dramatically reduce the search time that would be required for an operator to manually identify and assess those tracks.

References

[1] A. E. Abdel-Hakim and A. A. Farag. Csift: A sift descriptor with color invariant characteristics, 2006. IEEE Conference on Computer Vision and Pattern Recognition.

[2] N. Artner. A comparison of mean shift tracking methods, 2008. Central European Seminar on Computer Graphics.

[3] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–187, 2002.

[4] K. Barnard, B. Funt, and V. Cardei. A comparison of computational colour constancy algorithms; part one: Methodology and experiments with synthesized data. *IEEE Transactions in Image Processing*, 11:972–984, 2002.

[5] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. Burkitt. Performance of optical flow techniques, 1992. IEEE Conference on Computer Vision and Pattern Recognition.

[6] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:509 – 522, 2002.

[7] C. BenAbdekader, R. Cultler, and L. Davis. Person identification using automatic height and stride estimation, 2002. International Conference on Image Processing.

[8] A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin Calcutta Mathematical Society*, 35:99–110, 1943.

[9] S. Birchfield and S. Rangarajan. Spatial histograms for region-based tracking. *ETRI Journal*, 29(5), 2007.

[10] Q. Cai and J. Aggarwal. Automatic tracking of human motion in indoor scenes across multiple synchronized video streams, 1998. International Conference on Computer Vision.

[11] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[12] R. T. Collins, A. J. Lipton, H. Fujiyoshi, and T. Kanade. Algorithms for cooperative multisensor surveillance. *Proceedings of the IEEE*, 89(10):1456–1477, 2001.

[13] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(5):564–575, 2003.

[14] S. Cotton. Colour, colour spaces and the human visual system, 1996. University of Birmingham Technical Report, B15-2TT.

[15] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, 1991.

[16] A. Criminisi, I. Reid, and A. Zisserman. Single view metrology. *International Journal on Computer Vision*, 40(2):123–148, 2000.

[17] R. Cucchiara, C. Grana, M. Piccardi, A. Prati, and S. Sirotti. Improving shadow suppression in moving object detection with hsv colour information, 2001. IEEE International Conference on Intelligent Transport Systems.

[18] R. Cutler and L. S. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781–796, 2000.

[19] T. Darrell, G.Gordon, M. Harveille, and J. Woodfill. Integrated person tracking using stereo, colour, and pattern detection. *International Journal on Computer Vision*, 37(2):175–185, 2000.

[20] J. Deutscher, M. Isard, and J. MacCormick. Automatic camera calibration from a single manhattan image, 2002. Proceedings of European Conference Computer Vision.

[21] P. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[22] A. Doshi and M. Trivedi. 'hybrid cone-cylinder' codebook model for foreground detection with shadow and highlight suppression, 2006. IEEE International Conference on Advanced Video and Signal Based Surveillance.

[23] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2001.

[24] A. Elgammal and L. Davis. Probabilistic framework for segmenting people under occlusion, 2001. International Conference on Computer Vision.

[25] A. Elgammal, R. Duraiswami, and L. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance, 2001. IEEE Conference on Computer Vision and Pattern Recognition.

[26] A. Elgammal, R. Duraiswami, and L. Davis. Efficient kernel density estimation using the fast gauss transform with applications to color modeling and tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1499–1504, 2003.

[27] A. Elgammal, R. Duraiswami, and L. Davis. Probabilistic tracking in joint feature-spatial spaces, 2003. IEEE Conference on Computer Vision and Pattern Recognition.

[28] A. Elgammal, R. Duraiswami, D. Harwood, and L. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, 2002.

[29] C. E. Erdem, F. Ernst, A. Redert, and E. Hendriks. Temporal stabilization of video object segmentation for 3d-tv applications, 2004. International Conference on Image Processing.

[30] M. Fairchild. *Colour Appearance Models*. Addison Wesley, 1998.

[31] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[32] G. Finlayson, S. Hordley, C. Lu, and M. Drew. On the removal of shadows from images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:59–68, 2006.

[33] G. Finlayson, S. Hordley, G. Schaefer, and G. Y. Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern Recognition*, 38:179–190, 2005.

[34] A. Ford and A. Roberts. *Colour Space Conversions*. Westminster University, London, 1998.

