Surveillance Video Data Fusion

Author: Simi Wang

Director of Studies: Dr. James Orwell

Supervision team: Dr.Gordon Hunter, Prof. Tim Ellis

This Thesis is being submitted in partial fulfilment of the requirements of
Kingston University for the Degree of

Doctor of Philosophy (Ph.D.)

February 2016

Digital Imaging Research Centre

Faculty of Science, Engineering & Computing

Kingston University

Penrhyn Road, Kingston-Upon-Thames

KT1 2EE, London, U.K.

Collaborating partner:

BAe Systems and EPSRC Sponsorship Number: EP/H501266/1.

Declaration

This report is submitted as requirement for the Ph.D. Degree in 'Surveillance Video Data Fusion' in the School of Computing and Information Systems, Faculty of Science, Engineering and Computing at Kingston University. It is substantially the result of my own work except where explicitly indicated in the text.

No portion of the work referred to in this report has been submitted in support of an application for another degree or qualification of this or any other UK or foreign examination board, university or other institute of learning.

The thesis work was conducted from October 2010 to December 2015 under the supervision of Dr. James Orwell in the Digital Imaging Research Centre (DIRC) of Kingston University in London.

Kingston-upon-Thames, London, United Kingdom.

Copyright Statement

Abstract

The overall objective under consideration is the design of a system capable of automatic inference about events occurring in the scene under surveillance. Using established video processing techniques, low level inferences are relatively straightforward to establish as they only determine activities of some description. The challenge is to design a system that is capable of higher-level inference, that can be used to notify stakeholders about events having semantic importance. It is argued that re-identification of the entities present in the scene (such as vehicles and pedestrians) is an important intermediate objective, to support many of the types of higher level inference required.

The input video can be processed in a number of ways to obtain estimates of the attributes of the objects and events in the scene. These attributes can then be analysed, or 'fused', to enable this high-level inference. One particular challenge is the management of the uncertainties, which are associated with the estimates, and hence with the overall inferences. Another challenge is obtaining accurate estimates of prior probabilities, which can have a significant impact on the final inferences.

This thesis makes the following contributions. Firstly, a review of the nature of the uncertainties present in a visual surveillance system and quantification of the uncertainties associated with current techniques.

Secondly, an investigation into the benefits of using a new high resolution dataset for the problem of pedestrian re-identification under various scenarios including occlusion. This is done by combining state-of-art techniques with low level fusion techniques.

Thirdly, a multi-class classification approach to solve the classification of vehicle manufacture logos. The approach uses the Fisher Discriminative classifier and decision fusion techniques to identify and classify logos into its correct categories.

Fourthly, two probabilistic fusion frameworks were developed, using Bayesian and Evidential Dempster-Shafer methodologies, respectively, to allow inferences about multiple objectives and to reduce the uncertainty by combining multiple infomration sources.

Fifthly, an evaluation framework was developed, based on the Kelly Betting Strategy, to effectively accommodate the additional information offered by the Dempster-Shafer approach, hence allowing comparisons with the single probabilistic output provided by a Bayesian analysis.

Acknowledgements

Working as a Ph.D student at Kingston University was a magnificent and a challenging experience. As painful and stressful as the whole process was, it definitely made me a better person, not only because of the additional mountain of knowledge I've digested, but also because of the people I met on the way. Without the help of a great many people directly or indirectly, I would not have written this thesis. Here is a small tribute to all those people.

The first person to thank is my advisor, Dr. James Orwell, for giving me the opportunity and introducing me to the world of Fusion research. It is due to his valuable guidance, cheerful enthusiasm and ever-friendly nature that I am able to complete my research work. I am especially thankful to him for the unwavering support and understanding during the recovery of my illness. I am also very grateful to my second supervisor Dr. Gordon Hunter. Without his supervision, eagle eye for detail and handwork, I would have been unable to publish the papers and completed this thesis in such a tight schedule.

I wish to express my gratitude to the funding support offered by the EPSRC, Kingston University and BAe Systems for me to complete my research without any financial woes. A special thank you also goes out to the Mark Drake and Zsolt Husz at BAe System. they have supported my research at various stages and instructed me to work with some of the state-of-art instruments at BAe Systems.

The last 4 years in the Kingston University have not always been easy. The lab partners, now close friends, have been extremely helpful in getting me through all the challenges of the theoretical and practical sides. I couldn't have gotten through this research without the many technical discussion between myself, Charlie Mallah and Emilio Almazan. I have especially enjoyed the arguments no matter where and when we are, going to lunch or ending at the same spot everyday. My specially thanks go to Sateesh Pedagadi who have been working more closely and collaborating with me on research papers. I am grateful for you to help and teach me.

My parents deserve as much thank as anyone. I always told people that the greatest blessing is my parents' dedication to my educations.They gave up uncountable things so that I could have every opportunity to grow as a person both academically and athletically. I am especially grateful for their unwavering support and encouragement throughout my life. I must also thank my beautiful wife. Without her I could not have finished the thesis as she has taken a lot of my responsibilities on her shoulders so as to allow me the time and freedom to concentrate on my thesis. The eternal support and encourage from my mother-in-law should also be paid with a big thank.

Furthermore, I would like to thank my grandparents, especially my granddad, encouraged and convinced me to start my doctoral studies. Sadly, my granddad is now unable to see me finish it. This is a specially dedication to him. I thank him for all the love, attention and education offered to me. He will be greatly missed and will live forever in my heart.

I apologize to anyone I've forgotten, or anything people have helped me with that I've left off the list.

Finally, I'd like to thank my thesis examination committee: Dr. Soodamani Ramalingam (University of Hertfordshire), Dr. Puspha Kumarapeli (Kingston University) and Prof. Anthony Walker (Chair, Kingston University) for their valuable time and constructive feedback.

Simi Wang
January 2016

List of Publications & Submissions

Conferences

- Wang, S.; Lewandowski, M.; Annesley, J. and Orwell, J.: *Re-identification of pedestrians with variable occlusion and scale.* In Computer Vision Workshops (ICCV-Workshops), 2011 IEEE International Conference on, pp. 1876 –1882 (2011).
- Wang, S.; Pedagadi, S.; Orwell, J. and Hunter, G.: *Vehicle Logo Recognition Using Local Fisher Discriminant Analysis.* In 5th International Conference on Imaging for Crime Detection and Prevention (ICDP) on, pp 1–14 (2013).
- Wang, S.; Orwell, J.; and Hunter, G.: *Evaluation of Bayesian and Dempster-Shafer approaches to fusion of video surveillance information.* In Information Fusion (FUSION), 2014 17th International Conference on, pp. 1–7. IEEE, (2014).

Video Programs

- *Real-Life Big Brother: Secret NSA Data Centres And Global Surveillance*, Produced By Journeyman Pictures, May 2013, Available at www.youtube.com/watch?v=VZxd8w11YSA, Contribution start at 03:00 minutes and ends at 04:30 minutes .
- Demonstration Films For SurPRISE Citizen's Summit, in Surveillance, Privacy and Security: A large scale participatory assessment of criteria and factors determining acceptability and acceptance of security technologies in Europe (SurPRISE), Produced By Two Cat Can, 2014 Available at https://www.youtube.com/watch?v=OaP3L8S0R3A.

Contents

List of Figures

List of Tables

Acronyms

ANPR - Automated Number Plate Recognition

BN - Bayesian Network

BS - Background Subtraction

CCTV - Closed Circuit Television

CMC - Cumulative Match Characteristic

DBM - Dynamic Bayesian Network

DNA - Deoxyribonucleic Acid

DS - Dempster Shafer

DST - Dempster Shafer Theory

FDA - Fischer Discriminant Analysis

GMM - Gaussian Mixture Model

HOG - Histogram of Oriented Gradients

HSV - Hue, Saturation, Value

IF - Information Fusion

JDL - Joint Directors of Laboratories

LBP - Local Binary Patterns

LF - Local Fischer

LFDA - Local Fischer Discriminant Analysis

LMNN - Large Margin Nearest Neighbour

LMNN-R - Large Margin Nearest Neighbour classifier with Rejection

NP - Number Plate

OCR - Optical Character Recognition

PCA - Principle Components Analysis

PDF - Probability Density Function

pHOG - Pyramid Histogram of Oriented Gradients

RGB - Red Green Blue

ROC - Receiver Operating Characteristics

SIFT - Scale-Invariant Feature Transform

SNR - Signal-to-Noise Ratio

SVM - Support Vector Machines

VCA - Video Content Analytic

Chapter 1

Thesis Introduction

1.1 Introduction

With the aim of reducing crime and increasing public safety, millions of closed-circuit television (CCTV) cameras have been installed in streets throughout the world. The United Kingdom is one of the greatest proponents, with an estimated 5.9 million cameras in 2013. This is an increase of 4 million since 2011. A large proportion of the CCTV cameras were installed in the capital city of London. Norris and Armstrong (1999) claimed that a person could be observed from 300 different cameras every day in London. This number will continually increase, whilst cameras are being installed in public transport, in business and even on drones.

The large volume of data collected by these distributed CCTV networks have proven their effectiveness in a range of recent high profile investigations, such as identifying and tracking the Boston Bombing Terrorist and prosecuting the rioters after the London Riots. Although CCTV networks also assist law enforcement agencies to detect live crimes on a daily basis, the effectiveness is limited by the ability of the CCTV operator. Donald (2010) stated that the concentration level of capable operators, dealing with high activity videos is around 90 minutes. In low activity and static environments, the concentration level drops to about 20 minutes. The operation is further limited by the number of different cameras that an operator can monitor simultaneously to effectively detect events of interest. This means that the development of video analysis techniques for the detection of predefined events is necessary. This has led the Analytic Software to become the fastest growing component of a video

surveillance system, as addressed in the Global Video Surveillance Market Report (Network, 2011).

The Market Report also showed that the use of surveillance systems is expected to divert from traditional surveillance application to other function such as Retail. One major consideration for commercial users is the cost because the largest contributor of cost in a video surveillance system is the camera. Therefore to lower the investment, many users choose to use lower resolution and intensity cameras, such as the recent Pan-Tilt-Zoom cameras deployed at the Shanghai Airport. With the improvement of technology in the clarity of images, the cost of high resolution cameras will reduce and their popularity will increase. However, the benefits of higher resolution data have not been fully explored in the video analysis research domain due to the slow uptake of this new technology and the lack of availability of high resolution datasets.

In the available datasets, almost all of the data are captured using traditional intensity-based cameras. These cameras are highly sensitive to variations in illumination. Although sophisticated techniques are employed to mitigate these effects, the issue cannot be entirely removed. Illumination is, therefore, one of the major sources of uncertainty in the outcome of many video analytical systems. In general all analytical systems will have their limitations and will have been optimised to solve a particular aspect of uncertainty associated with either the data or technique used.

1.1.1 Definition of Uncertainty

With reference to the information theory introduced by Shannon and Weaver (1949), uncertainty is the number of alternative outcomes to an event and is measured by the probability of an outcome occurring. If there is only one possible outcome and the probability of occurring is 1, there is no uncertainty. As information can only be gained when there is uncertainty, an event and the outcome that occurs with the lowest probability will convey more information. Therefore the information gained is indirectly measured as the amount of reduction of uncertainty. The amount of uncertainty is measured by Shannon Entropy, and maximised when all

alternative outputs have equal probability.

The above discussion was conducted in detail by Singh (2013). Singh stated that uncertainty could be understood as a form of information deficiency or reflecting information reductions, which was supported by Smets (1983). The manifestations of information deficiency are summarised as the followings:

- Incomplete Information: Refers to a result when some of the information is missing, therefore the information has an unknown degree of confidence but an upper limit of confidence is known (Florea et al., 2007).
- Imprecise Information: The information that could be used to describe a number of different instances, instead of just referring to one particular instance (Florea et al., 2007).
- Fragmentary Information: Refers to information that is not continuous and is only available under certain instances, time or conditions (Krell et al., 2013).
- Vague Information: Refers to information that is ill-defined, therefore it could be interpreted subjectively from one observer to the other.
- Contradictory Information: The result when information from different sources, measuring the same environment, reports opposite result (Dong and Naumann, 2009).

In the context of this thesis, the uncertainty will refer to the inaccuracy in the estimated outcomes that may be produced by the information deficiency, as outlined above. In theory the inaccuracy can be measured in terms of Shannon Entropy. In practice there may be scenarios in which this measurement is difficult to obtain as it requires a detailed statistical model. For example, changes of illumination, as the intensity varies during the time of day, could produce fragmented information.

1.2 Thesis Motivation

The motivation for this project is, therefore, to devise a framework that would reduce the level of uncertainty to improve the accuracy of the outcome, by combining different video ana-

lytic approaches. To complete it, the adoption of information fusion techniques is investigated.

A video analytical system is normally developed and optimised for a particular objective. The framework should be able to combine the results from these different analytical systems to infer a single objective and extend this inference to other related objectives.

1.3 Thesis Aims

In this thesis an investigation will be conducted on the feasibility of constructing an information fusion framework to reduce uncertainty and investigate the possibility of inferencing a range different objectives from one single framework. The investigation will be aided by the use of high resolution data to explore and utilise the benefits it brings to tackle a range of surveillance objectives.

1.4 Thesis Objectives

The following items are the main objectives of the thesis:

- To conduct a comprehensive review of information fusion techniques that are currently available ,identifying suitable candidates to be adopted within this thesis.
- To examine the suitability of publicly available datasets to aid the investigations conducted within this thesis.
- To identify and understand the nature of uncertainty involved within a video surveillance system and associated uncertainty with the current analysis techniques to resolve surveillance objectives under investigation.
- To investigate the use of high resolution data to reduce uncertainty when solving traditional surveillance objectives under various challenging environments.
- To combine state-of-the-art video analysis techniques with information fusion techniques to solve challenging surveillance tasks.

- To design, construct and theoretically evaluate suitable fusion frameworks for the reduction of uncertainty and the inference of multiple surveillance objectives.
- To construct and theoretically assess an evaluation framework which would allow an effective and direct comparison between the developed fusion frameworks.

1.5 Thesis Outline

The following chapter will give an overview for the application of information fusion in visual surveillance domain and the approaches to tackle two traditional security objectives. It includes a comprehensive introduction of the Information Fusion Model and the popular techniques used to reduce uncertainty in the outcome. This will be followed by an analysis of the evaluation metrics to measure the uncertainty removed by the different techniques.

Chapter 3 presents results from an experiment designed to measure the accuracy at which human beings can be re-identified using a colour feature vectors only. The experiments further investigates the advantages of a high resolution dataset by using a state-of-the-art classifier to examine the performance with respect to the level of occlusion, the training regime, specificity of the domain and the resolution of the observations. A method is proposed to reduce the adverse impact of occlusions when present and to increase the beneficial impact of higher resolution data, when available.

In Chapter 4 the types of objective that a video surveillance system can deal with are evaluated. This will aid the design of an experimental test-bed for capturing the required data for the investigation of this project. The chapter also outlines and evaluates the techniques for reducing the redundant information with the captured data. This is done to lower the cost and requirement to store a large amount of data. The technique was combined with tracking algorithms to produce demonstration videos , which were published in two TV programs. Following the creation of the experimental data in Chapter 5, an examination of the types of uncertainties within the visual surveillance context is given. The nature and quantity of uncertainty associated with the various video surveillance techniques are then reviewed and

examined. Using the data acquired from the test bed, Chapter 6 shows a method for localising and recognising vehicle manufacturer logos in both the front and rear views. The features are constructed from local histograms of gradients in both conventional and hierarchical arrangements. The dimensionality of these vectors are reduced by an unsupervised Principle Component Analysis and by a subsequently supervised method based on Local Fisher Discriminant Analysis. It also includes an introduction of a suitable metric for multi-class classification, combining the result with fusion techniques. The challenge of logo detection and localisation are finally conducted in this chapter.

Chapter 7 presents the application of fusion methods for a developed visual surveillance scenario. Two statistical parametric fusion methods, Bayesian Networks and the Dempster Shafer method, are developed and a theoretical investigation is conducted. This chapter also presents the development of a metric for a direct comparison of the benefits between the two methods. This metric provides a method to quantify the extra information produced by using the Dempster-Shafer method with comparison to a Bayesian Fusion approach. In the final chapter the main contributions and achievements of this thesis are discussed. The conclusion and future direction of research then follows.

Chapter 2

Review

2.1 Introduction

This chapter reviews the models and techniques of Information Fusion. The review bridges the terminologies in the Fusion research community with the researches that are conducted in the image processing community, specifically for Visual Surveillance. As the review shows, there are various techniques used to achieve video surveillance objectives and for fusion research to conduct fusion objectives. There does not seem to be a formal link between the two objectives, therefore, a review of the techniques used in Information theory, to decrease the effect of uncertainty and current applications in the video surveillance context, should be conducted.

As a key benefit of information fusion is to decrease the overall extent of uncertainty, the chapter first outlines the definition of information fusion and discusses the application of fusion models for a video analytic system. It also gives a description of the extensively used fusion techniques and their applications in visual surveillance.

2.2 Information Fusion Overview

The human eye is a powerful tool that allows the human brain to assess situations quickly. However the human eye has limitations which influence the judgement of situations, such as low visibility. In these situations, the human brain relies on the information from the other

senses (hearing, touching, smelling and tasting) to support the judgement. The brain fuses all of the available information from all of the sources to improve the certainty of the judgement. By obtaining information from multiple sources, the amount of uncertainty generated from one single source is reduced. The field of study that tries to simulate this process is called Information Fusion (IF) [1], this is a branch of Information Theory.

Since the Information Fusion's first use in military application, there has been an abundant variation of its definition depending on the specific activity or a given field of application. Boström et al. (2007) reviewed and discussed the strengths and weaknesses of IF's definition. The authors concluded a generalised definition of information fusion as:

> *"Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision making."*

The above generalised definition could be applied to all the various terms, including data fusion, sensor fusion, image fusion, decision fusion and classifier fusion. The benefits of IF can be summarised with the type of information by combining the ideas of Hall and Llinas (1997) with Durrant-Whyte (1988) as the following:

- Complementary information: can improve the spatial and temporal coverage. The information provided by a single source can only provide a fragment of the global space. By combining complimentary information from independent sources measuring different aspects of the same space, a better construction of the global space can be achieved, for instances, the information from two cameras measuring the same target from different viewpoints.

- Redundant Information: when fusing information provided by different sources (or same sources over time) to measure an aspect of the same space, the redundant information can reduce the overall uncertainty and enhance accuracy. Multiple sources providing

[1] As Hall and Llinas (1997) stated that sensor and data fusion are sometimes equivalent to Information Fusion. However in some situations, as demonstrated later, the term "data fusion" is also used for the fusion of raw data from the sensors. The fusion of raw data from a sensor is considered to be a special case of Information Fusion.

redundant information can also serve to increase reliability in case of source failure (Blum et al., 2005), such as the overlapping section of multi-camera network

- Cooperative Information: increases the performance robustness by fusing information from multi-spectral or multi-modal sensors, such as the combination of audio and video.

2.3 Information Fusion Models

2.3.1 Introduction

As mentioned above, Information Fusion was first adopted in military applications. Therefore, the most common and most popular conceptualisation of information fusion was proposed by White (1987) at the Joint Directors of Laboratories (JDL) and the American Department of Defence. White's fusion process includes an associated database with five processing levels, and an information bus that connects between the five levels. The proposed fusion system was divided into four increasing abstraction levels; object, situation, impact and process refinement. These terminologies were tailored toward the military application, and were, therefore, very restrictive. To alleviate these restrictions and to resolve limitations such as uncertainty, extensions on the JDL model have been proposed by Llinas et al. (2004), by Steinberg et al. (1999), and by Blasch and Plano (2002) who added an additional user interface model on top of the JDL model.

Dasarathy (1997) developed another popular fusion model that was employed from the engineering prospective. Unlike the JDL model, Dasarathy focused on the difference between the input and output results, independent of the fusion process. The data flow of the Dasarathy model was characterised by input and output as well as the functional process, therefore the abstraction levels were specified as whether an input or output. A simplified fusion abstraction level was given by Luo and Kay (1992), where the authors divided fusion into three levels: low, medium and high, depending on the output of fusion process. Based on these previous models, many authors have tried to generalise the fusion process based on mathematical notions, such as Goodman (1997) using random sets, and more recently Kokar et al. (2004) using category theory.

Based on the analysis of the above fusion models, information fusion, when adopted into the image processing domain can be generalised into two distinctive abstract fusion level: Low Level and High Level.

2.3.2 Low Level Fusion

Traditionally the Low Level fusion is the process to fuse raw data from multiple physical sensors. Characterised by Data level fusion as shown in Figure 2.1. It involves the fusion of raw data from different sensors, such as radar, to measure the same object before any processing is conducted. The fused information can provide data of higher accuracy by lowering the signal-to-noise ratio of individual sensors. However, its main drawback is that the data from the different sensors must be commensurate and it can be properly associated.

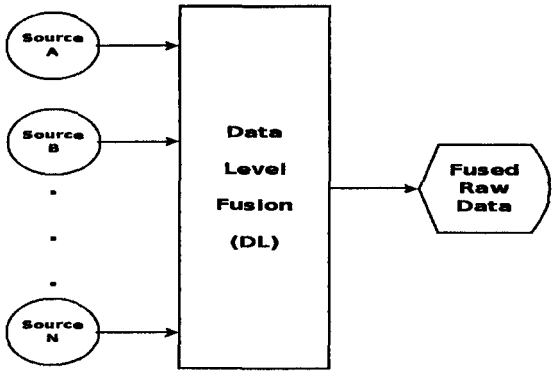


Figure 2.1: Illustration of Data Level Fusion, where raw data from the multiple physical sensor are combined to lower errors within the measured signal

In a visual domain, the raw input to any visual system is the image. The fusion of raw data is known as Image or Pixel-Level Fusion. A definition of Image Fusion is given by Blum et al. (2005) as a procedure for generating a fused image, in which, each pixel is determined by a set of pixels in each source image. The purpose is to generate a single image that contains a more accurate description of the scene than any individual source. Traditionally the application of image fusion is to produce images that could assist the human visual system to make better judgement such as medical diagnosis (Constantinos et al., 2001) and defect inspection (Leon and Kammel, 2003). Recently there has been a move to use of image

fusion for achieving results with a higher level of information such as highlighting landscape changes, as proposed by Dong et al. (2009). As the paper by Pohl and Van Genderen (1998) illustrates, there is a number of researchers which concentrate on fusion at the image level. As the authors states, the main advantage of image level fusion is to increase the clarity of the original input image in order to allow it to be processed more efficiently by humans. However these advantages can also be propagated down the image processing pipeline to assist the fusion process at a higher level, although the advantages are very difficult to measure.

In the field of video analytics, low level fusion can also refer to the fusion of raw pixels, as described above, as well as the fusion of the features related to the image. In feature level fusing, as illustrated in Figure 2.2, each information source provides some observational raw data, where some distinguishable feature vector is extracted. The features from the sources are concatenated together to form a single feature vector that is used as the input to an analysis unit in order to achieve the target objective. Fusion of the features allows for the utilisation of the correlation between them and the combined feature only requires one training phase (Snoek et al., 2005). Some drawbacks are that the features to be fused must be of a common format and that it cannot deal with fragmented information (Das et al., 2008), as features from different sources may not be available at the same instance in time.

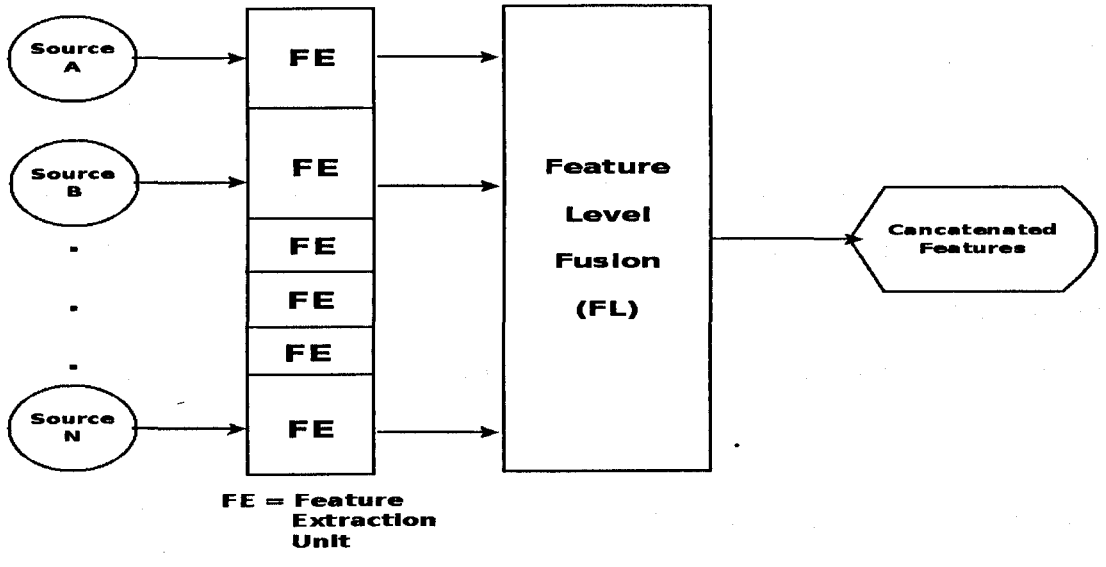


Figure 2.2: Illustration of Feature Level Fusion, where different descriptors of the same image are fused together to create a new high dimensional feature

In the visual domain, the features that can be extracted may be summarised as [2]:

1. Visual Feature: These features can be extracted from the whole image, patches within the image or from segmented blobs. The features may include histograms of a colour spaces, texture features and shape information.

2. Text Features: These features can be extracted through optical character recognition or automated license plate recognition processes.

3. Motion Features: These can be represented in the form of kinetic energy, motion direction, magnitude histograms, optical flow and motion patterns in a specific direction.

4. Metadata :Features that are associated with the captured information such as a time-stamped, global positions of the information sources, which are used to supplement the above features in a fusion system.

The features listed above have been fused in a range of visual surveillance scenarios, two examples are:

1. Face recognition. Ekenel and Stiefelhagen (2005) fused the wavelets of sub-bands of the same image. It improves the classification performance as information resulting from the sub-bands that attain individually high correct recognition rates, is fused. Tan and Triggs (2010) illustrated that combining two of the most successful local face representations, Gabor wavelets and Local Binary Patterns (LBP), gives considerably better performance than either alone.

2. Human Tracking. Foresti and Snidaro (2002) fused information from both optical and infra-red source with the trajectory information to achieve the tracking tasks under challenging conditions. Wang et al. (2003) fused motion, colour and texture cues at the feature level to perform human facial tracking and vehicular tracking in a range of environments

Like image fusion, feature fusion is generally considered a way to provide extra dimensions of information to generate more accurate outcomes by higher level process. However the

[2]The use of an audio as a feature has been omitted from the list as it is not always available with a visual surveillance scenario. But the following section would show the combination of audio and visual features used by a different researcher

term "feature fusion" has not been well documented within the literature, even though many researchers have endorsed the concatenation of various features, such as Dikmen et al. (2011) who concatenated colour histograms in the HSV and RGB colour space.

2.3.2.1 Example: Fusing of Histograms

There are various features that can be used for the pedestrian re-identification challenge. One of the most popular features is the use of colour, in particular colour histograms, which this review will concentrate on.

The seminal work by Swain and Ballard (1990) demonstrated colour indexing for retrieval; histograms were compared by using an 'intersection' operator that is similar to the L_1 norm [3]. The histogram is a non-parametric, quantised representation of the accumulated values. One alternative is the parametric family of representations, e.g. second order statistics (Metternich et al., 2010), possibly with mixture estimation (Tuzel et al., 2006). Another alternative is to find and represent multiple salient points in the observation, e.g. using SIFT features (Sivic and Zisserman, 2003).

Park et al. (2006) extended the histogram-based representation by dividing the region of interest into horizontal partitions to form histograms concatenated into a fused feature vector. Each partition can be considered a raw data sensor, from which, features are extracted before it is fused through the concatenation process. This is a special case of a more general set of robust computer vision methods, in which, overlapping regions are used to achieve spatial selections with spatially tolerant accumulators, as presented by Dalal and Triggs (2005).

Gray et al. (2007) introduced colour histograms based on three predefined regions of a human body: one fifth for the top, two fifths for the middle and two fifths for the bottom. The division, as outlined by Gray et al. , presumes a creation of a better descriptor by segregating more noisy background pixels in the head region. The two largest regions would occupy a larger region of the divided images therefore less noisy background pixels. The combined histograms for all three regions are used as the descriptor for the whole image. An improved

[3] L_1 norm minimise the sum of the absolute differences between the target value and the estimated values

descriptor is proposed by Alahi et al. (2010), using a grid collection of region descriptors. Each grid segments the objects into a different number of sub-rectangles of equal sizes.

Various methods have been proposed to generalise this approach. First, Gray and Tao (2008) used a boosting technique to optimise a set of histogram features from a large combinatorial space. Zheng et al. (2011) constructed histograms for each of over twenty types of features for six horizontal stripes across the bounding box. Finally, Dikmen et al. (2011) applied an array of histogram responses, extracted from overlapping regions, to form the fused feature. More details on the the Dikmen et al. approach will be given in Chapter 3.

Due to the considerable increase in the dimensions of the fused features, there is generally a limit to the number of features to be fused, around about 3, as it would require an increase in the processing of requirements . To reduce the processing power, the fused features would generally require its dimensions to be reduced before the analysis is done, to increase the rate of convergence. However, the dimensional reduction, in some instances, will reduce useful information. the loss can be reduced by using a higher level processing system, such as Decision level Fusion.

2.3.3 Decision Level Fusion

Decision level fusion, as illustrated in 2.3, is the highest level of abstraction. Each extracted feature is first analysed by a processing unit to acquire a decision on its identity. A decision analysis unit makes a final decision on the hypothesis by fusing each individual decision based on the provided feature or features. Decision level fusion avoids a majority of the shortcomings in the feature fusion, such as the requirement that different features needs to be in a common format. However the analysis unit's decisions usually have the same representation, such as probability of a hypothesis. Decision level fusion allows the fusion of unlimited features and offers a level of flexibility as the most suitable analysis procedure for each feature can be chosen. Although there is more freedom offered by fusion at the decision level, the main drawbacks are that the training of the processing units for each feature would increase the overall processing time of the fusion system.

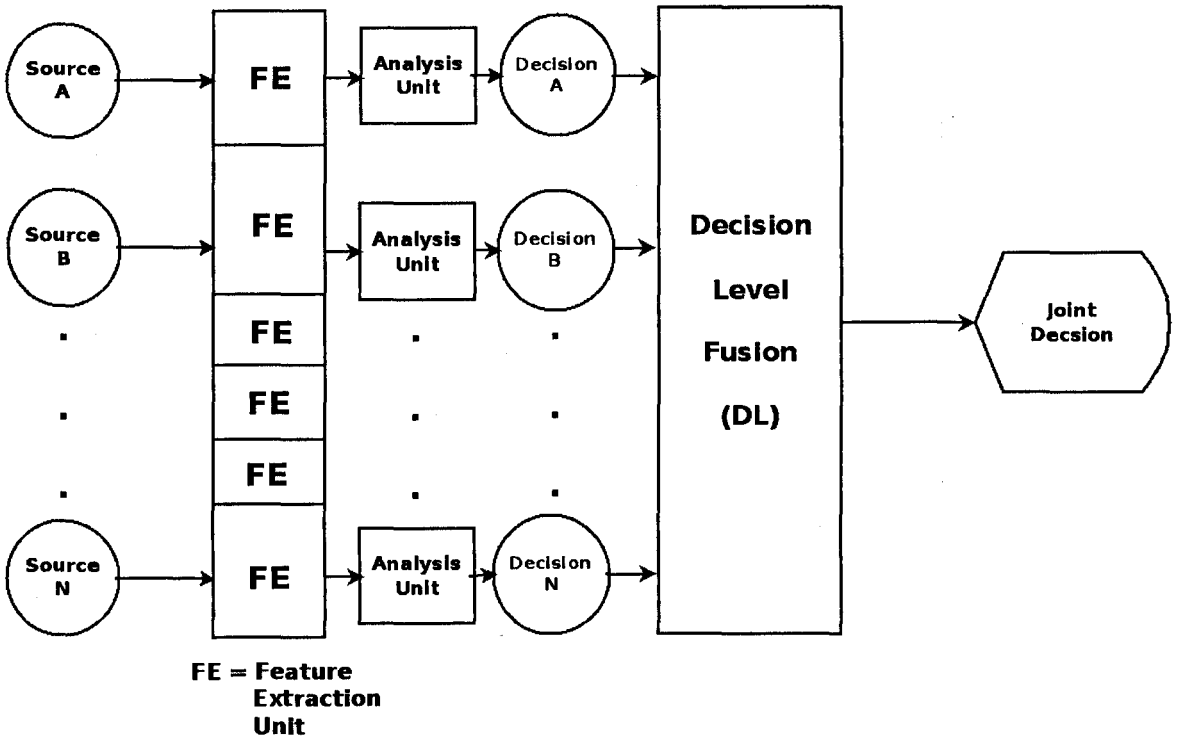


Figure 2.3: Illustration of Decision Level Fusion, where decisions calculate the different video analysis units, based on different raw features, are fused to acquire a final decision

As the information provided by the decision level would require the least amount of human processing power for situation assessment, it has been widely adopted across different fields of research. A review of the techniques used is broken down, and the adoption in the surveillance domain is analysed in Section 2.4.

2.3.4 Hybrid Fusion Model

Each increase in the fusion abstraction level decreases the amount of human processing required for situational assessment. This has focused the research community on the decision level, although each level has its own benefits. To utilise these benefits such as the correlation of the feature at the feature level, some researchers have used a hybrid model that fused together different levels of fusion abstractions. An example offered by Wu et al. (2004) is the fusion of the multiple independent features as one input modality and then fusing multiple classifier results at the decision level. Other uses of the hybrid model in the video domain can also be found in event detection (Xu and Chua, 2006), and pedestrian tracking (Snidaro et al., 2004).

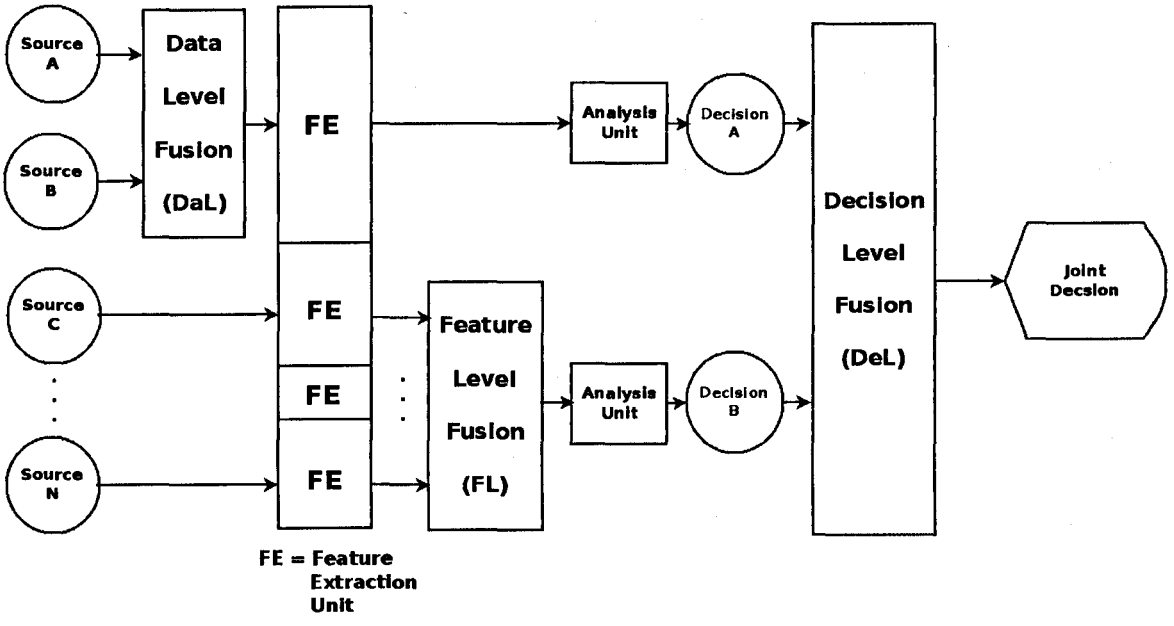


Figure 2.4: Illustration of Hybrid Level Fusion, where different benefits at all abstracts level could be combined to create a more accurate fusion framework.

2.4 Decision Fusion Techniques

2.4.1 Introduction

Unlike other research fields where sensor information fusion have been used extensively, such as fault diagnostics (Basir and Yuan, 2007), computer intrusion detection (Giacinto et al., 2003) and a range of military applications (Hall and Llinas, 1997), the IF is a relatively new technique for automated video surveillance with comparatively few publications. Although various classifications of information fusion methods and techniques have been proposed by various researchers (Castanedo, 2013; Pohl and Van Genderen, 1998; Khaleghi et al., 2013; Nakamura et al., 2007; Bloch, 1994) based on a range of criteria such as the type of data, purpose of the techniques, parameters, and mathematical foundation. The categorisation of the fusion techniques in the surveillance vision field has not been conducted.

At the heart of many video analytic applications is a classifier, fusing different results from various classifiers using a broad range of information sources. This can be used to increase

the accuracy of the final classification results as outlined by Ruta and Gabrys (2000).The choice of the methods used to fuse the classifiers may depend on the output of each of the classifiers. Xu et al. (1992) distinguished three types of classifier output:

- Abstract Level - The classifier would only output one unique label. At this level there is no information about the certainty of the guessed labels, nor are any alternative labels suggested.
- Rank Level - The classifier outputs a ranked list of all possible labels. The highest label is the first choice and the alternatives rank in order of plausibility of the correct label.
- Measurement Level - The classifier attributes each label a measurement value to represent the supporting probability for the hypothesis. The input vector submitted for classification comes from each of the classes.

Based on the three categories of classifier outputs, the review of the fusion techniques can be broken down into two main sections:

1. Logical Reasoning: Mainly uses the classifier outputs from the Abstract and Rank Level.
2. Evidential Reasoning: Uses outputs from the measurement level.

The mathematical foundation of the methods described in the following section will be covered in the relevant chapters. Therefore, the review will give an overview of how these methods are currently being used in the vision community.

2.4.2 Logical Reasoning Techniques

These techniques relate to logical methods that assist with the decision making process. The simplest method is the majority vote. The identity that receives the largest number of votes from the individual processing units is selected as the consensus decision. Oliveira et al. (2010) presented a pedestrian detection system by employing multiple classifiers of different extracted features. The authors fused class labels from both Support Vector Machines (SVM) and a Multilayer Perceptions classifier using two features; Histogram of Oriented Gradients (HOG) and local receptive fields. The output scores of all of the classifiers were fused to

obtain a majority decision regarding the identity of the object.

The majority voting process is used when no prior information is known, therefore, all classifiers are assumed to have equal accuracy. However if some prior knowledge is known such as changes in lighting condition due to time of day, some classifiers would have increased reliability over others. In these situations, it is more suitable to assign a higher confidence to the more competent processing units in the decision making process. These methods are called Weighted Majority Voting. The weights assigned to the classifiers are normalised to 1. There are various normalisation methods that can be adopted, as outlined by Han et al. (2006). The most important aspect of using weighted majority voting is to determine and adjust the weights to achieve the optimal accomplishment of the decision. The weighted voting process was adopted by Foresti and Snidaro (2002) for the tracking problem. The authors fused the position of the blob from multiple information sensors at a certain time. Each sensor has a corresponding weight according to the reliability factors, and the point that best represents the object's position is given by the weighted majority voting process.

These techniques require the results of the classifiers to be of a common format, for example, if all of the classifiers were to investigate "Is a probe vehicle probe image the same as a target vehicle?". Under the condition, the voting process can be used to acquire a more accurate result. However if the used classifier outputs attributes of the vehicle such as the colour or shape, these attributes alone , employed to determine the similarity of two vehicles, would be difficult when the voting techniques is used. In this case, the evidential reasoning technique can be adopted because it is able to accept these different types of attributes to build a likelihood model to determine the outcome.