[35] H. Freeman and L. Davis. A corner-finding algorithm for chain-coded curves. *IEEE Transactions on Computing*, 26:297–303, 1997.

[36] N. Friedman and S. Russel. Image segmentation in video sequences: a probabalistic approach, 1997. Conference on Uncertainty in Artificial Intelligence.

[37] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 27(6):942–956, 2005.

[38] T. Gandhi and M. Trivedhi. Panoramic appearance map (pam) for multi-camera based person re-identification, 2006. Advanced Video and Signal Based Surveillance.

[39] T. Gandhi and M. Trivedhi. Panoramic appearance map (pam) for multi-camera based person re-identification. *Machine Vision and Applications Journal*, pages 207–220, 2007.

[40] R. Gonzales and R. Woods. *Digital Image Processing*. Prentice Hall, 2002.

[41] W. Grimson, C. Stauffer, R. Romano, and L. Lee. Using adaptive tracking to classify and monitor activities in a site, 1998. IEEE Conference on Computer Vision and Pattern Recognition.

[42] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti. Smart video surveillance: Exploring the concept of multiscale spatiotemporal tracking. *IEEE Signal Processing Magazine*, 22(2):38–51, 2005.

[43] I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-time surveillance of people and their activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):809–830, 1998.

[44] I. Haritaoglu, D. Harwood, and L. Davis. An appearance-based body model for multiple people tracking, 2000. International Conference on Pattern Recognition.

[45] I. Haritaoglu, D. Harwood, and L. Davis. A fast background scene modeling and maintenance for outdoor surveillance, 2000. International Conference on Pattern Recognition.

[46] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2004.

[47] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics*, 34:334–352, 2004.

[48] T. Huang and S. Russel. Object identification: A bayesian analysis with application to traffic surveillance. *Artificial Intelligence*, 103(1–2):77–93, 1998.

[49] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density, 1996. European Conference on Computer Vision.

[50] M. Isard and A. Blake. Condensation-conditional density propagation for visual tracking. *International Journal on Computer Vision*, 29:5–28, 1998.

[51] D.-S. Jang and H.-I. Choi. Active models for tracking moving objects. *Pattern Recognition*, 33(7):1135–1146, 2000.

[52] O. Javed, Z. Rasheed, K. Shafique, and M. Shah. Tracking across multiple cameras with disjoint views, 2003. International Conference on Computer Vision.

[53] O. Javed, K. Shafique, and M. Shah. Appearance modeling for tracking in multiple non-overlapping cameras, 2005. IEEE Conference on Computer Vision and Pattern Recognition.

[54] T. Kailath. The divergance and bhattacharyya distance measures in signal selection. *IEEE Transactions on Communications Technology*, 15(1):52–60, 1967.

[55] V. Kettnaker and R. Zabih. Bayesian multi-camera surveillance, 1999. International Conference on Computer Vision and Pattern Recognition.

[56] M. Kilger. A shadow handler in a video-based real-time traffic monitoring system, 1992. Workshop on Applications of Computer Vision.

[57] K. Kim, T. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground-background segmentation using codebook model. *Real-Time Imaging*, 11:172–185, 2005.

[58] J. Kittler, M. Hatef, and R. D. J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239, 1998.

[59] M. Kohle, D. Merkl, and J. Kastner. Clinical gait analysis by neural networks: Issues and experiences, 1997. Proceedings IEEE Symposium on Computer-Based Medical Systems.

[60] D. Koller, J. Weber, T. Huang, J. Malik, G. Ogasawara, B. Rao, and S. Russell. Towards robust automatic traffic scene analysis in real-time, 1994. International Conference on Pattern Recognition.

[61] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[62] P. Kumar, K. Sengupta, and A. Lee. A comparitive study of different colour spaces for foreground and detection for traffic monitoring system, 2002. International Conference on Intelligent Transport Systems.

[63] L. Li, W. Huang, I. Gu, K. Tian, and Q. Tian. Principal color representation for tracking persons, 2003. International Conference on Systems, Man, and Cybernetics.

[64] L. Li, W. Huang, I.-H. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

[65] A. Lipton. Local application of optic flow to analyze rigid versus nonrigid motion, 1999. International Conference on Computer Vision Workshop Frame-Rate Vision.