2.4.3 Evidential Reasoning

Evidential reasoning methods are based on the knowledge of the perceived situation. Evidence, thus, refers to the transition from one likely true proposition to another. The truth is believed to result from the previous one, as stated by Nakamura et al. (2007). Classic evidence methods are based on subjective probabilities. Two very popular concepts have been chosen by the

wider research community: Beyesian and Dempster Shafer.

2.4.3.1 Bayesian Inference

Information fusion based on Bayesian Inference offers a formalism to combine evidences according to rules of probability theory. Uncertainty is represented in terms of posterior conditional probabilities describing the belief of a hypothesis It can assume values in the interval $[0, 1]$, where 0 is the lack of belief and 1 is absolute belief. The posterior probability density function is produced by using Bayes rule, and relies on the prior belief of the hypothesis and the probabilistic likelihood function that describes the probability of the hypothesis given an observation, (Castanedo, 2013). Bayesian inference allows a range of prior knowledge about the likelihood of the hypothesis utilised in the inference. It also allows for the probability of the hypothesis being true to be computed incrementally. As the posterior probability from one likelihood function can be used as the new prior probability, it can be used to update the posterior probability of hypotheses.

Applications of the Bayesian inference in a surveillance domain include the following: Atrey et al. (2006) adopted a Bayesian inference fusion approach to fuse audio features and video features to detect pre-defined events, and Stolkin et al. (2012) applied a Bayesian method to the tracking problem by combining images from a thermal imaging camera and conventional colour cameras. The Bayesian approach is able to adjust the relevance of the cameras in the tracking decision process. The tracking of vehicles by fusing different features using a camera network for monitoring the highway was presented by Huang and Russell (1998). The authors computed the probability of any two objects being the same, given a stream of the features of the observation such as timestamps, mean colour, and forward velocity.

An extension of the Bayesian inference, popular in the research community, is Bayesian Network. It is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. The nodes within a Bayesian network represent the random variables or observation from different information sources and the edges represent the conditional dependencies. Dynamic Bayesian network (DBN)

models are an attractive modelling choice when fusing information from multiple sources, as they combine an intuitive graphical representation with efficient algorithms for inference and learning (Choudhury et al., 2002). Choudhury et al. proposed an application of the Bayesian Network to combine different speech and visual cues for the classification of speaker interaction. Toyama and Horvitz (2000) performed head tracking by fusing different trackers of colour and motion features. The authors also made use of the random variables in a Bayesian network to serve as context sensitive indicators of the reliability of the different trackers. Town (2007) also utilised the Bayesian networks to model the probabilistic dependencies and reliabilities of different sources of information. In Town's work, integrating visual information from video cameras with ultrasonic sensor data at the decision level allowed for the tracking of people within an office environment.

2.4.3.2 Dempster-Shafer

Hall and Llinas (1997) illustrated that Bayesian inference has two major shortcomings: firstly it requires a well-defined, prior and conditional probabilities of the hypothesis. Secondly the hypothesis needs to be mutually exclusive, to solve these limitations in probabilistic methods, a number of alternative techniques have been proposed. One of the popular technique is based on evidence, called as the Dempster-Shafer Theory (DST).

The Dempster-Shafer Theory is based on the mathematical theory introduced by Dempster (1967) and mathematically formalised by Shafer (1976) toward a general theory of reasoning based on evidence. It became a popular method because it could be considered as a generalisation of the Bayesian Inference that deals with probability mass functions. Unlike Bayesian inference, DST does not require a well assignment of the prior probability; instead the probability is assigned, only when the supporting information is available. DST also relaxes on Bayesian's restriction on mutually exclusive hypothesis so that it is able to assign evidence to the union of hypothesis.

As the heart of DST, a hypothesis in a set of all possible, mutually exclusive hypotheses is characterised by a belief and a plausibility that represents the lower and the upper bounds,

respectively, of the hypotheses being true. The interval bounded by the belief and plausibility values defines the true belief in the hypothesis.

Some applications of DST in the video surveillance domain include: Maguire and Desai (2012) explored DST's ability to allow each source to contribute information with different levels of detail, to fuse a wide range of sensors to the problem of intrusion detection. Ma et al. (2013) built a fusion framework based on DST to increase the event detection rate when event detectors have conflicting decisions. Morbee et al. (2010) used DST to fuse ground occupancies computed from a set of cameras so that the occupancy detection results outperformed probabilistic methods. Zhao et al. (2007) applied DST to the vehicle detection problem by fusing signals from a video sensor and the magnetic sensor. Li et al. (2013) reported human tracking that explored the spatio-temporal visual information and used DST to fuse different classifiers to perform better tracking.

2.4.3.3 Fuzzy Set Theory

Fuzzy set theory is an alternative reasoning scheme for the fusion of uncertain information. It was first introduced by Zadeh (1965) and is an extension of the Set Theory. In contrast to Set Theory where the membership of elements in a set is assessed in a binary value according to the bivalent conditions, Fuzzy set theory introduces the novel notion of partial set membership, which enables imprecise reasoning.

Similar to Bayesian theory where prior knowledge of probability distributions is required, fuzzy sets theory requires prior membership functions for different fuzzy sets. Therefore, fuzzy set theory deals in gradual membership functions of an element in the interval $[0, 1]$, where the higher the membership function is, the more an element will belong to the set. Fuzzy data can then be combined using fuzzy rules to produced fuzzy fusion outputs. Fuzzy fusion rules are classed into three main categories:

- Conjunctive - are considered appropriate when fusing data provided by equally reliable and homogeneous sources.
- Disjunctive - are deployed when at least one of the sources is deemed reliable, though which one is not known, or when fusing highly conflictual data.

- Adaptive - are used as a comprise between the two above therefore these can be applied in both cases.

As concluded by Khaleghi et al. (2013), in contrast to the Bayesian and Dempster theories, which are well suited to modelling the uncertainty of membership of a target in a well-defined class of objects. Fuzzy sets theory, however, is well suited to modelling the fuzzy membership of a target in an ill-defined class. As such it is often integrated with both the Bayesian and Dempster-Shafer fusion algorithms rather than being used independently in challenges relating to uncertainty reduction.

2.4.4 Uncertainty Evaluation Metric

The aim of a performance evaluation metric is to establish the advantages and disadvantages of a technique based on a set of measures or metrics. These metrics can also be used to effectively compare and evaluate the outcomes of different techniques operating to the same objective. As Khaleghi et al. (2013) suggests, the results from Information Fusion Systems are typically a mapping of different techniques into different real values or partial orders for ranking.

In these goal-orientated challenges, a range of methods can be used to evaluate the merits of a fusion framework. Some applications used in the video surveillance domain, to evaluate the performance of object identification, are the Precision and Recall, Receiver Operating Characteristics (ROC) and Cumulative Match Curve (CMC). These measure the performance of tracking analysis metrics such as mean distance from track, detection rate, and false positive rate (Town, 2007). For event detection, a multi-metric evaluation procedure including certainty, accuracy and timeliness has been proposed by Hossain et al. (2011).

However, as the goal of many fusion systems is to reduce uncertainty, a metric to evaluate the uncertainty reduction is needed. With the logical reasoning techniques, the uncertainty reduction is normally carried out by the improvement of accuracy using the metrics mentioned above. The evidential reasoning techniques can adopt a similar measurement strategy, however since there is probability associated with the outputs, information theory should be

used, such as Shannon's Entropy that is a measure of uncertainty, as highlighted in Chapter 1. However Shannon's Entropy can't be applied to the Dempster-Shafer model as no prior information is available; therefore, the reduction of uncertainty, before fusion, is unknown.

The review of the evaluations suggested a need to develop standardised measurements of uncertainty reduction that can be applied to different evidential reasoning techniques, to provide an effective comparison between the techniques under investigation.

2.5 Summary

This section reviewed some of the techniques that have been used in the vision community to fuse information from different sources for achieving better results. The choice of the techniques depends on the information provided and the availability of key information.

As mentioned in Section 2.4.2, when the information provided by a range of sensors is common, a logical reasoning technique can be used. If the information is not common but with some supporting evidence, the evidence reasoning techniques can be adopted to combine these types of information. In the different evidential reasoning techniques, the Bayesian inference would require some types of prior information about the outcome hypothesis. When prior information is unavailable, the Dempster-Shafer technique might be more beneficial. The section also highlighted the use of the Fuzzy set theory; however it is often used in combination with Bayesian and Dempster-Shafer theory to solve challenges involving uncertainty. As such the research will concentrate on Bayesian and Dempster-Shafer fusion frameworks.

This section also highlighted that the fusion approaches mainly concentrated on the fusion of multiple physical sensors and the fusion of different information provided by the same sensor are limited. With the vehicle re-identification problem, Sumalee et al. (2012) tried to combine different physical sensors with vehicle features such as colour, shape and size derived from video image data using different image processing techniques. These features are fused by using an evidential reasoning fusion framework, in order to provide a probabilistic measure for the re-identification decision. Similarly, Kumar et al. (2010) applied Fuzzy Logic modelling

to produce a 'belief mass' for each of the sensors, and addressed the pedestrian tracking issue in varied illumination conditions.

Recently, Torabi et al. (2012) improved Kumar's idea by applying fusion techniques to track multiple people walking in close proximity. In their work, data from colour-based and thermal sensors were fused to achieve the task of tracking peoples in both indoor and outdoor environments in varied lighting conditions.

For the identification of vehicle types in surveillance data, Sun et al. (2004) used an 'inductive loop signatures' system, built with some specialised equipment and reported an accuracy of 90% on a three-category problem. Sumalee et al. (2012) employed similar ideas and applied them to the problem of vehicle re-identification within video. The authors also introduced other vehicle features such as colour, shape and size, which were derived from video image data using different image processing techniques. These features were fused by using a probabilistic fusion technique in order to provide a probabilistic measure for the re-identification decision. They claimed that the overall re-identification accuracy was about 54%, which represents the current state-of-the-art technique. This highlights the need to design and implement a fusion system that can handle multiple features from a single physical sensor.

As illustrated throughout this review, the motivation for most researchers to use information fusion techniques is to increase the accuracy by utilising its uncertainty reduction capabilities. The performance of the fusion system has been compared to standard non-fusion techniques, which use standard goal-oriented metrics. There are, however, limited means of uncertainty reduction measurement and comparison. The available metrics are inadequate when comparing some fusion methods, especially between different evidence-based approaches. This leads to a need for the development of an evaluation framework that would effectively measure and compare uncertainty reduction.

Chapter 3

Person Re-identification

3.1 Introduction

Pedestrian re-identification is an important component of visual surveillance analysis in public space. The ability to assign a single correct identifier to multiple observations of an individual, improves the semantic coherence of the analysis. This, in turn, is useful to construct descriptions of behaviour and to facilitate the retrieval of data relevant to a given individual.

The difficulty of the problem is partly determined by the extent of time and space, over which, these "multiple observations" are recorded. At one extreme is part of the pedestrian's tracking problem within a single view; this is particularly important in more crowded or occluded scenes. Similarity of appearance is used (Breitenstein et al., 2009) alongside spatio-temporal measurements to estimate trajectories of individual person. In these cases, the appearance and pose of a person being observed are relatively similar. The problem is more difficult when using observations from multiple cameras and with discontinuous trajectories, not only pose, viewpoint and illumination vary, but the number of possible alternative candidates is most likely increased. In the extreme case, re-identifying a pedestrian at some arbitrary location and future point in time is indeed a challenging problem. Some aspects of the pedestrian's appearance such as clothing and hair may have changed, and the less changeable aspects such as face and gait are more difficult to analyse (assuming that the pedestrian is not actively co-operating with the analysis and that the environment is relatively uncontrolled). Furthermore,

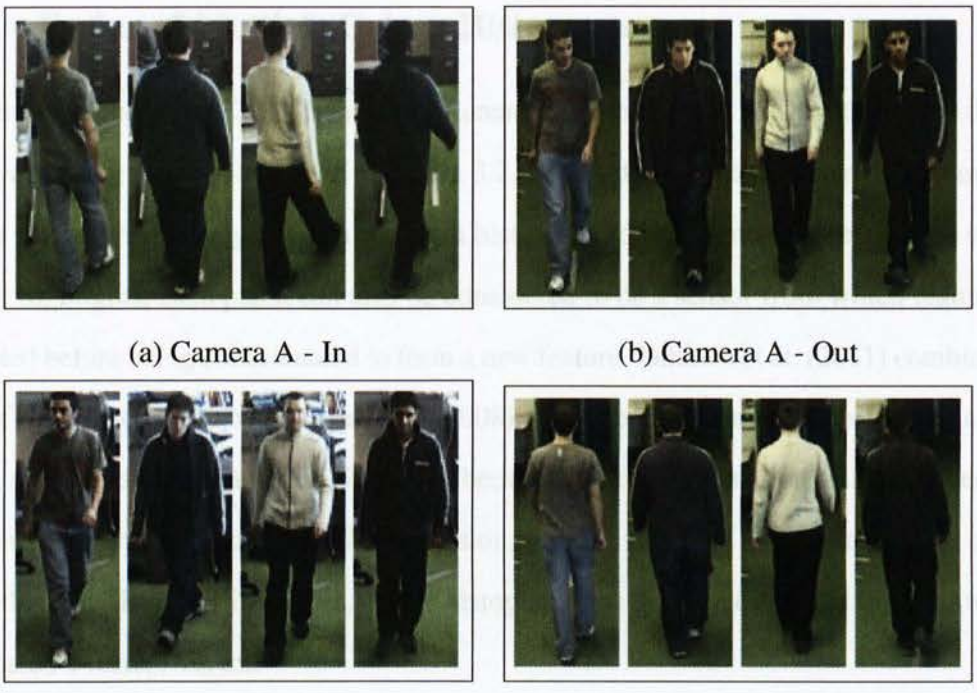
re-identifying people "in the wild" will require robust processing of partial and incomplete observations due to crowds and clutter.

Dikmen et al. (2011) methodology is regarded as the current benchmark method for pedestrian re-identification, since it provides the best performance when evaluating an unseen a subset of the VIPeR dataset. This chapter will examine how the current state-of-the-art techniques performance can be retained and improved under varying experimental conditions. The two principle sources of variations are the resolution of the images in dataset and the presence of occlusions in the probe and target observations.

3.2 The V-47 dataset

In this investigation a new higher resolution dataset is going to be used. One motivation for the proposed dataset is to block a gap within the publicly available datasets which is of a lower resolution Other benefits of the dataset is that it includes occlusion that blocks part of the body, this is useful to solving re-identification problems in crowded environments where only parts of the person could be observed.

The V-47 dataset comprises videos of 47 pedestrian walking in and out of a room through a pre-defined indoor route, observed by two progressive scan high resolution cameras. The scene had both artificial and natural lighting, which varies throughout the duration of the filming activity. There are 4 video sequences for each pedestrian (two cameras, two directions), with each sequence around 30 seconds. For each of the participants, 4 images were extracted to one for each view and some examples of the images are illustrated in Figure 3.1.



(a) Camera A - In (b) Camera A - Out

(c)Camera B - In (d) Camera B - Out

Figure 3.1: Example Images of Participants Viewed from Both Cameras, both coming in and going out of the room. All the data used will be captured at a relatively similar distance away from the camera.

3.2.1 Variable Image Resolution

In the V-47 dataset, the height of each pedestrian (in image scan-lines) varies from 140 to 480. The method outlined in Section 3.3.2 was designed and tested on a vertical resolution so that each pedestrian was 128 lines tall. Higher resolution input can always be down sampled as necessary to fit the expected input size, but this will discard higher spatial frequency signals that may aid the discrimination task. Operating directly on the original resolution signal gives the opportunity to preserve this information. The operation to accumulate RGB (or HSV) values from overlapping patches into histograms can be generalised to accept input of any size, by scaling the size of these input patches accordingly. However, the aggregation of pixel values into histograms also discards information, as mentioned in Section 5.3.1. Nevertheless, it is hypothesised that by using smaller scale patches, this effect is mitigated and better use is made of the higher resolution input.

3.2.2 Fusion of Multiple Colour Histograms

As mentioned in Section 2.3.2.1, in order to create a feature vector, the input image is divided into overlapping patches, as shown in Figure 3.2. Each patch can be considered as a bounding box of set width and height, from which, a histogram for that patch is created. As section 2.3.2.1 highlights, each patch can also be considered to be a sensor from which features are extracted before being concatenated to form a new feature. Dikmen et al. (2011) combined the idea of patches, outlined by Gray and Tao (2008), with the idea of multiple features in a single patch, as offered by Zheng et al. (2011), to better aid the discrimination task. To investigate the benefits of this additional fused information, for each patch an 8 bin histogram is created for both the RGB and HSV spaces. These histograms are then concatenated together to form each patch's feature vector.

3.3 Methodology

3.3.1 Overview

An overview of the method used for the investigation is illustrated in Figure 3.2.

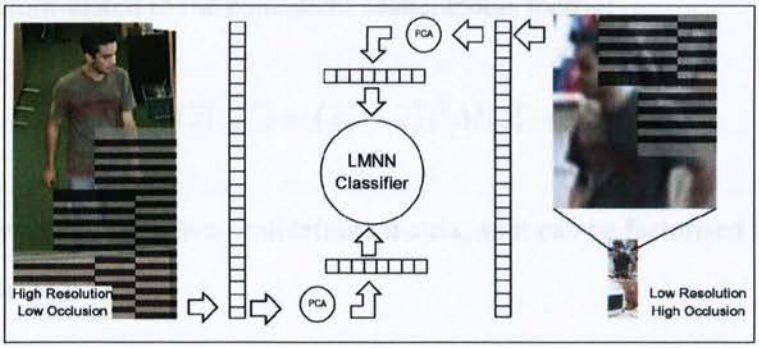


Figure 3.2: Overview of the Techniques Used. The fused feature vector for each patch (black/grey boxes) would be concatenated together to create an image vector. The image vector's dimensions will be reduced, via PCA, before the learning metrics are applied.

3.3.2 Large Margin Nearest Neighbour Classifier with Rejection

This section makes a brief description of the usage of the Large Margin Nearest Neighbour classifier with Rejection (LMNN-R), as introduced by Dikmen et al. (2011), to learn a metric

suitable for pedestrian re-identification. This is the current state-of-the-art approach for person re-identification on the benchmark VIPeR (Gray et al., 2007) dataset and it is an enhancement of the Largest Margin Nearest Neighbour techniques offered by Weinberger et al. (2006).

To increase the performance, the fused feature vector is reduced to a smaller metric learning space by applying a Principle Components Analysis (PCA) (Jolliffe, 2005), thus giving an input vector $\vec{x}\ \varepsilon\ \Re^d$.

The objective is to learn a linear transformation $L : \Re^d \rightarrow \Re^d$, where $\Re^d$ is the real number that minimises a distance between each low dimensional training point and its K nearest similarly-labelled neighbours. This is done while also maximising the distance between all differently labelled points, and while maintaining a constant minimum margin between differently labelled points. As a consequence, a similarity measure of the pairwise feature vectors would follow the weighted squared distance:

$$D(\overrightarrow{x_i}, \overrightarrow{x_j}) = \|L(\overrightarrow{x_i} - \overrightarrow{x_j}\|^2 \tag{3.1}$$

which can be reformulated to the equivalent Mahalanobis metric:

$$D(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} - \overrightarrow{x_j})^T M(\overrightarrow{x_i} - \overrightarrow{x_j}) \tag{3.2}$$

where M is a symmetric positive-semidefinite matrix, so it can be factorised into real-valued matrices L as $M = L^T L$.

The objective function over the distance metrics parametrised by equation 3.1 or equation 3.2 has two competing terms: $\epsilon(M) = \epsilon_1(M) + \epsilon_2(M)$. The first term penalises large distances between each point i and its neighbours j according to Euclidean norm:

$$\epsilon_1(M) = \sum_{i,j \rightsquigarrow i} D(\overrightarrow{x_i}, \overrightarrow{x_j}) \tag{3.3}$$

where the $j \rightsquigarrow i$ denotes that x_j is one of the K similarly labelled nearest neighbours of x_i. while the second term penalises small distances between each point and all other differently

labelled ones:

$$\epsilon_2(M) = \sum_{i,k}(1-\delta_{i,k})\left[1+\frac{1}{NK}\left(\sum_{m,l\rightsquigarrow m} D(\overrightarrow{x_m},\overrightarrow{x_l})\right) - D(\overrightarrow{x_i},\overrightarrow{x_k})\right]_+ \tag{3.4}$$

Here, $\delta_{i,k}$ is an indicator variable which is 1 if and only if x_i and x_k belong to the same class, and 0 otherwise. The x_k for which $\delta_{i,k} = 0$ are so called "impostors" for x_i. The closest impostors of a training point x_i are forced to be at least a certain specified distance, called the margin, away from the considered point x_k. This distance is computed using the average distance between all K nearest neighbour pairs (m, l) in the training set The expression $[z]_+ = \max(z, 0)$, in equation 3.4 denotes the standard hinge loss, which is a loss function used for training of maximum margin classifier (Rosasco et al., 2004). This optimisation process can be solved as an instance of a positive semi-definite program when distance D is given by the equation 3.2.

3.4 Generalising Over Occlusions

CCTV observations of pedestrians are often partly occluded due to crowded environments and obstacles. It is important for re-identification methods, such as LMNN-R, to perform robustly in these cases. However, all pedestrians of the VIPeR dataset are fully visible. To investigate the re-identification performance in the presence of occlusions, a set of partially occluded pedestrians was synthesised from the VIPeR dataset, and an occluded subset of the V-47 dataset was also used.

The VIPeR dataset was produced by overlaying another randomly selected pedestrian (from the same dataset) on top of the target pedestrian, using a feathered elliptical mask similar to that described by Weinhaus (2013). The overlaid placement was varied stochastically, with a mean occlusion level of 50%, to simulate typical observations taken from crowded scenes. In the V-47 dataset, the nature of the scene resulted in a specific subset of the pedestrian observations included real occlusions from approximately the waist down, and these instances formed the V47 partially occluded dataset. A few examples of occluded images are depicted in Figure 3.3. Such occluded images can be directly fed into training/testing procedure of

LMNN-R classifier as described in Figure 3.2.

(a) Synthesised from VIPeR dataset

(b) Real occluded images from V-47 dataset

Figure 3.3: Examples of occluded observations, where the occlusions have all been limited to the lower half of body only.

Firstly, an experimental process is designed to investigate whether a classifier trained on occluded examples will improve the performance when applied on an occluded test set compared with a classifier that is trained only on non-occluded data. However the performance of occlusion trained classifier may deteriorate (compared with the benchmark) when tested on the original (un-occluded) test set. To balance these opposing factors, a second experiment process using a mixed test set of (for example) 50% each of occluded and non-occluded data is conducted.

Two strategies for constructing a classifier to work on this mixed dataset are considered. These are:

1. A single hybrid classifier, trained with a mixed training set.

2. A joint system with two classifiers, one trained on occluded and another trained on non-occluded data. This strategy will need an additional component to detect if either probe or target observation is occluded. Even though it is currently unavailable, its performance can be estimated by simulating:

   (a) Upper Bound - simulating a perfect occlusion detector.

   (b) Lower Bound - simulating a random occlusion detector.

   This will provide a preliminary indication of relative performance between these strategies.

3.5 Experimental Results

A common evaluation framework, as presented by Gray et al. (2007) and Dikmen et al. (2011), is to offer a single 'probe' image together with a 'target' set that exactly contains one observation of the probe which may be captured at a different instance in time. The output is a ranked list of elements of the target set which can be aggregated over a test set into a Cumulative Match Characteristic (CMC) curve. The normalised area under this curve is a straightforward and intuitive performance indicator. Under the following experiments, these values would be expressed in the legend where possible.

The results of five different experiments are presented here to investigate the effects of occlusion, change in resolution, any dependency on pose, similarity of the training set to the test set and the effect of feature fusing. Where possible, reference is made to the benchmark experiment provided by Dikmen et al. (2011). In all experiments, the pedestrians in the test sets were never included in the training sets. To generate statistically significant results, a cross validation procedure is used. The data is randomly divided into training and testing datasets before a result is acquired. The random split is accomplished by repeating the iteration of a random number generator. The generator's range is limited by the total number of pedestrians and the number of iterations is guided by the number of training samples required. The final result is the average over 10 iterations of the random split process. Another configuration constant in the experiments is the retention of the first 60 principle components.

3.5.1 Occluded and Non-Occluded data

As discussed in Section 3.4, this experiment process measures the re-identification performance on occluded observations. The first step is to measure the performance on datasets, with and without occlusions. The classifiers in Table 3.1 are trained only on one type of data, and the normalised area under CMC (%) is reported in each case.

Table 3.1: Comparison of performance employing data with & without occlusions, where **C** denotes Occlusion data

notation	Training set	Test set	Performance	
TrTe	no occlusions	no occlusions	95 %	benchmark result
TrTe**C**	no occlusions	with occlusions	80 %	
Tr**C**Te	with occlusions	no occlusions	74 %	
Tr**C**Te**C**	with occlusions	with occlusions	87 %	

From the results in Table 3.1, it is clear that the underlined results with occluded data pose a more challenging problem. The classifiers achieved their best performance when the testing and training data contained the same non-mismatched data. When classifiers use the mismatched data for training or testing, the best performance is achieved by classifiers that are trained on non-occluded data.

Taking the problems of achieving the best possible performance, with occluded and un-occluded data into consideration, three results are presented: the hybrid classifier strategy, the upper bound joint strategy (perfect detector) and the lower bound joint strategy (50/50 detector), as described in section 3.4. The upper bound simulates a perfect detection between occluded and un-occluded input, while the lower bound simulates a randomly generate detection results, roughly 50%, between the two cases. In addition the hybrid classifier was trained with data that was half of each type.

To allow comparison of the experiments for each iteration, the same test data was used and it contained a random mixture of occluded and non-occluded data, roughly 50/50.

Table 3.2: Comparison of performance between the employment of various strategies to re-identify data containing both occluded and un-occluded data

Strategy 1: hybrid classifier	83 %	benchmark result
Strategy 2: joint classifiers (perfect occlusion detection):	91 %	
Strategy 3: joint classifiers (random occlusion detection):	84 %	

The results presented in Table 3.2 reveal that it is worthwhile to pursue a strategy of training specific classifiers with occluded and non-occluded observations respectively, rather than training a single hybrid classifier with both types of data.

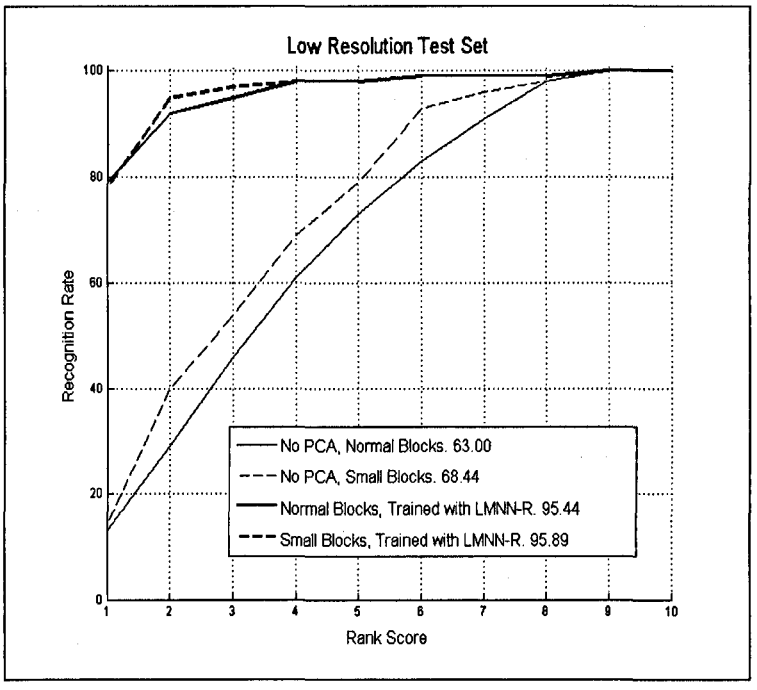
3.5.2 Higher Resolution Observations

As the original VIPeR image dimensions is only 128x48 pixels, the feature vector created in the benchmark method (Dikmen et al., 2011) makes use of 38 vertical and 4 horizontal overlapping patches, as the best performance was achieved. However the V47 dataset has 480x264 pixels therefore the performance might be improved by increasing the number of vertical and horizontal overlapping patches. For simplicity the results refer to the following:

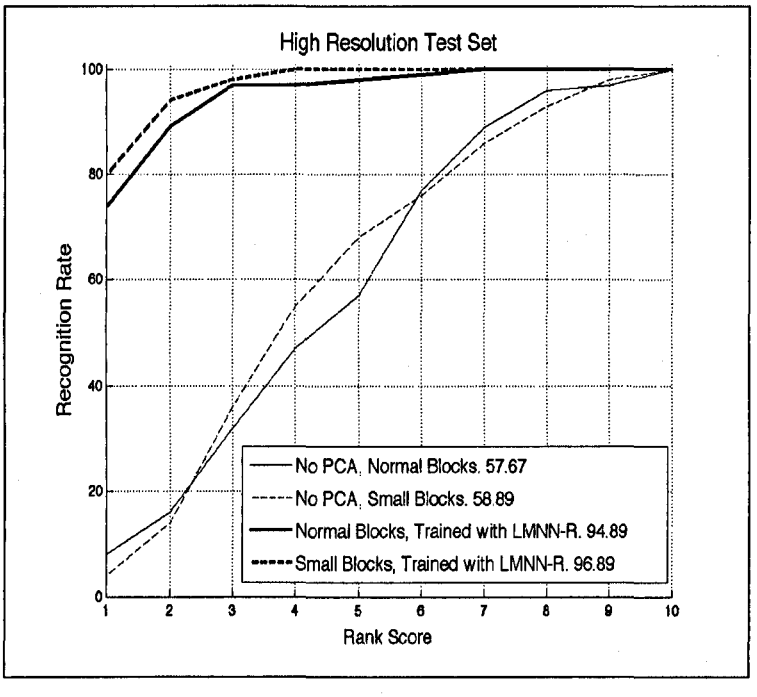
- Normal Blocks = 38 vertical and 4 horizontal overlapping patches
- Small Blocks = 47 vertical and 5 horizontal overlapping patches

To investigate and clearly identify if an improvement can be made by increasing the number of overlapping patches. A bi-cubic interpolation (Keys, 1981) was carried out to create a low resolution version of the V-47 dataset, which is the same size as the VIPeR dataset.

Experiments were conducted to compare the Normal Blocks as used in Dikmen et al. (2011) with the Small Block. These alternate schemes were tested on both low and high resolution versions of the V-47 datasets, and the results were plotted in Figure 3.4. The numbers in the legend describe the normalised percentage area under the CMC curve. The training set consisted of 37 individuals with the remaining 10 used in the test set, and all data used in this experiment did not contain any occlusions.



(a)



(b)

Figure 3.4: CMC Curves For Changes in Feature Vector; (a) Low Resolution V-47 Data (b) High Resolution V-47 Data. Both sets of results have shown drastic improvements when the feature vector is reduced by PCA

Firstly, it is worth noting that the results obtained on the 'raw' feature vectors (57-68 %) are significantly inferior to those obtained from the trained classifier (95-97 %) Overall comparison of the two blocks suggests that increasing the number of overlapping patches would improve the performance. The improvement of performance is more significant when using the higher resolution data.

In addition, when using the higher resolution data, it is important to choose the correct number of patches. As each patch will contain more pixels, compared to a lower resolution, there is an opportunity for the patch to capture more noisy pixels which can mislead the feature vector. This causes the results to deteriorate slightly when the Normal Blocks are used in the higher resolution data compared to the lower resolution using the same blocks.

3.5.3 Dependency on Viewpoint

Experiments have been conducted to investigate any dependency on the viewpoint of the pedestrian. As shown in Figure 3.1, the V47 were captured at both the front view and rear view of the pedestrians. This experiment would investigate the matching of the a probe set, which contains the view of the pedestrians from one instance in time, to the target set which are views of the same pedestrian at a different instance in time from the same camera. In this experiment Camera A was used. To control the variations in the results, the chosen data are when the pedestrian are within full view of the camera and without occlusions.

Table 3.3: Area under the CMC curve for different poses: FV = Front View, BV = Back View.

Pose Variation (Probe Versus Target)		
FV Versus FV	BV Versus BV	FV Versus BV or BV Versus FV
96.56 ± 2.79	99.00 ± 1.26	96.89 ± 1.63

These experiments use the higher resolution V-47 data, adopting the increased number of overlapping patches (small blocks). The performance is listed in Table 3.3, the highest

accuracy was achieved when the rear view data was used and all experiments with the front view data reached similar results. One explanation for this difference can be attributed to colour pattern on the subject's clothing, from the frontal view, as there is a large variability on the front of the clothing compared to the more consistent patterns on the back of the clothing.

3.5.4 Domain Specificity

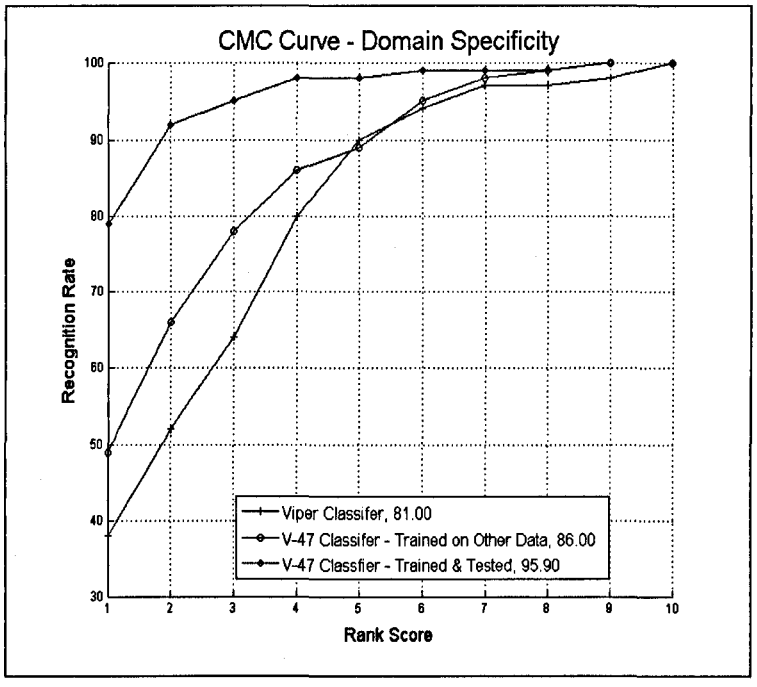


Figure 3.5: CMC Curves of Domain Specificity, the results shown here proves that there is a degree of exclusivity between the data domains, as the best results is achieved when the training and testing data is captured from the same domain.

This experiment evaluates the performance of three classifiers on 10 unseen pedestrians from high resolution V-47 dataset. Since the VIPeR dataset contains pedestrians viewed from different angles, in order to make the results comparable the training and testing VIPeR dataset will contain both front and rear views as well. The differences between the classifiers are:

1. Viper Classifier - the first classifier was trained on 316 pairs from the lower resolution VIPeR dataset.

2. V47 Classifier on other data - this classifier was trained using 37 pedestrians, observed from **different** camera from those used in the test set.

3. V47 Classier same camera - this classifier was trained using 37 pedestrians, observed from the **same** camera views from those used in the test set.

The results in Figure 3.5 reveals that there is a significant dependence on the type of training data: the best results are obtained when the classifier is training data is the captured in the same domain as the test data. This suggests that both the VIPeR and V-47 training sets exhibit some degree of exclusive properties that are not shared. These exclusive properties are also present across different cameras monitoring similar domain.

3.5.5 Use of Additional Training Samples

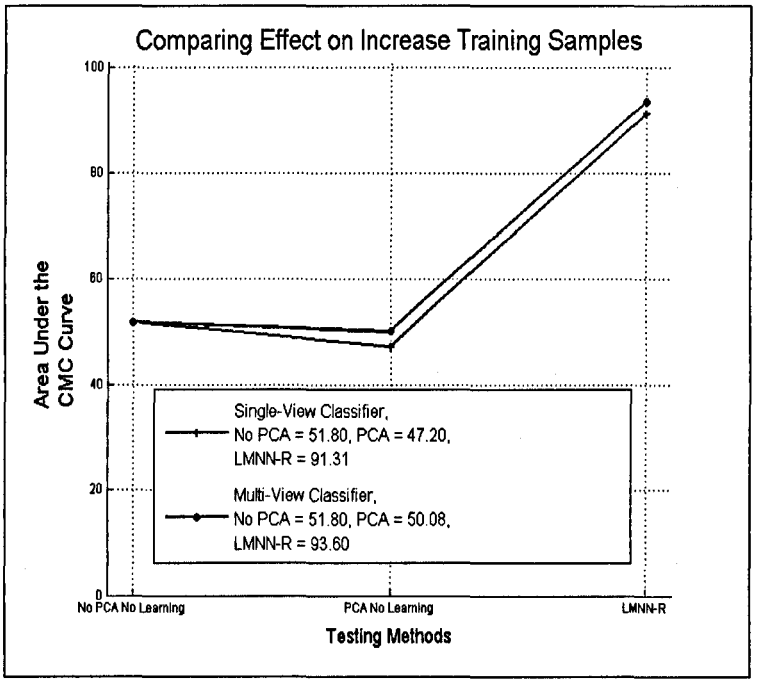


Figure 3.6: Performance of two classifiers, trained using 37 individuals, where multiple instances of the same individuals were supplied; 1) Single View classifier uses 3 instances of the one individual. 2) Multi-view classifier using 9 instances of the one individual

This experiment is designed to investigate whether the cross-camera re-identification performance can be improved by injecting more data of the 'target' set into the training sets. This is achieved through two strategies:

1. Single View - since there are 2 cameras to capture a pedestrian coming and leaving the environment which is equal to 4 viewpoints of the pedestrian. Therefore the training set can be provided with 3 images for each "target" out of the 4 viewpoints.

2. Multiple Views - since a video is available, from each of the 3 viewpoints of the "target", from above. 3 additional images of the same target from each of the 3 viewpoint can be extracted. This is equal to 9 instances of the pedestrian for each 'target' within the training set.

To control the experiment, the viewpoints of pedestrian would vary in the "probe" and only 1 viewpoint of the pedestrian would be included. To demonstrate the performance, the area under the CMC curve is plotted at several stages of the classifier construction: (i) Prior to the PCA and Training, i.e. the raw feature vector; (ii) after the PCA but prior to training; and (iii) After PCA and Training.

The results in Figure 3.6 shows the improvement of cross-camera re-identification compared with results of the V47 classifier trained on other data (86%) as shown in Figure 3.5. However, the results cannot match that achieved by using data from the same camera (96%), as illustrated in Figure 3.5.

In addition, small improvement in the performance can be achieved by using multiple instances from the same viewpoint, which was observed at both prior and after the learning phase. Another observation is the small dip in performance after PCA was applied. This is due to the number of components being limited to 60 so that some of the variance (discriminative feature) with the raw data is removed during the transformation, as outlined in Section 5.2.2.2. However, through the learning phase, the decrease in performance was removed.

3.5.6 Feature Fusion

The experiment in Section 3.5.2 has already shown that fusing with the increased number of patches (sensors) is beneficial to the re-identification challenge. The Figure 3.7 further demonstrates that a concatenated feature based on two colour spaces will perform better than a single feature does, even when is used with fused patches.

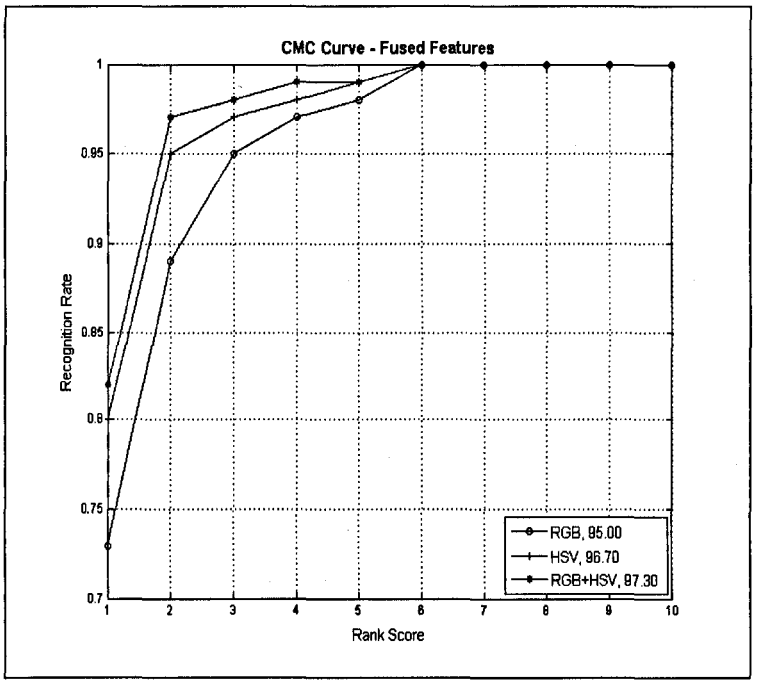


Figure 3.7: Comparison of using single colour space features with fused colour spaces, which shows additional improvement can be achieved when two independent colour spaces are fused, compared to any single feature alone.