[66] A. Lipton, H. Fujiyoshi, and R. Patil. Moving target classification and tracking from real-time video, 1998. IEEE Workshop on Applications of Computer Vision.

[67] T. Lissack and K.-S. Fu. Error estimation in pattern recognition via c-distance between posterior density functions. *IEEE Workshop on Applications of Computer Vision*, 22(1):34–45, 1976.

[68] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[69] W. Lu and Y. Tan. A color histogram based people tracking system, 2001. International Symposium on Circuits and Systems.

[70] C. Madden, E. D. Cheng, and M. Piccardi. Tracking people across disjoint camera views by an illumination-tolerant appearance representation. *Machine Vision and Applications*, 18:233–247, 2007.

[71] C. Madden and M. Piccardi. Height measurement as a session-based biometric for people matching across disjoint camera views, 2005. Image and Vision Computing New Zealand.

[72] C. Madden and M. Piccardi. Comparison of techniques for mitigating illumination changes on human objects in video surveillance, 2007. International Symposium on Visual Computing.

[73] C. Madden and M. Piccardi. A framework for track matching across disjoint cameras using robust shape and appearance features, 2007. Advanced Video and Signal based Surveillance Conference.

[74] Y. Matsushita, K. Nishino, K. Ikeuchi, and M. Sakauchi. Illumination normalization with time-dependent intrinsic images for video surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1336–1347, 2004.

[75] K. Matusita. Decision rules based on distance for problems of fit, two samples and estimation. *Ann. Mathematical Statistics*, 26:631–641, 1955.

[76] B. Maxwell, R. Friedhoff, and A. Smith. Bi-illuminant dichromatic reflectance model for image manipulation, 2007. International Patent Application No. PCT/US2007/002238.

[77] S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80(1):42–56, 2000.

[78] D. Meyer, J. Denzler, and H. Niemann. Model based extraction of articulated objects in image sequences for gait analysis, 1997. International Conference on Image Processing.

[79] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4):349–361, 2001.

[80] J. H. A. Mojsolovic. Optimum color composition matching of images, 2000. International Conference on Pattern Recognition.

[81] C. F. Mosteller and J. W. Tukey. *Data Analysis and Regression: A Second Course in Statistics*. Addison-Wesley, 1977.

[82] H. Ning, L. Wang, W. Hu, and T. Tan. Articulated model based people tracking using motion models, 2002. International Conference on Multimodal Interfaces.

[83] C. OConaire, N. OConnor, and A. Smeaton. An improved spatiogram similarity measure for robust object localization, 2007. International Conference on Acoustics, Speech, and Signal Processing.

[84] E. Ong and S. Gong. A dynamic human model using hybrid 2d-3d representation in hierarchical pca space, 1999. British Machine Vision Conference.

[85] J. Orwell, P. Remagnino, and G. Jones. Multi-camera colour tracking, 1999. International Workshop on Visual Surveillance.

[86] H. Palus. *Colour Spaces*. Chapmann and Hall, 1998.

[87] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):266–280, 2000.

[88] H. Pasula, S. Russell, M. Ostland, and Y. Ritov. Tracking many objects with many sensors, 1999. International Conference on Artificial Intelligence.

[89] E. A. Patrick and F. P. Fischer. Nonparametric feature selection. *IEEE Transactions on Information Theory*, 15(5):577–584, 1969.

[90] M. Piccardi. Background subtraction techniques: a review, 2004. International Conference on Systems, Man and Cybernetics.

[91] R. Plankers and P. Fua. Articulated soft objects for video-based body modeling, 2001. International Conference on Computer Vision.

[92] R. Polana and R. Neilson. Low level recognition of human motion (or how to get your man without finding his body parts), 1994. Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects.

[93] A. F. Police. Identity of an offender - australian federal police. $http : //www.afp.gov.au/act/report\_crime/identify\_offender.html$, November 2007.

[94] N. S. W. Police. Nsw police - reporting knowledge of criminal activity online. $https : //www.police.nsw.gov.au/crime\_report$, November 2007.

[95] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara. Detecting moving shadows: Algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3):294–307, 2005.

[96] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005.

[97] F. Remondino and C. Fraser. Digital camera calibration methods: Considerations and comparisons, 2006. ISPRS Commission V Symposium 'Image Engineering and Vision Metrology'.