3.6 Summary

This work has investigated the use of the Large Margin Nearest Neighbour with Rejection (LMNN-R) classifier to re-identify pedestrians viewed from different cameras at various resolutions and in situations involving partial occlusion. As far as the author knows, this is the best-performing method to evaluate on the VIPeR dataset, and the results can be used as the benchmark for the V-47 dataset.

The experimental results described above support the following conclusions. Firstly, for potentially occluded observations, the best strategy is to attempt the detection of an occlusion and then deploy the appropriate classifier. This achieves better results than training a classifier on a mixed occlusion dataset. Secondly, increasing the number of overlapping patches will improve results; this improvement is more dominant as higher resolution data are used. In addition, the results can be further improved by fusing multiple representations of the subject. Thirdly, the best performance is achieved when re-identifying process uses only the rear-view

of the pedestrian; and when training set is from the same domain as the test set. Finally, the challenge of cross-camera re-identification can be improved by supplying the target set with multiple instances of the probe, which will lower the cross-camera domain specificity effect.

The investigation has shown that effects such as the domain specificity, variation of the viewpoint and the number of patches used, can all influence the accuracy of the re-identification results. This is true even when the datasets are collected in a moderately controlled environment compared with the data captured in the wild. Even though these effects are undesirable, the experiment does show that some improvements can be achieved by adopting feature fusion techniques, which demonstrate the possibility of the high fusion techniques in improving the results.

Chapter 4

A Video Surveillance Scenario

4.1 Introduction

As outlined in *Chapter 1*, the popularity of the video surveillance system, in public and commercial domain, has grown considerably. Although the main applications of these systems are for security objectives, opportunities exist to fully utilise the system, for certain commercial objectives, in parallel. Therefore in this chapter the security and commercial objectives will be discussed and the common denominator to achieving both of these objectives in image processing domain will be described. An examination of the available dataset used to aid our investigation will be further investigated. Finally, the design and implementation of an experiment test bed that incorporates all the identified benefits in the available datasets will be introduced in detail.

4.2 Video Surveillance Objectives

4.2.1 Security Objectives

Traditionally, the objectives of the video surveillance system were to capture a relativity small localised environment. Its applications were limited to act as a deterrent and the captured video was employed as evidence for prosecution. These objectives are currently still the main functions of the distributed surveillance network with the added advantage of capturing a wider environment and increasing the amount of evidence. As discussed in *Chapter 1*, these

distributed surveillance networks are now treated centrally so that those interesting objectives can be detected and analysed whilst they are happening, thus improving the traditional system with in-time function.

However, as highlighted in *Chapter 1*, with the vast amount of information available in time, some events might be missed by the human operator. In order to resolve this issue, an automated tool, combining image processing techniques with certain behaviour rules, are developed to identify certain predefined security objectives in a distributed surveillance network. The techniques of typical predefined security objectives are:

- Matching - The technique involves the identification of objects of interest in the monitored environment and alerts the relevant security personnel of its presence.
- Tracking - The technique involves the continuous tracking of an object of interest through a network of cameras, which could fix the location where evidence could have been located or the crime location.
- Anomaly Detection - The technique for the identification of uncharacterised user behaviour in a controlled environment. This would assist the detection of potential crimes to stop them before happening.

4.2.2 Commercial Objectives

Meeting the requirements of the security objectives is very challenging because the events of interest normally happen very quickly. Comparatively, this will occupy a small amount of footage. In these cases, a large amount of redundant information is created, but it normally contains vast amounts of valuable data that can be utilised to generate statistics of the monitored environment. Commercial enterprises often need statistics about monitored environments. The statistical data includes:

- Usage Statistics - In a commercial environment, the number of users using a facility can be used for various purposes:
  - Statics to evaluate the successfulness of a business.

  - Identify times of high usage to adjust staffing levels
  - Identify the availability of resources

- Tracking - Generating users' travel patterns to help design a more ergonomically commercial facility.
- Area Usage - Tracking users to generate a heat map of facility used to allocate effective product and advertising placement.
- Different Users - Identifying the user groups and the usage patterns of each group to help complete anomaly detection in the security context and effective product placement in a commercial context.

4.2.3 Discussion

The content in this section covers the variety of challenges that a distributed surveillance system can solve in both the security and commercial domains. This,therefore, validates the trend that a surveillance system is no longer limited to meet security demands.

The analysis also demonstrates that an objective, such as tracking, is common in both domain. That is the reason why almost all video content analysis software will share similar Video Content Analysis (VCA) blocks. An example of a general VCA pipeline is illustrated in Figure 4.1.

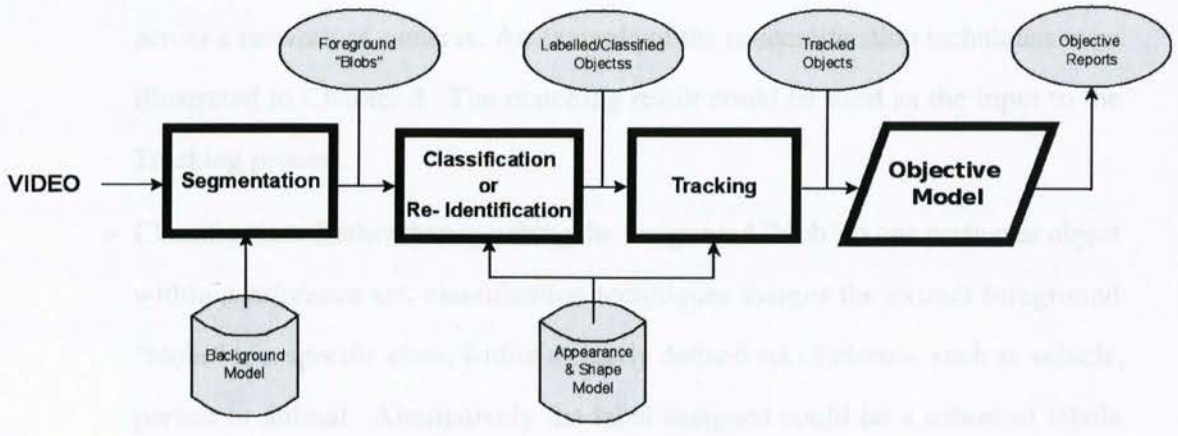


Figure 4.1: Video Analytic Processing Pipeline, an example of a typical processing steps with a VCA, where some process is often supported by a supporting model.

The sequence of the processing steps is shown by the rectangular blocks. They are the common parts within a VCA system. The objective that a particular VCA system attempts to realise is defined by an objective model illustrated by the parallelogram in Figure 4.1. The model would extract information in the scene and combine it with rules to raise an alarm when a pre-defined situation arises or is expected to arise.

Depending on the model or the application requirements, the processing blocks may be supported by data models, as illustrated by the cylinder blocks in Figure 4.1. Generally these models are first developed by some training data and can be subsequently updated, over time, through learning. The operations and the intermediate results produced by each of the processing blocks are:

- Segmentation: The extraction of salient areas or pixels distinctive from the background model. The extracted foreground pixels are further grouped into foreground "blobs", each of which is connected set sharing some features in common. The foregrounds are then passed to the next processing block.
- Re-identification and Classification:
  - Re-Identification: Each foreground "blob" would be matched against a set of reference "blobs". If a match is found the new "blobs" would be given the same object label as that in the reference set. The matching may need to be completed from different viewpoint of a single camera as well as the matching across a network of cameras. An example of the re-identification techniques were illustrated in Chapter 3. The matching result could be used as the input to the Tracking process.
  - Classification: Rather than matching the foreground "blob" to one particular object within a reference set, classification techniques assigns the extract foreground "blobs" to a specific class, within a clearly defined set of classes, such as vehicle, person or animal. Alternatively the label assigned could be a subset of labels within one specific class, such as one of a set of standard human poses. The output is normally the most likely class label. In most cases, the class label is not used

for the tracking process, but can be used to identify if a certain class of object is available within the scene. An example of classification techniques is illustrated in Chapter 6.

- Tracking: By using the labelled objects, tracking algorithms establish correspondence between the same blobs in successive video frames, from either a single camera or from multiple cameras and hence obtain a temporal sequence of the blob.

The pipeline illustrated in Figure 4.1 and the functions described above is generic. The processing steps and the processing order may be normally be defined by analytical application or the objective model.

From the analysis of the objectives and the example of the VAC pipeline, the completion of objectives would all require an object of interest defined by the "blobs". In both the security and commercial context, these objects are either People or Vehicles. Therefore, at the heart of applications, the correct Re-Identification or Classification of the foreground "Blobs" is necessary in order to obtain the expected objective.

The above analysis helps the design of an experimental test-bed and examination of the currently available datasets, including either People or Vehicles, is conducted in the following section.

4.3 Datasets

4.3.1 People

There are several datasets that are targeted at different challenges related to people. For this review, all the datasets listed in Table 4.1 relate to the public datasets that have been used in the pedestrian re-identification challenge.

These datasets provide variations of pose, viewpoint, lighting variation and occlusion. As mentioned in section 3.2, a common issue with these datasets is their modest resolution of the extracted pedestrians with an average 128 of image lines in height. These dataset are captured

Table 4.1: Pedestrian Re-identification Datasets

Name	Media	Resolution	#People	#Images	Occlusion	Location
VIPeR (Gray et al., 2007)	Still	128 by 48	316	632	No	Outdoor
CAVIAR (Hamdoun et al., 2008)	Video	384 by 288	n/a	n/a	Yes	Indoor
ETHZ (Ess et al., 2007)	Video	64 by 32	83	4857	No	Outdoor
iLIDS (UK-Home-Office, 2008)	Video	various	600	600	No	Outdoor
GRID (Loy et al., 2010)	Still	100 x 200	250	500	No	Indoor
PRID (Hirzer et al., 2011)	Video	various	200	400	No	Outdoor
V-47* Chapter 3	Video	576 by 720	47	752	Yes	Indoor

under various conditions such as varied lighting condition, and the challenge with occlusions are only tackled in a limited number of datasets.

It is for these reasons the application of V-47 data, a higher resolution dataset for the pedestrian re-identification problem, were introduced in Chapter 3 and have been bench marked. Although the dataset is captured indoors, it still contains challenges present in the other datasets with extra bonus of natural and artificial occlusion that blocked part of the pedestrian. The occlusion is often useful to solve re-identification problems in crowded environments where only part of the person could be observed.

4.3.2 Vehicles

Unlike pedestrian re-identification, the datasets for vehicles are inadequate. The publicly available datasets are summarised below:

- PETS 2001 (PETS, 2001) and PETS 2000 (PETS, 2000) - The video is of 352x288 pixels and monitors road users in a car park. The dataset is mainly used for vehicle tracking problems.
- i-LID Parked Vehicle (UK-Home-Office, 2008) - It films users of a main road and is used for the detection of the parked vehicles.
- ViSOR (Vezzan, 2010) - This video is of 320 x 256 pixels and captures vehicles using a

car park. The vehicles are used to tackle loitering problems.

- Vehicle Silhouettes (Turing-Institute and Siebert, 1987) - The dataset contains images of 128x128 pixel for different classes of vehicle, which is used for the classification of vehicles.

- VIRAT (Oh et al., 2011) - The focus of the dataset is to detect people's behaviours within the monitored environment, but part of the dataset can be used for vehicles' car park usage.

Examination of the datasets shows that the vision challenges related to vehicles, focused more on tracking; less attention is paid to re-identify the same vehicle in different spatio-temporal environments. This challenge is assumed to have been solved by the correct recognition of the pattern on the number plate. However as Section 5.3.2 there are still various challenges around the correct recognition of the number plates.

Another limitation of these datasets is that they only capture a limited spatio-temporal environment, therefore they can only be used to infer limited amount of queries.

4.3.3 Discussion

From the review of the datasets used to tackle surveillance challenges, the below lists restrictive factors:

- Vehicles as the object of interest are partly captured and are often a by-product of datasets that focused on people.

- The resolution of the image is low, it effectively excludes the use of high spatial frequency features.

- Datasets are normally produced to tackle particular issues, therefore it is hard to use them for tackling other problems.

- Datasets are captured by snapshots of the monitored environment, therefore it cannot show how an object is using the environment as only particular instances of the object are included.

The above analysis shows that using the current available datasets to tackle both security and commercial challenges are difficult. Therefore, it is necessary to design an experimental dataset that can mitigate the restrictions listed above.

Furthermore, an effective dataset also needs to encapsulate many of the challenging variables present in the current datasets, such as variation in lighting, environmental variables, and occlusions. Therefore the dataset can be used by a wider community to deal with different video surveillance challenges.

4.4 Experimental Test-Bed

4.4.1 Design

With the objective outlined in Section 4.3.3, the surveillance scenario illustrated by Figure 4.2 is proposed.

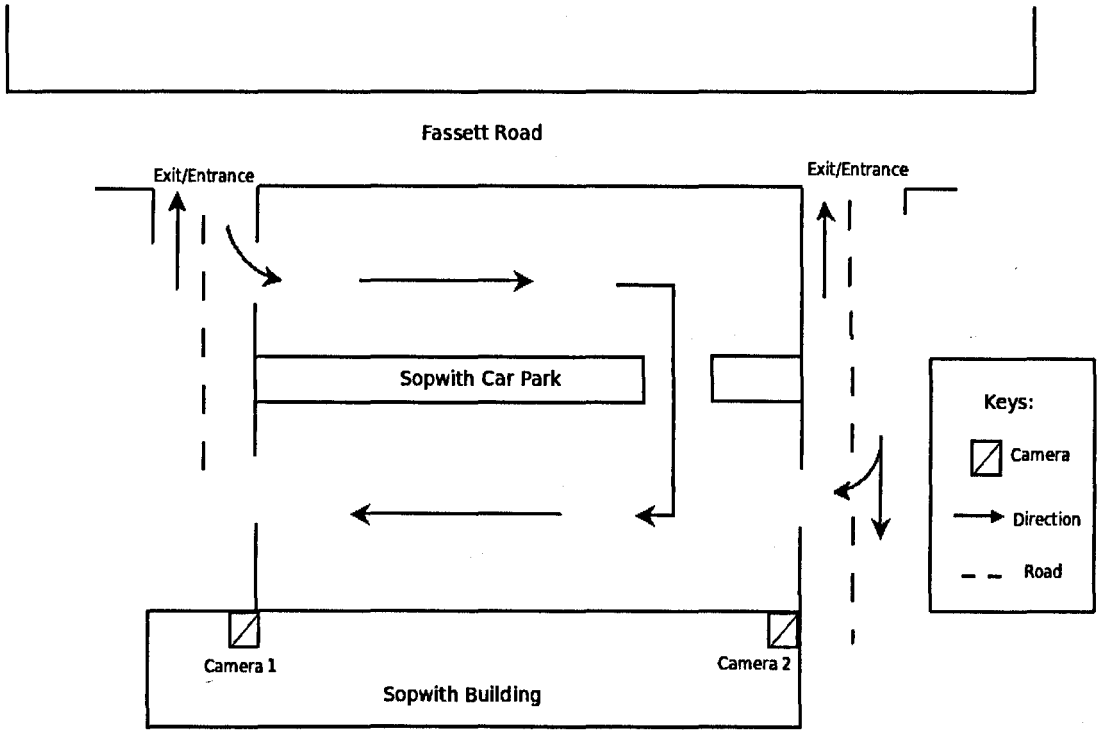


Figure 4.2: Layout of Sopwith Car Park with Camera Position

This proposed scenario is ideal for this investigation as the car park is a closed looped system whereby a vehicle's entry and exit activities can be fully monitored. To utilise the closed

looped system, the object of interest in this scenario would be vehicles. In some scenes, however, vehicles are mixed with peopleand alough the dataset also captures people using the carpark, pedestrians who use other exits can't be fully monitored.

The main function of many car park monitoring systems is to determine if a particular vehicle is allowed to use the monitored car park. This is normally achieved by using a controlled barrier. It not only needs special hardware and software, but it can also cause traffic problems. As the car park in the project is not barrier controlled, the proposed scenario allows the investigation the possibility of solving the problem whilst eliminating the shortcomings of the barrier system. This capturing plan can also be used to query a range of different objectives:

- Security
    1. Is this vehicle breaching car park usage polices?
    2. Is the identified vehicle loitering?
    3. Is the identified vehicle using the car park?
    4. Is the usage pattern of a particular vehicle identifiable?
    5. etc...
- Commercial
    1. Identifying whether the vehicle is using car park for delivery or parking.
    2. Identifying when the car park is full.
    3. Identifying the time when the car park is in high demand.
    4. Identifying the popularities of certain exit/entry point.
    5. etc...

4.4.2 Equipment

As stated in Section 4.3, the object of interest in the available dataset has a relatively low resolution. The low resolution is due to the fact that popular surveillance cameras need to capture a large field-of-view, the information from video needs to be transmitted in real-time,

to the monitoring stations and that there is a limited data storage facility. Although the captured images are suitable for humans to process, they are less than ideal for a computer vision system to process. As a lot of the edge and colour information is lost, almost all computer vision systems rely heavily on these key pieces of information, in order to distinguish objects in the image.

With advancements in the transmission and storage capabilities, higher resolution cameras are now being used in some surveillance scenarios. An increase in the resolution and improvement in quality of images are key characteristics, brought by the high definition cameras which are ideally suited for computer vision systems. However, these benefits are not being utilised by the research community, mainly due to the relatively high cost. This project aims to utilise these advantages, for this reason the dataset will be captured by using a full high definition 1080 progressive scan colour video cameras during the car park's operational times over a four month period. The data captured through this test bed is going to be used for the investigation in Chapter 6.

The dataset is captured at 25 frames-per-second and the data is first stored in camera's storage. To reduce the storage requirement, the raw footage is compressed via H.264 format to limit the loss of the quality. During the non-operational times, the data are transferred to other storage facilities for off-line processing in order to free up the camera's storage. To consolidate the two cameras, they are time synced to a control computer and the meta-data associated with the footage are used to meet any time-plexing requirements.

4.4.3 Redundant Data Reduction

The captured footage includes a significant quantity of redundant information because the car park is not being continuously used during its operational time. The removal of the redundant information is inevitable, in order to extract the most useful events regarding vehicles entering and exiting the car park. Another benefit of removing the redundant information is the saving of storage, considering the fact that 5 minutes of raw footage needs 300 mega-bytes of storage.

The automated event extraction algorithm needs the capability to overcome the following issues in the raw data:

1. Lighting Variation - Because the footage is captured from before sunrise to after sunset, the algorithm needs to adapt the gradual changes with the lighting condition as well as sudden changes, such as those caused by fast moving clouds. The algorithm is also required to perform well for a broad spectrum of lighting intensity levels.

2. Miscellaneous Movements - The algorithm should be robust so that it does not segment events caused by other moving objects within the field of view, such as swaying trees which is a specific artefact within the monitored data used.

3. Object Entering and Exiting - The algorithm needs to determine when a new object has entered the scene, as well as when the object is leaving the scene. In addition, it should not be affected by the permanent shadows in the scene.

4. Slow Moving Objects - Because there are restrictions in the road, vehicles are often moving slowly. The algorithm should be able to identify the difference between a vehicle moving slowly and a vehicle entirely stopped in the scene.

5. Adaptive Background - the car park entrance can sometime be used as a loading bay. Therefore the algorithm needs to adapt quickly to classify the parked vehicle as temporary background.

The above list is all of the challenges that need to be overcome by a single algorithm. The techniques that are chosen need to perform well under all of these scenarios, rather than the best performing one, under a single challenge.

4.4.3.1 Background Subtraction

Taking the above analysis into consideration, the most suitable approach should be one that can determine changes between frames within a video sequence. The most widely used approach is the Background Subtraction (BS).

Background subtraction is a very active research field where various methods have been proposed. They all form a common characteristic where a background reference image is

constructed over a number of training frames. Each pixel in the new image is subtracted from the corresponding reference pixel and a threshold is applied to determine if the pixel belongs to a foreground or a background.

The difference between the methods is the techniques used to construct and update the reference background image. The most commonly used approaches are outlined below:

- Temporal Filters - In this approach, each pixel value in the reference image is determined based on either the maximum, median and minimum values of the corresponding pixels within the training frames. Lo and Velastin (2001), suggested using the median value of the last few frames to construct and subsequently update the reference image.

- Gaussian - This approach is based on fitting a Gaussian probability density function to each individual pixel of the reference image by using the corresponding pixel values in the training images. For each pixel, there will be a mean and a standard deviation. Because the rate that the background image is updated is often controlled by a weight, it allows variation in the lighting conditions. Wren et al. (1997) proposed a single Gaussian. Other researchers developed a single adaptive Gaussian to overcome gradual changes in variation in lighting condition. Furthermore, Stauffer and Grimson (1999) reported a more sophisticated approach known as a Gaussian Mixture Model (GMM). The GMM approach has the ability to model the multiple background objects, thus made it possible to detect the background objects that are not permanent and appear at a rate faster than that of the background update.

- Kernel Density Estimation - This approach aims to eliminate the drawbacks caused by the limited number of training data that is used to approximate the histograms used to create the Gaussian probability density function (PDF). Elgammal et al. (2000) modelled the background distribution by using a non-parametric model based on Kernel Density Estimation on the buffer of the last n background values.

- Eigenvalue - The method is not so popular due to the computational cost, thus making it inefficient in a real-time application. The method utilises the eigenvalue decompositions. Seki et al. (2003) applied the eigenvectors of blocks of the pixels, but Oliver et al. (2000)

introduced the decomposition of the whole image to avoid the tiling effect caused by the block partitioning in Seki's methods.

The above approaches all have their advantages and disadvantages, as outlined in the review paper of Piccardi (2004). Ideally the detection of the foreground object should be accomplished in real-time with a visual surveillance system. Therefore when considering the choice of BS method, a balance between accuracy and speed needs to be made. The most suitable method is to use the GMM method that was supported by Piccardi's comparison. Furthermore, GMM has been shown to be the most versatile and robust across a range of different video sequences, as illustrated in the experimental comparison conducted by Benezeth et al. (2008).

This section only covers a selection of available techniques. Some of the new techniques with improved performance have also been suggested by Barnich and Van Droogenbroeck (2011). However, the robustness of these new methods on different video sequences has not been shown. In addition in a recent experimental comparison conducted by Brutzer et al. (2011), the GMM has shown to outperform some of the new state-of-the-art techniques. This further justifies the choice of adopting the GMM as the BS approach within this framework. Furthermore, the approach described in Stauffer and Grimson (1999) paper makes GMM method ideal to solve all requirements listed in 4.4.4.

GMM Implementation

This section will outline the mathematical and assumption made for the GMM, based on the Stauffer and Grimson's Stauffer and Grimson (1999) paper. The approach models every pixel with a mixture of Gaussian and describes the probability of observing a certain pixel value, x, at time t

$$P(x_t) = \sum_{i=1}^{K} \omega_{i,t}\, \eta\left(x_t, \mu_{i,t}, \Sigma_{i,j}\right) \tag{4.1}$$

where K is the number of Gaussian distributions and describes an observable objects, it can be either a foreground or background. $\eta\left(x_t, \mu_{i,t}, \Sigma_{i,j}\right)$ is the i^{th} Gaussian probability density and is represented by the mean $\mu_{i,t}$, and covariance matrix $\Sigma_{i,j}$. As the Gaussian's multi-variant removes the costly matrix inversion and the three colour channels are assumed

to be independent and have the same variance, therefore, simplify the covariance matrix to be a diagonal with the form $\Sigma_{i,j} = \sigma_i^2 I$.

Matching is conducted with each new pixel value x_t, against each of K Gaussian distributions. It is defined as:

$$\frac{(x_t - \mu_{i,t})}{\sigma_{i,t}} < T \tag{4.2}$$

where T is the threshold currently set to be 2.5 standard deviations. The parameters of the matched components are updated as follows:

$$\omega_{i,t} = (1 - \alpha)\,\omega_{i,t-1} + \alpha \tag{4.3}$$

$$\mu_t = (1 - \rho)\,\mu_{t-1} + \rho.x_t \tag{4.4}$$

$$\sigma_t^2 = (1 - \rho)\,\sigma_{t-1}^2 + \rho.\,(x_t - \mu_t)^T\,(x_t - \mu_t) \tag{4.5}$$

where α is a predefined learning rate and ρ is learning rate defined by α and the closest Gaussian distribution.

For unmatched distributions the σ and μ do not get updated and the weight is updated by:

$$\omega_{i,t} = (1 - \alpha)\,\omega_{i,t-1} \tag{4.6}$$

Allowing decay to occur, the least probable distribution (the one with the largest standard deviation) is replaced with a distribution of the current value as the mean, a large initial variance and a small weight.

To determine the portion of the mixture model that best represents background process, the K distributions are ordered based on a fitness value defined by $\frac{\omega_{i,t}}{\sigma_{i,t}}$ and only the most reliable B

is chosen as the background model:

$$B = argmin_b \left(\sum_{b}^{k=1} \omega_k > \tau \right) \tag{4.7}$$

where τ is a user defined threshold.

Connected Components

The GMM method outputs foreground pixels for each new frame. The labelled foreground pixels can be segmented into regions using a connect component algorithm. As it is considered to be an object of interest, the size of the labelled components needs to be greater or equal than a threshold τ_c that is set to be 3000 pixels.

Once it has been determined to be an object of interest, the total pixel value, centre of mass, and bounding box size coordinates are also stored. The main function of the stored information is to extract the useful information and remove the redundant information. The information could also be used for later processing, such as tracking.

4.4.4 Discussion

In the data capturing design introduced above, the data meeting the requirements explained in Section 4.3.3 are used as the foundations of this project. One of the key steps is the visual surveillance challenges in the extraction of the object of interest from the background images. The popular GMM background subtraction techniques is extended to remove some of the redundant information with the captured data and to create the extraction database.

An extension of the background subtraction is to combine it with the Kalman filter to create a tracking algorithm. This extension algorithm has been used to create the content of two surveillance TV programs using the data captured from the experiment test-bed.

Although the GMM is a very popular technique, there are some limitations when applied to this dataset, these are:

1. Low Lighting Intensity - As natural light is the dominant source for the car park so when the intensity is low, the algorithm is unable to identify the vehicles.

2. Shadows - This causes the increase of the number of false-positives, mainly by the shadows as pedestrians walk by and through the car park. The shadows from vehicles are less of an issue at the BS stage as they are already the dominant feature.

3. Group Pedestrians - When a large group of people walk in close proximity, the algorithms is unable to separate them out, which will cause a false-positive vehicle identification.

These limitation are the results of using a single algorithm to meet all the challenges listed in Section . Due to this requirement a balanced threshold, as shown in Equation 4.2 was chosen for the GMM process. However, even with the limitations, the algorithm has successfully met our original requirements of removing the redundant information,as a manual audit all of the events of interests were captured and the false-positive results were also removed, albeit manually.

4.5 Summary

This chapter states that the video surveillance network is a valuable tool to achieve a range of security purposes, as well as a range of commercial challenges. An examination of the available datasets has shown that the datasets are unsuitable for solving challenges in both domains simultaneously. It also inherits a range of limitations which decreases the effectiveness of the imaging techniques used to solve the objectives.

This leads to the design of an experimental test-bed to fully utilise the scenario. The objects of interest chosen for investigation are "Vehicles". Although a lower interest of using vehicles as the object within the research community exists, due to the lack of available datasets, the analysis in Section 4.4.1 claims that vehicles are still an important subject in a range of security and commercial objectives.

The review of the objectives has also shown tha,t in a video analytic system, the correct re-identification and/or classification of the subject is necessary. As Chapter 3 has already demonstrated the re-identification challenges have various obstacles which effect the results a further examination of the challenges in the classification scenario is explored in 6. As the completion of these tasks would require some types of features related to the object of interest, Chapter 5 will examine the features that can be extracted from vehicles and the associated uncertainties related to each of the features and within the various stages of the video surveillance system.

This chapter also demonstrates the application of segmentation techniques to extract the object of interest from the background, which is believed to be the first step in any VAC application. Previous discussion also reveals that there are a number of limitations that are associated with the techniques and influence the accuracy. This proves that uncertainty exists in every processing step and may propagate down the pipeline. Furthermore extending the segmentation techniques with Kalman filtering, makes a tracking algorithm which was used to produce two demonstration videos that have been viewed to a broad audience and forms part of the publications that accompanies this thesis.

Chapter 5

Uncertainty Within Video Surveillance Systems

5.1 Introduction

Based on the design of the experiment test-bed in Chapter 4, the aim of this chapter is to first outline the uncertainties within a complete video analytical system, from converting the analogue world to its digital representation to meeting the video analysis objective.

Chapter 4 also concluded that the successful completion of re-identification and classification will depend on the extractable features related with the object of interest. This chapter will also examine features related with vehicles and their associated uncertainties when being processed. Although the examination concentrates on vehicles, some of the features may become useful for the researches related with people. The review taken place in this chapter will take a generic view on the approaches, where reference to a specific approach a link to the relevant chapters will be given

5.2 Uncertainty Within Video Surveillance Systems

This section will first outline uncertainty caused by the hardware that is used by a video analytic system to convert the real world to its digital representation. This is followed by an exploration into the uncertainty created by the generic activities within the software used to

process the captured footage to achieve the VCA objective.

5.2.1 Data Uncertainty

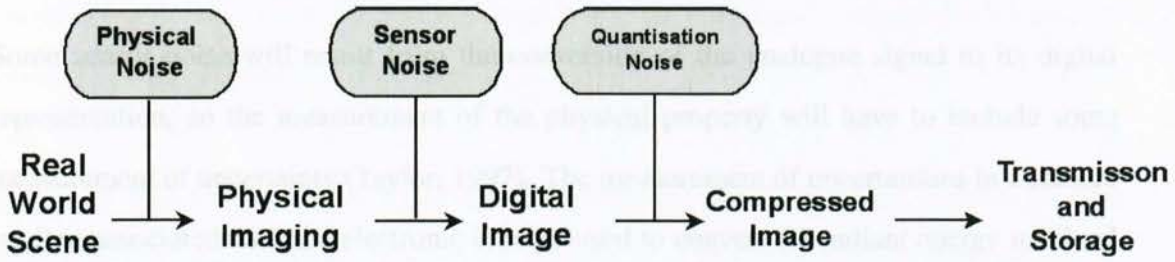


Figure 5.1: Image Processing Pipeline, converting the analogue to the digital representation creates various artefacts which can contribute to the uncertainty, even before it is processed by a VCA.

Figure 5.1 shows a generic processing pipeline to transform real world scenes into video data that can be processed by a video content analysis system. The figure also shows the type of noise that would increase the uncertainty of the video data.

5.2.1.1 Physical Noise

Physical noise is caused by the defects of the hardware involved with a video camera, such as the lens. The camera is similar to the human eye. In the eye there are hundreds of millions of light-sensitive cells which is also equipped with the ability to perceive intensity (brightness) in a remarkable range of nine orders of magnitude (Sonka et al., 1999). In contrast, a current state-of-the-art surveillance video camera only has a few million photosensitive sensors and a range of intensity sensitivity to about 4 orders of magnitude. The amount of information available from a VCA system has already decreased greatly compared with the human eye. This is a key contributor to some of the defeat in segmentation systems, highlighted in Section 4.4.4, especially in a low light condition.

Other sources degrading of the image in a video camera were addressed by Morris (2004). These sources include Geometric distortion, refractive index of the lens and the uneven sensitivity of the image sensors. As Morris mentioned, although these degrading effects are

difficult to measure, most camera manufacturers will combine these effects into the Signal to Noise Ratio (SNR).

5.2.1.2 Sensor Noise

Some sensor noise will result from the conversion of the analogue signal to its digital representation, so the measurement of the physical property will have to include some measurement of uncertainty (Taylor, 1997). The measurement of uncertainties in a camera is often associated with the electronic sensors used to convert the radiant energy involved with an electrical signal of the photosensitive cells. The uncertainties are also quoted by the camera manufacturers as the device's SNR. Two main types of noise are:

1. Salt and Pepper- This randomly introduces pure white or black pixels into the image. Although it is minimised when high resolution images are used, it still contributes to the overall information imperfection.

2. White Noise and Gaussian Noise - This arises due to randomness superimposed on the signal as being captured or processed.

5.2.1.3 Quantisation Noise

Quantisation noise is introduced by the need to compress the video data in order for effective transmission and storage, which is a similar process as the neurons linking the eye to the brain.

The compression of the information would often cause further degradation of the image. According to Richardson (2004), video compression is realised by removing the subjective redundancies which are elements of the video sequence, removed without significantly affecting the viewer's perception of visual quality. However, as Chen et al. (2008) mentioned, the measurement of perception depends on the "viewer". When the "viewer" is an image processing algorithm, the measure is the sharpness of the decompressed image. Chen et al. revealed that decrease in the sharpness is unavoidable, even when a lower compression rate is used. The loss of sharpness will affect the edge information, as mentioned in Section 4.4.2, therefore degrading the quality of information available to any VCA system causing a reduction in the overall accuracy of the result.

5.2.2 Software Uncertainty

Here, the software refers to a video processing system such as the pipeline illustrated in Figure 4.1. Due to the existence of these factors within almost every system, a very specialised field known as Soft Computing exits to develop low cost methods to tolerate the uncertainty created by these factors.

Generally there are three main types of factors that may cause uncertainty in video analytic software. They are known as Data Model, Training Data and Mathematical assumptions.

5.2.2.1 Data Model

As discussed in Section 4.2.3, the majority of the processing blocks in a VCA system require the support of some data model. For instance, in a segmentation process, the detection of the foreground blobs is supported by background model. A simple model could be a threshold value for the difference between frames to define the foreground pixels. A complicated model could be the creation and updating of the reference background image, which is subsequently subtracted from the subsequent image to acquire the foreground pixels, as addressed in Section 4.4.4.

Both of these models will cause a change in the appearance of the blob. If it was incorrectly created or updated, the change would contribute to uncertainties, which is propagated down the processing pipeline. In a simple model, if the threshold is incorrect, the "blobs" will appear either larger or smaller than expected. In a complex model, the frequency at which the background model is updated is important to eliminate the effects of changes in illuminations, and may influence the detection of the foreground pixels. As the illumination would normally create shadows, it changes the appearance of the "blobs". Although various methods tried to sort out the issues, as reviewed by Sanin et al. (2012), the entire removal of shadow seems to be impossible.

5.2.2.2 Training Data

Another effect of a data model for processing blocks is the training data. Typically, a model is created by maximising the performance with respect to the training data. In these situations, the over-fitting of the data would occur. As Tetko et al. (1995) stated when over-fitting occurs, the model will learn the error of the training data rather than generalising a model to represent the environment it expects to evaluate. This increases the uncertainty in the system outcome when conducting analysis for previously unseen data.

Prince (2012) expressed that the benefit of using data with reduced dimensions is that the model would require fewer parameters therefore it would be faster to learn and to use for inference. However, Bevington and Robinson (1969) stated that due to the reduction of dimension some of the information would be lost and the lost information is difficult to measure. This is sometimes associated with the errors in the data model.

5.2.2.3 Mathematical Assumptions

The development of a video analysis system is driven by a computer of high computational power with a relatively low cost. Such a computer would support the computation of a complex mathematical program for processing blocks of models and solving difficult problems in a real life. The use of mathematical models to define real life scenarios can be extremely difficult as described by Kennedy and O'Hagan (2001). To simplify the mathematics calculation and complete the process in a reasonable time frame, designers sometimes make some assumptions or modify the parameters to achieve the best results. As a consequence, the accuracy of the results would be affected when data violating these pre-defined parameters is used.

5.3 Vehicle Features Uncertainties

This section will also examine features related to vehicles and their associated uncertainties when being processed. The examination will mainly cover the feature usage when identifying vehicles, however, some of the features may be more predominantly used in researching challenges containing people.

5.3.1 Colour

Colour is one of the most important factors in any computer vision challenge because it contains a lot of edge information that is useful in segmenting the foreground and background pixels for recognising the object of interest.

Once the object of interest has been extracted, colour is also valuable to describe both people, as shown in 3, and vehicles. The simplest form of colour description is a grey scale image, which is a measure of intensity with a range from total black to white. As mentioned in Section 5.2.1.1, the human eye senses intensity in nine orders of magnitude. In order to digitally encapsulate it, each pixel would require 4 bytes and a typical high definition image would require 8 megabytes of storage, which is impractical within a surveillance environment for both transmission and storage. To improve the storage requirement, a compromise is made and the typical intensity level is reduced to 1 byte (0-255). The compromise means that a large proportion of intensity information, normally required in low light situations, may be lost. Meanwhile, it may increase the level of uncertainty in visual challenges under low light situation, as described in Chapter 4.

Grey scale image only uses one colour channel to get better description. The multiple channel descriptors are used to create a colour space. The most popular colour space commonly used is the Red, Green and Blue (RGB) colour space, as the same colour space is used by the human eye. With 1 byte for each colour channel, there are 16777216 possible combinations of colours, which give far better description ranges than grey scale alone without demands of large storage.

To create the description of an object, an image is converted to create histograms representing the number of pixels that have the same intensity level in each colour space. This will create a sparse histogram where every possible colour is represented. This will become uninformative, as most of the colours will never occur and those that do occur will mainly occur once or twice. This sparse histogram will be difficult to use in a reasonable way and would require large processing power. A potential solution is to divide the histograms into smaller discrete bins.

Each represents a range of possible values. There are some uncertainties, possibly created by the bins, depending on the size of the bin. For instance, if the range is too big, the key colour characteristics are merged together. If the range is too small, the problem of sparse histograms would resurface. Another technique that could be employed here is the use of a knowledge-based histogram binning process, similar to that used by Gray et al. (2007). Here, the object of interest is broken into known regions based on the knowledge that certain areas will have more variation than other, different sizes of bins can be chosen. Although this approach can create a better descriptor, therefore lowering the uncertainty, it still requires the correct size of bins to be chosen otherwise the histogram issues above would still exist. Although the uncertainty created is not ideal, it is outweighed by the advantages created by the colour histograms.

Another contributor of uncertainties when using RGB colour space is the variation of colour observations with respect to changes in various environmental conditions. One possible solution is to use the Hue-Saturation-Value (HSV) space, as HSV is more tolerant (Sumalee et al., 2012) to such change. All of the colour spaces, including the other popular colour space, have their own merits and could be transformed to the RGB space with ease. Although there are different tolerance levels to the variations, the continuous nature of the colour spaces makes the comparison of two objects very difficult, as it is unlikely that the same object would have the exact same measurement due to the uncertainty of the measurement results. To reduce the uncertainty produced by the continuous characteristics, some researchers quantised the colour spaces to produce discrete subsets of colour known as colour codebooks. Similar to the histogram, the uncertainty in the codebook arises from the range of colours used.

Within the surveillance context, colour is extensively used for the re-identification of people, as demonstrated in Chapter 3. Chapter 3 has also shown the accuracy level is effected by various aspect of the data that a single colour feature cannot resolve.

Compared to people, the re-identification of vehicles based on colour is less extensively researched. Possible reasons are:

1. Colour alone could not re-identify the vehicle as many vehicles share the same colour.

Hence it is not a distinctive characteristic.

2. An appropriate colour model is not straightforward to determine because sometimes a single value is useful, in other situations, colour can be described by several colour channels.

3. The overall colour of a vehicle is difficult to summarise as windows and wheels have large contrasting colour schemes compared to the body. Psyllos et al. (2011a) suggested a solution. They collected a RGB histogram for a range of patches on the vehicle and chose the peak of each of the colour channels as the components to represent the overall colour.