[98] S. Ribaric, G. Adrinek, and S. Segvic. Real-time active visual tracking system, 2004. Proceedings of IEEE Mediterranean Electrotechnical Conference.

[99] A. Rizzi, C. Gatta, and D. Marini. From retinex to automatic color equalization: Issues in developing a new algorithm for unsupervised color equalization. *Journal of Electronic Imaging*, 13(1):75–84, 2004.

[100] Y. Rubner, C. Tomasi, and L. Guibas. The earth movers distance as a metric for image retrieval. *International Journal on Computer Vision*, 40(2):99–121, 2000.

[101] A. B. S. How australians measure up, australian bureau of statistics, 1995. National Health Survey.

[102] E. Sangineto. An abstract representation of geometric knowledge for object classification. *Pattern Recognition Letters*, 24(9–10):1241–1250, 2003.

[103] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:162–177, 2005.

[104] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle. Appearance models for occlusion handling, 2001. International Workshop on Performance Evaluation of Tracking and Surveillance Systems.

[105] S. V. Stevenage, M. S. Nixon, and K. Vince. Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, 13:513–526, 1999.

[106] M. Tan and S. Ranganath. Multi-camera people tracking using bayesian networks, 2003. Joint Conference of the Fourth International Conference on Information, Communications and Signal Processing and the Fourth Pacific Rim Conference on Multimedia.

[107] H. Tao, H. S. Sawhney, and R. Kumar. Object tracking with bayesian estimation of dynamic layer representations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):75–89, 2002.

[108] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9):1459–1472, 2005.

[109] T. Tian, C. Tomasi, and D. Heeger. Comparison of approaches to egomotion computation, 1996. IEEE Conference on Computer Vision and Pattern Recognition.

[110] M. Tkalčič and J. F. Tasič. Colour spaces - perceptual, historical and applicational background, 2003. EUROCON.

[111] D. Toth, T. Aach, and V. Metzler. Bayesian spatiotemporal motion detection under varying illumination. *European Signal Processing Conference*, pages 2081–2084, 2000.

[112] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance, 1999. International Conference on Computer Vision.

[113] R. Tsai. A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, 1987.

[114] J. Wang, Y. Chung, C. Chang, and S. Chen. Shadow detection and removal for traffic images, 2004. International Conference on Networking, Sensing and Control.

[115] Y. Weiss. Deriving intrinsic images from image sequences. *International Conference on Computer Vision*, 2:68–75, 2001.

[116] C. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis And Machine Intelligence*, 19(7):780–785, 1997.

[117] G. Wyszcecki and W. Stilts. *Color Science Concepts and Methods, Quantitative Data and Formulae*. John Wiley and Sons Inc., 2000.

[118] H. Yan and T. Tjahjadi. Optical flow estimation and segmentation through surface fitting and robust statistics, 2003. International Conference on Systems, Man and Cybernetics.

[119] Y. Yang, D. Harwood, K. Yoon, and L. Davis. Human appearance modeling for matching across video sequences. *Machine Vision and Applications*, 18(3–4):139–149, 2007.

[120] J. Yao and Z. Zhang. Systematic static shadow detection, 2004. International Conference on Pattern Recognition.

[121] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computer Surveys*, 38(13):1–45, 2006.

[122] E. Zaharescu, M. Zamfir, and C. Vertan. Color morphology-like operators based on color geometric shape characteristics, 2003. International Symposium on Signals, Circuits and Systems.

[123] W. Zajdel and B. Krose. A sequential algorithm for surveillance with non-overlapping cameras. *International Journal of Pattern Recognition and Artifcial Intelligence*, 19(9):977–996, 2005.

[124] H. Zhang, J. Wu, D. Zhong, and S. Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern Recognition*, 30(4):643–658, 1997.

[125] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, 2004.

[126] T. Zhao, T. Wang, and H. Shum. Learning a highly structured motion model for 3d human tracking, 2002. Asian Conference on Computer Vision.

[127] S. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):917–929, 2006.

[128] Z. Zivkovic and B. Krose. An em-like algorithm for color-histogram-based object tracking, 2004. IEEE Conference on Computer Vision and Pattern Recognition.