Considering the challenges related to colour and using colour for re-identification vehicles, the measurements from two "vehicle colour sensors" could be used to produce an estimate of the probability whether these two observations refer to the same vehicle, or the two observations refer to two vehicles with the same colour model. In this scenario where colour may have a high variance, techniques such as Fuzzy logic can be used to create the similarity probability measurement required for the input into the high level fusion system.

5.3.2 Automated Number Plate Recognition (ANPR)

Although a vehicle's number plate (NP) is its most discriminating feature, there are a number of issues with the correct recognition of these number plates. One of the major issues is the non-uniform standard across the world. Even in the United Kingdom, there are different alternatives depending on where vehicle is registered, as stated by Rhead et al. (2012). Although ANPR has been widely used for law enforcement, a range of general issues still require better solutions, such as:

- Poor Resolution - These issues can be due to the low resolution cameras, when the plate location is far away from the camera.

- Blurring - Caused by the motion of vehicle being faster than the frame rate of the camera, thus creating motion blur.

- Lighting - Number plates, within the UK, have a retro-reflectivity coating that was introduced to improve the visibility of unlit vehicles parked on roads. However it can

affect the captured image if the light intensity is high therefore the reflected cause of the images to be blurred.

Due to these problems, many of ANPR systems require specialised equipment and may work better, only when the vehicles in question are travelling at a relatively lower speed. However, even when this equipment is used, there are specific struggles associated with the ANPR procedures. An ANPR procedure involves two main problems:

1. NP Localisation - The ANPR procedure relies heavily on correctly locating the bounding box that contains the possible NP. Generally, it is quite accurate. The uncertainty potentially introduced is that multiple areas can be classified as NP, such as stickers containing letters. These uncertainties can be reduced by enforcing government regulations to eliminate false-positive areas.

2. NP Recognition - It relies on the Optical Character Recognition (OCR) techniques to recognise the number on the plate.

Compared to using OCR techniques for handwriting, the level of uncertainty in recognition of NP has been assumed to be very small because the NP needs to conform to government standards. However, the uncertainty is increased due to the decreasing clarity as the issues mentioned above, as well as the introduction of foreign bodies such as screws and mud on the NP, as addressed by Rhead et al. (2012).

The above highlights many factors that may affect the accuracy of recognising a NP. These challenges have been categorised into plate variation and environmental variations by Du et al. (2013). The authors outlined that the accuracy rates of the current state-of-the art techniques are typically between 90 - 97%, depending on plate formation. The majority of the techniques described are designed for vehicles that are almost stationary, and the rate of accuracy decreases when moving-image (video) data is used.

Some uncertainties are also found when comparing the outcomes of the ANPR that is displayed by strings. A popular metric of string comparison is the Hamming Distance metric (Hamming, 1950), which outputs the number's position at which the corresponding string are

different. A major shortfall of this metric is that the strings need to be of the same size because a real NP string may have variable length. Even in the UK, the NP length varies between 2-7 letters long. To resolve this shortfall, fuzzy string searching methods adopted for DNA comparisons can be employed. One popular metric is the Damerau - Levenshtein distance (Damerau, 1964; Levenshtein, 1966). It measures the number of editing operations needed to make the strings identical by taking 4 different editing operations; insertion, substitution, deletion and transposition. Since all editing operations, by default, are assigned the same weight, some comparisons may result in identical scores. For example to compare the score of ABC, the following strings will score the same AB (one insertion), ABCD (one deletion). Thus a degree of uncertainty still exists in the measured metric.

Other challenges resulting in the uncertainty, when using the NP alone to fix a particular vehicle, are the partial blockage of the NP and circumvention techniques to change or remove the NP. In these circumstances the challenge of using one object feature to re-identify the object is difficult. This further illustrates the need to fuse multiple features to reduce the uncertainty in re-identification results.

5.3.3 Vehicle Manufacturer's Logo

A vehicle manufacturer' logo is another distinctive feature related to the vehicle that can assist with the vehicle re-identification process. Unlike the NP, a logo can't be modified as easily, it is therefore ideal as a secondary feature.

The importance of correctly identifying a manufacturer logo is not restricted to the security context. Detection of the logo is also important in identifying counterfeit products, as illustrated by Lei et al. (2012). In addition, it can also be used to track on-screen time of sponsor logos of a sporting event. For a visual system, the contour of the logo is very importantin its idendification; ideally it should be unique so that it can be easily classified. as shown in Figure 5.2.

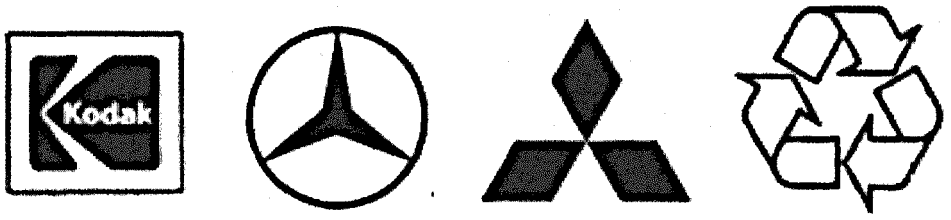


Figure 5.2: Example of Typical Brands, showing the contours of the logos which are typically used to identify the brand.

The distinctive contour means that the shapes can be easily categorised by the edge information alone. However the logos used by some vehicle manufacturers have very similar contours, as illustrated in Figure 5.3.

Figure 5.3: Similar Vehicle Brands, where contours that very similar but can be distinguished by the detailed patterns within the logo

To distinguish the logos illustrated in Figure 5.3, the information contained in the contour should be compared. If images shown in Figure 5.3 are used, the level of uncertainty is lower as the edge information inside the contour can be distinguished easily. However, the quality of the logo image captured by the surveillance system would significantly reduce the capability of edge detection system for identifying all the information inside the contour. Major causes for the reduction are:

- Size - Compared with the vehicle, a logo only occupies a very small percentage of the vehicle's area. The size will also vary depending distance between the camera and the vehicle.

- Orientation - Depending on the view point a logo's edges would appear to be merged, so as to loss some of the distinguishable edge information.

The challenges above are not restricted to classify similar contoured logos. They also contribute to the uncertainty when classifying unique contoured logos. Another challenge that results in the uncertainty is the number of variants of a vehicle's logo that is currently in circulation as demonstrated in Figure 5.4.

Figure 5.4: The various incarnations of the Vauxhall logos which are still actively being used on UK roads

Future uncertainty is also introduced by the classification approach. Two popular techniques are suggested below:

1. One-against-all classifier - Each manufacturer's logo needs its own classifier, n manufacturers require n classifiers. Therefore, for a given logo there will be n estimates. For similar logos, as illustrated in Figure 5.3, the difference between the estimates might be very small, therefore, certainty of the results is reduced. The current best result is Psyllos et al. (2010) and works with averaging 91% overall classification success for 10 categories. However it has some bias in the data used, such as capturing the data close up.

2. Multi-class Classification - There is a limited exploration of using this technique for vehicle classification, partly due to the subtle difference between some of manufacture logos and it is made difficult by the low resolution image. However benefits exist in the use of a single multi-class classifier rather than multiple classifiers, such as the reduction in the number of classifiers that should be trained. A state-of-the-art classification technique is employed for manufacturer's logo classification, in Chapter 6.

Apart from the challenges discussed above, the correct classification of the vehicle's logo is heavily dependent on the correct location of key reference points on the vehicle. In the

majority of techniques, the key reference is the location of the NP, and alternatively, the location of the headlight or tail lights are also used. Therefore, the amount of uncertainty introduced by the decency on the correct location of these key features should be controlled.

5.3.4 Vehicle Body Type

Classifying the vehicle's body shape into distinctive categories gives a characteristic that is unable to be modified. Examples of the silhouette of popular vehicles are represented in Figure 5.5.

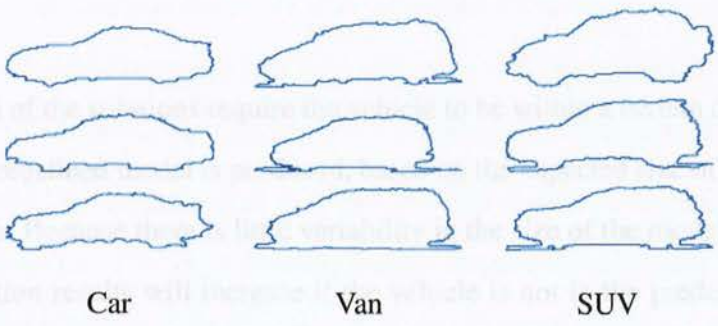


Figure 5.5: Vehicle Silhouette Example, extracted from the side view of the vehicle showing distinctive characteristics of the car, van and SUV categories

The key differentiators between the categories are the shapes made by the silhouette. The extraction of the silhouette requires the extraction of the foreground pixels from the background. However the extraction imperfections, such as shadows, as identified in Chapter 4, is a major source to the uncertainty of the classification result, as shown in the Van and SUV contours in Figure 5.5.

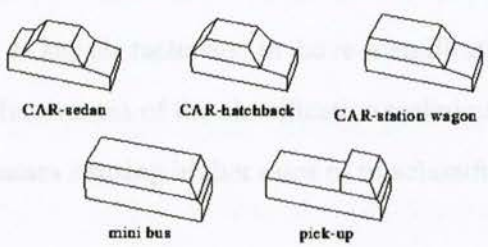


Figure 5.6: Vehicle 3D silhouette models used for the classification of vehicle types, by projecting it into a 2 dimensional image, as suggested by Koller et al. (1992).

The silhouettes displayed in Figure 5.5 are only present when viewed from the side, however, the position of the majority of surveillance cameras is of a top-down view of the vehicle. In a 2 dimensional scene, the vehicle will appear to be flat. A possible solution is to project a 3 dimensional silhouette, as those illustrated in Figure 5.6. Koller et al. (1992) suggested a predefined model of the image to see if the vehicle fits or not. Alternatively, Buch et al. reported the projection of the 2D view into a 3D representative before fitting the model. Both of these solutions offered a degree of freedom in the viewing angle, but as Buch stated, due to the line detection techniques used to fit the model, there are still limitations to the viewing angle.

In addition, both of the solutions require the vehicle to be within a certain distance from the camera, as the predefined model is produced, based on the expected size of the vehicle in an area of the scene. Because there is little variability in the size of the model, the uncertainty in the classification results will increase if the vehicle is not in the predefined area in the scene. Kanwal et al. (2013) conducted a review of the state-of-the-art techniques used for the classification of the body shapes. It concentrated on the various software for vehicle classification techniques with accuracies between 82% – 95%. A source of the uncertainty is the subtle difference between the vehicle classes, as demonstrated in Figure 5.5. When comparing SUV and Van categories, it leads to some incorrect classifications.

Some researchers believed that direct comparisons between approaches are an inappropriate definition of the vehicle, as body shapes are different between each of the methods. The lack of cohesion is mainly due to the different variations in the subcategories of the "car" category, as shown in Figure 5.6 there are three different definitions of cars. Although a large variation in category is ideally used as key characteristic in the re-identification process, a large number of categories dilute the effectiveness of the classification technique. This is due to the subtle difference between the classes causing higher rates of misclassification and increases to the level of uncertainty.

5.3.5 Vehicle Trajectory

A vehicle's trajectory is not a characteristic commonly used for the re-identification of vehicle. However, it is a measure that could record the driver's usage pattern in driving a given vehicle to judge if a vehicle is being driven by a different driver. From a psychological viewpoint, a driving pattern could be used as a key identifier for the driver of a given vehicle. If the driver is a repetitive user of the car park, the route and driving pattern stays quite consistent, especially in a condition that has limited variability, i.e. only have two alternative routes. Therefore, these features can be used as part of the re-identification problem for answering queries such as "is the vehicle driven by someone else?". The trajectory information relies heavily on the tracking information that is usually obtained through the tracking algorithms dealing with the video data. There are two main classes of tracking algorithms. They are:

1. Data Association - Kalman Filter (Kalman, 1960) proposed the famous approach known as data association. His target tracking process consists of a recursive process where at each frame of the object's location was predicted by using a motion model and then was updated based on the latest observation. The appearance of the target model would be compared with all the records in order to find the closest model that matches in target.

2. Data Driven - Comaniciu et al. (2000) reported the popular mean-shift approach. The mean-shift algorithm does not first segment objects but rather uses information, only retrieved from the data itself to build the target model. The mean-shift method tracks a given target by searching for its model in every image of the sequence.

Both of the tracking models perform equally effectively when tracking a single target in an uncluttered environment. However, when the environment is complicated and multiple targets are in close proximity, the uncertainty associated with the tracked target increases. The reasons may due to those listed below:

- Limited ability to effectively distinguish multiple targets in dense space.
- Occlusions caused by an unexpected change to the expected model by changes of the appearance of the target, which may be caused by:
  - Split - A set of pixels are defined as background, when a tracked target passes through these pixels, they split the target.

- Merge - If two or more targets move close proximity to each other, they could be merged together.

The trajectories can be mapped within an x-y coordinate map, the similarity between two or more trajectories can be the closeness of each of the points on the trajectory. In the experimental bed, the number of alternative routes is limited. Therefore, a lot of users will run similar trajectories so that it can be used as an effective discriminator for the re-identification process. It can also be used as secondary feature to help achieve other security or commercial objectives.

There are situations where there is a road block, this will alter the effectiveness of this feature in cases where the drivers are forced to drive in alternative patterns so that a lot of uncertainty arises when using this feature to identify the driver under these circumstances.

5.3.6 Spatio-Temporal Information

Like trajectory, Spatio-temporal information can also be used to infer the behaviour of the drivers. Two cues inherited with the experimental test-bed are the time of the observation and the gate that the vehicle enters (or exits). Both of the cues are associated with the automaticity of human behaviour. In addition if there is a short time scale between entering and exiting a vehicle, this may demonstrate other behaviour patterns, such as cars driving away because a car park is full or the car is used for delivery.

Both of these cues will have negligible measurement noise as there are multiple cameras involved. However, a very low level of uncertainty may be introduced if the internal clocks are out of sync.

5.4 Summary

This chapter discusses the factors that may influence the uncertainties, from both the hardware and software, in a complete video analytic system. Some of the uncertainties are unavoidable such as the intensity restriction of the camera, and some are introduced by the process to

improve the performance of the system such as mathematical assumption. The level of uncertainty would increase as the information is propagated down the system pipeline, thus decreasing the level of accuracy of the outcome of the system. Current research regarding the measurement (Taylor, 2009; Lira, 2002), quantification (Kennedy and O'Hagan, 2001; Matthies, 2007) and propagation (Ku, 1969; Lee and Chen, 2009) of uncertainty are active research areas for a variety of research fields.

The chapter also carries out a discussion for the type of uncertainties associated with features involved with vehicle. Although some of the challenges have already achieved a very high level of accuracy, uncertainty associated with the result still exists, due to a range of limitations and variations. The discussion reveals that under these limitations, the use of additional features can help to lower the uncertainty, for example, when there is a misreading of the number plate, the vehicle manufacturer's logo can be used as verification to the vehicle identification.

These findings conclude that fusion techniques have merits in combining different features for improving the accuracy of a range of different challenges. They also show that the majority of the video analytic challenges would require either re-identification or classification, therefore it is necessary to have a deeper study for these techniques.

Chapter 6

Vehicle Logo Categorisation

6.1 Introduction

As stated in Chapter 4, the correct identification of vehicles is important in a wide range of surveillance situations. Currently, the completion of this task relies predominantly on the correct identification of the characters on the vehicle's number plate (NP). As Section 5.3.2 indicated, however, there are many challenges in correctly reading the NP A secondary key attribute of the cars is needed to identify the vehicle in these situations. One possible solution is the recognition and classification of a vehicle's manufacturer logo, as these are fixed in the front and rear of the vehicle, similar to the NP. In addition, although there are variations of the same manufacturer's logo, as shown in Figure 5.4in Section 5.3.3, these can't be easily altered once they are installed.

Recent research has relied on the texture information of the vehicle's grille to find a coarse Region of Interest (RoI) where the logo could be finely located, as the majority of manufacturers like to install their logo at the centre of the vehicle's grille. Some manufacturers will, however, also place their logos on top of the bonnet, where grille information becomes less relevant. This is even more apparent when attempting to locate the logo on the rear view of the vehicles, where the grille information is unavailable. This increases the difficulty of the research. To break the limitation, the investigation will devise a new process to locate the logos from both the front and rear views.

The data used for this challenge has been captured by using the apparatus in our experimental test-bed, as demonstrated in Section 4.4. The data will, therefore, simulate vehicles in a real-world situation, where the logos are of varying sizes and viewed from both the front and rear, under varying lighting and environmental conditions. Using the novel logo localisation approach, the classified logo regions are read by the Fisher Discriminative multi-class classifier to determine the most likely category. In this project, the vehicle logo categorisation process is divided into two stages: Logo Localisation and Logo Classification.

6.2 Related Work

This section will first evaluate the current research activities that have been undertaken to tackle both the NP localisation and NP classification challenges, followed by an examination of the multi-classification techniques that have been implemented within a wider research area.

6.2.1 Localisation

The localisation of the logo, in a view of the vehicle, is an essential first step for the achievement of its accurate classification. The majority of localisation research relies heavily on some prior knowledge such as the position of the number plate (NP). Once the position of the NP is located, researchers can define a Region of Interest (RoI) relative to it. A number of authors (Li and Li, 2009; Liu and Li, 2011; Yang et al., 2012) have assumed that the RoI for the vehicle logo is a patch above the NP with a size relative to the extracted NP. Other researchers (Dlagnekov and Belongie, 2005; Lee, 2006; Psyllos et al., 2011b; Petrovic and Cootes, 2004; Psyllos et al., 2010; Wang et al., 2007) defined the RoI as an area on the front of the car specified relative to the size and location of the NP, incorporating the NP and other dominant features such as grille and head lights. Furthermore, instead of one reference point, Lu et al. (2010) adopted three reference points to define a RoI containing the vehicle logo and the grille. Their three reference points are the NP, and the left and right headlights.

From a large RoI, Lee (2006) extracted a smaller area of interest that incorporates the texture

information of the grille and logo. The texture information of the grille was also employed by the other authors (Li and Li, 2009; Liu and Li, 2011; Lu et al., 2010; Yang et al., 2012) which was combined with edge information obtained from different edge detectors and filters. The authors reduced the size of the RoI to incorporate the logo alone. Wang et al. (2007) suggested the use of the peaks in the edges' vertical direction projection as the initial location, to start a symmetry search in order to locate the logo. Psyllos et al. (2011b) and Psyllos et al. (2010) stated a completely different method known as the Phase Congruency Feature Map and its derivatives to divide the RoI into smaller areas such as left/right light, grille and logo.

An attempt to remove the dependency on the NP is described by Sam and Tian (2012), who utilised the Modest Adaboost (Freund and Schapire, 1995) algorithm to search for vehicle logos, represented by extended Haar-like features. However, the gradually sliding window used in the search makes the method sensitive to the complicated background, thereby limiting its application. The Zhang and Zhou (2012) method applied the frontal images of the vehicles and adopted a bilateral symmetry detection based on a set of Size-Invariant Feature Transform (SIFT) features (Lowe, 1999). Although this method has claimed a localisation accuracy of 98.91%, its reliance on the grille information makes the method unsuitable for rear-view logo localisation.

6.2.2 Vehicle Logo Classification

Previous research on the categorisation of a vehicle manufacturer's logo is inadequate, even though it is an attribute which is useful in a vehicle identification system. Early work by Dlagnekov and Belongie (2005) used SIFT features to re-identify a vehicle from the whole rear-view of the vehicle, not just the logo. This approach attained 89.5% re-identification rate of total 38 test samples. Psyllos et al. (2011b, 2010) elaborated Dlagnekov and Belongie (2005) work of proposing a SIFT-based enhanced matching scheme, which only concentrated on the logo. The scheme boosted the categorisation accuracy higher than the standard SIFT-based feature-matching method developed by Dlagnekov and Belongie (2005). Wang et al. (2007) presented a method for logo categorisation that exploited a template matching and a histogram of orientation gradients (HOG) of the logo. The methods proposed by Wang et al.

(2007), Psyllos et al. (2010) and Psyllos et al. (2011b) faced the issue of reduced robustness under variation of the environmental conditions, such as lighting levels. An improved solution was recommended by Burkhard et al. (2011). The solution relied on the Fourier shape descriptors, introduced by Zhang et al. (2001), who characterised shapes based on their curvature as they are not sensitive to distortion due to changes in environmental lighting effects. However, this method is highly dependent on the logo segmentation, as it requires the logo to be the dominant element in the RoI. Recently, Sam and Tian (2012) applied the invariance property of Radial Tchebichef moments (Mukundan, 2005) to recognise high resolution segmented vehicle logos. They achieved a recognition rate of 92% with aid of a manually extracted logo RoI. Their test sets contained 200 images of 10 different categories.

Although there has previously been limited focus on the recognition and classification of vehicle logos, some research based on a RoI from a frontal, or rear view of the vehicle to identify the vehicle model has been done. Lee (2006) advised a set of 16 texture descriptors of the RoI taken from the front view of the vehicle as the input to a 3-layer back propagation multi-layer perceptron neural network. This method was used to classify vehicles into 24 different models and achieved a recognition rate of 94%. Petrovic and Cootes (2004) recommended the RoI from the front of the vehicle in an approach based on HOG and launched a similarity measure between a test and target, i.e. dot product and euclidean distance, to determine the vehicle model. Petrovic and Cootes reported an identification rate of over 93% on parked cars. Their dataset contained 77 models. Zhang and Zhou (2012) proposed the use of a Rotation Forest Ensemble method, as introduced by Rodriguez et al. (2006), for vehicle classification. Zhang's method relied on the features from a Fast Discrete Curvelet Transform (Candes et al., 2006) and the Pyramid Histograms of Orientated Gradients (pHOG) (Bosch et al., 2007) of the RoI from the top view of the vehicle, with a success rate of 96.5% on a 21 model dataset. Similar to the approach of Dlagnekov and Belongie (2005), Bhanu and Kafai (2012) tried to classify vehicles using the rear view of the vehicle, with the vehicle being categorised into classes of vehicle type, such as vans, cars or trucks, rather than make or model, with a success rate of 95.7% for a four class dataset.

Iqbal et al. (2010) conducted a comparison of the techniques used previously for vehicle model classification, for both environmentally controlled and uncontrolled datasets. They noted that techniques such as SIFT are sufficient in controlled environments, where there is little variation of illumination, viewing angle and scale. Their research also concluded that for make or model recognition, the RoI from rear-view images performed better than the RoI of the frontal view due to fewer variations caused by the grille.

6.2.3 Metric Learning for Classification

Classification methods can be broadly categorised into feature-based and learning-based methods. Feature-based methods rely on the discriminative ability of the feature alone, while learning-based methods, such as that used in Chapter 3, estimate a discriminative model by analysing the training data, representative of the collected data.

SVMs (Joachims, 1999), Boosting (Freund and Schapire, 1997) and Neural networks (Bishop, 1996) have successfully been employed to learn two class classifiers in various vision related problems such as Pedestrian detection and Face recognition. Multi-class classification has been addressed successfully in learning methods by different authors (Weinberger and Saul, 2009), (Xing et al., 2003), (Ying et al., 2009) that mainly focused on metric learning that requires a Mahalanobis distance metric to be estimated in the feature space. The feature space is often non-linear in nature and needs a transformed feature space, in which, the Euclidean distance between data samples maintains the neighbourhood characteristics of data.

Metric learning has been considered as a data association problem when multiple classes are involved. The Mahalanobis metric is consistent with a positive semi-definite matrix, and the general set of such "metric matrices" – all of which are positive semi-definite, and can be considered to be the interior and surface of a cone with the apex at the origin. Other methods such as Local Distance Metric (Liu and Rong, 2006), LMNN (Weinberger and Saul, 2009) and that of Xing et al. (2003) estimated this metric by modelling the solution as an optimization problem where strategies like gradient descent approaches are employed. However, scalability with increasing feature dimensions tends to be problematic with such

approaches due to the computationally expensive, time consuming iterative steps involved. A few recent methods such as that proposed by et al. Kastinger et al. (2012) have modelled the solution for estimating the metric in eigenspaces. The solution in these cases can be easily programmed with aid of solving a formulated eigenvalue problem. The Local Fisher (LF) method (Pedagadi et al., 2013) showed an approach to produce better discrimination amongst sub space methods by using relatively simple features.

6.3 Vehicle Logo Categorisation System Design

This section is structured as follows: The strategy employed to extract the RoI and the relevant features are first described followed by an overview of the system processing pipeline. An overview of the Local Fisher Discrimination techniques used for the classification will then be given. The section concludes with the description of the decision fusion model that is going to be deployed in the classification stage.

6.3.1 Feature Extraction

As discussed in Section 6.2.1, the most common placement of a manufacturer's logo on the frontal and rear views is at a position with some distance above the NP. Therefore, to extract the RoI, the Automated Number Plate Recognition (ANPR) module is employed to accurately detect the position of the NP. The relationship between the coordinates of the corners of the NP and the logo's RoI is shown in Figure 6.1, and is expressed in the below set of Equations 6.1.

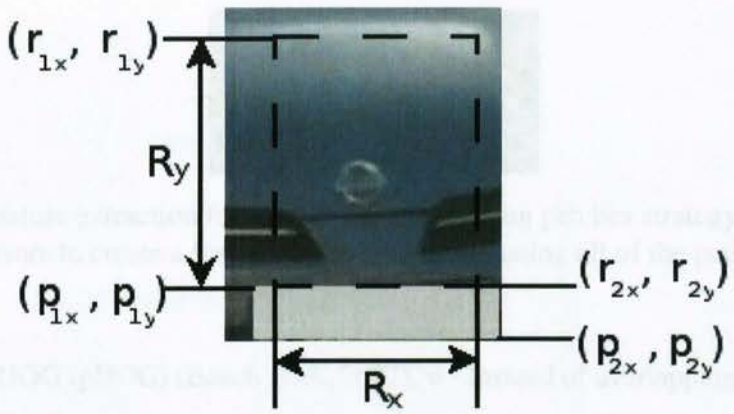


Figure 6.1: Logo RoI Extraction Referring to Equation 6.1, $\mathbf{p}_1$ and $\mathbf{p}_2$ are the respective bottom-left and bottom-right coordinates of the located NP. $\mathbf{r}_1$ and $\mathbf{r}_2$ are the respective top-left and bottom-right coordinates of the located RoI. R_x and R_y denote the width and height of the RoI with 128 and 152 pixels respectively. The current R_x and R_y values are set to allow the capture of the logo in our dataset at varying distances from the camera, and the values can be varied depending on the need.

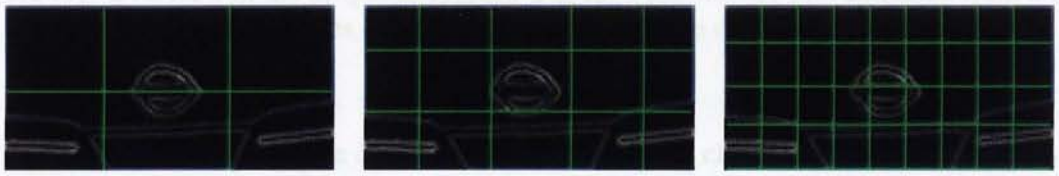
$$\begin{cases} R_x = r_{2x} - r_{1x} \\ R_y = p_{1y} - r_{1y} \\ where: \frac{\mathbf{r}_{2x}+\mathbf{r}_{1x}}{2} = \frac{\mathbf{p}_{2x}+\mathbf{p}_{1x}}{2},\ r_{1x} = p_{1x},\ r_{1y} = p_{1y} + R_y \end{cases} \tag{6.1}$$

Once a logo's RoI is extracted, the region is sub-divided into patches. The patches are converted into two types of feature vector, based on the Histogram of Orientated Gradient (HOG). The first stage of the operation is to convert the RGB patch to HSV space, followed by the application of a Sobel edge detector on the grey level image in the V space. The two different HOG feature vectors are finally extracted as follows:

1. Multiple Overlapping Patches u - By adopting the feature fusion strategy stated in Chapter 3 and illustrated in Figure 6.2. Each patch is divided into smaller overlapping bounding boxes of 16 horizontal and 32 vertical pixels. For each overlapping bounding box the pixels' orientations in the box are picked up to form an 8-bin edge histogram. All of the histograms are concatenated together to build the feature vector $\mathbf{u}$

Figure 6.2: Feature extraction for u, uses the overlapping patches strategy, which acts as different sensors to create a fused feature by concatenating all of the patches together.

2. Pyramid-HOG (pHOG) (Bosch et al., 2007), **v** - Instead of overlapping bounding boxes, the edge image is divided into grids, an 8-bin edge histogram is created for each grid and concatenated together. 3 different levels of grids are used as illustrated in Figure 6.3. The histograms for each grid levels are finally combined together to form the feature vector **v**



Level 1 Level 2 Level 3

Figure 6.3: An example of pHOG and the grid arrangement of the three different levels. The three levels are concatenated together to create the fused vector v

The different types of edge information are useful for estimating a reliable embedding space in the subsequent stages. Instead of doing the dimensionality reduction after combining the two feature vectors, as explained in Chapter 3, the reduction is done prior to fusing together **u** and **u**, which will be stated in the following section.

6.3.2 System Processing Overview

As mentioned in 6.3.1, once a logo's RoI is extracted, the region is sub-divided into patches. Each patch size is 128 by 64 pixels, and the first patch is at the top left of the RoI and successive patches are created by moving down with an interval of 5 pixels. The patch width is currently set as the whole width of the RoI to allow the logo to be located at various angles of the viewpoint, i.e. the logo does not have to be located at the horizontal centre of the patch, as shown in Figure 6.4.

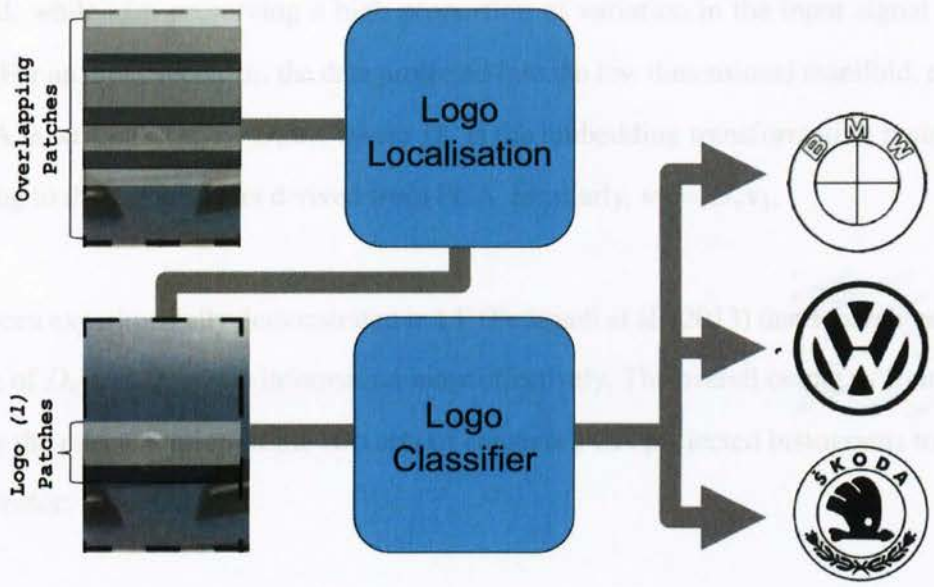


Figure 6.4: Overview of the proposed system. The logo patches are first extracted from all of the overlapping patches. Using the logo patches, the class decision is made.

As illustrated in Figure 6.4, the extracted patches are firstly classified into two categories in the localisation stage, logos (l) and background (b). A logo patch is defined as a patch that contains at least 50% of the logo in the vertical direction, and all other patches are defined as background patches. Only the logo patches are used in the logo classification process.

6.3.3 Local Fisher Discrimination

Local Fisher (LF) (Pedagadi et al., 2013) explores the idea of projecting feature data into two successive sub-spaces. The first sub-space is estimated by employing the dimensionality reduction technique of Principal Components Analysis (PCA)(Jolliffe, 2005) and the second subspace by the application of a supervised dimensionality reduction method of Local Fisher Discriminant Analysis (LFDA) (Masashi, 2006) on the PCA projected feature data. A brief review of the LF is presented below.

A low dimensional embedding space is obtained from the high dimensional feature space by firstly estimating a PCA transformation separately, for each of the two input feature vector types u and v. Principal Component Analysis enables the dimensionality of the data to be

reduced, while also preserving a high proportion of variation in the input signal (Jolliffe, 2005). For an input vector $\mathbf{u}_i$, the data projected into the low dimensional manifold, estimated by PCA is written as $\mathbf{u}'_i = D_u\mathbf{u}_i$, where D_u is the embedding transformation matrix corresponding to the eigenvectors derived from PCA. Similarly, $\mathbf{v}'_i = D_v\mathbf{v}_i$.

It has been experimentally demonstrated in LF (Pedagadi et al., 2013) that separate estimation and use of D_v and D_u retain information more effectively. The overall output $\mathbf{x}_i$ from the first stage is the concatenation of the two sets of separate PCA projected histograms to create a fused vector: $\mathbf{x}_i = \{\mathbf{u}'_i|\mathbf{v}'_i\}$.

LF combines the neighbourhood preserving property of Locality Preserving Projection (LNP) (Xiaofei, 2004) with the traditional Fisher Discriminant Analysis (FDA) (Fisher, 1936). It is very common for a multi-class dataset to be multi-modal in nature, i.e. to show a significant variation in class samples. LF captures this multi-modality in classes by constructing an affinity matrix A that estimates the neighbouringcharacteristics of the dataset. A local scaling method (Zelnik-Manor and Perona, 2004) is used for the estimation of A by choosing the n-th nearest neighbour and assigning individual scaling factors for samples from the same class.

The width between class S^W and class S^B of scatter matrices in traditional FDA is weighted with the affinity matrix A so that the far apart in-class samples do not contribute to the estimation.

$$S^W = \frac{1}{2}\sum_{i,j=1}^{n} A^w_{i,j}\left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^t \tag{6.2}$$

$$S^B = \frac{1}{2}\sum_{i,j=1}^{n} A^b_{i,j}\left(\mathbf{x}_i - x_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^t \tag{6.3}$$

where

$$A^w_{i,j} = \begin{cases} A_{i,j}/n_c & \text{if } y_i = y_j = c \\ 0 & \text{if } y_i \neq y_j \end{cases} \tag{6.4}$$

$$A^b_{i,j} = \begin{cases} A_{i,j}\left(\frac{1}{n} - \frac{1}{n_c}\right) & \text{if } y_i = y_j = c \\ \frac{1}{n} & \text{if } y_i \neq y_j \end{cases} \tag{6.5}$$

Here, n_c is the number of samples in class c and n is the total number of samples. The transformation matrix T_{lfda} is defined as:

$$T_{lfda} = \arg\max tr\left(\left(T^t S^W T\right)^{-1} T^t S^B T\right) \tag{6.6}$$

where $T \in \mathbb{R}^d \times \mathbb{R}^m$. Similar to FDA(Fisher, 1936), the estimation of T_{lfda} is achieved by representing the above as a generalised eigenvalue problem, $S^B\varphi = \lambda S^W\varphi$, here $\{\varphi_i\}$ and $\{\lambda_i\}$ are the eigenvectors and eigenvalues of this system. The final projection into the embedding space characterised by LFDA can be written as:

$$\mathbf{z}_i = T^t_{lfda}\mathbf{x}_i \tag{6.7}$$

The similarity measure between any two observations i and j is given by the Euclidean distance between the LFDA transformed vectors of each observation

$$D(i,j) = |\mathbf{z}_i - \mathbf{z}_j| \tag{6.8}$$

6.3.4 Logo Localisation Using LF

To localise the logo patches, the training set of logo (l) and background (b) patches is used, as defined in section 6.3.1, where ground-truth has been manually selected. At this section, the vehicle class label is not used any more. The training data is used to estimate the matrix T^t_{lfda} that transforms the feature vectors $\{\mathbf{x}_i\}$ to their representation in the embedded space $\{\mathbf{z}_i\}$. Let $\{\mathbf{z}^l_i\}$ and $\{\mathbf{z}^b_i\}$ to be the sets of n_l and n_b training vectors in the embedded space for logo and background patches, respectively. A new test vector $\mathbf{z}^*$ is classified as either logo or

background on a k-nearest neighbour basis:

$$\mathbf{z}^* \quad \varepsilon \quad \begin{cases} \text{logo} & \text{if } d_l < d_b \\ \text{background} & \text{otherwise} \end{cases} \tag{6.9}$$

where

$$d_{l,b} = \min_{\mathbf{z}_i \epsilon (\{\mathbf{z}_i^l\}, \{\mathbf{z}_i^b\})} (|\mathbf{z}_i - \mathbf{z}^*|) \tag{6.10}$$

Using a LF-based binary classifier on the ensemble of patches within a RoI, the patches categorised as a 'logo' are classified into one of the N manufacturer logos, which are input into a voting process to provide a final estimate for this vehicle. This will be described in the next section.

6.3.5 Logo Classification Using LF

Here, a fusion reasoning technique based on Voting, as introduced in Section 2.4.2, is used. For a given RoI, n patches are categorised as 'logo'. If $n > 0$, a LF-based multi-class classifier is adopted in an analogous manner to assign a predicated class y_i to each logo patch, where $y_i \in \{1, \ldots, N\}$, where there are N categories corresponding to the different vehicle manufacturers. Otherwise, if $n = 0$, no suitable patches are available and the classification cannot proceed. The overall manufacturer class assigned to the RoI, $y_{\max}$ is the class that the largest number of individual logo patches belongs, as follows:

$$y_{\max} = \max_{1 \leq j \leq N} \left(\sum_{i=1}^{n} \delta(y_i, j) \right) \tag{6.11}$$

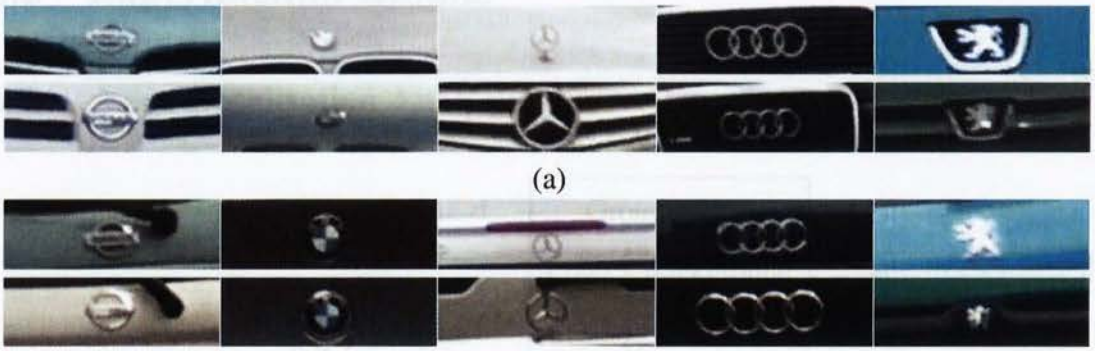
$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{Otherwise} \end{cases} \tag{6.12}$$

If the voting process does not result in an outright winner, a second weighted voting procedure is adopted. For each logo patch, there is a logo confidence measure represented by the value of d_l. For all the equal top ranking classes, from the first vote, their corresponding patch d_l values are summed together with the class that has lowest cumulative value to become the

overall winner.

6.4 Logo Dataset

As mentioned before, the data involved was captured in our experimental test-bed. The footage captured vehicles entering and exiting the car park at different velocities, trajectories and under varying environmental conditions such as lighting. The footage was first segmented by the process described in Section 4.4.3.1 to extract the objects of interest, and then processed by existing ANPR software to acquire the NP coordinates for the vehicles in this study.



(a)

(b)

Figure 6.5: Examples of Vehicle Logo Patches. a) Logos' Frontal View, and, b) Logos' Rear-View. Showing that the current logos are of various sizes and its location is no longer limited to being in the front of the vehicle and within the grille

As Section 5.3.3 outlined, one issue with the manufacturer's logo is that some have quite similar outlines. In order to effectively test the classification, the following five commonly used classes of manufactures were selected for the experimentation. They are: Nissan, BMW, Mercedes (Merc), Audi, and Peugeot (PG). Examples of the logos are given in Figure 6.5, the figure illustrates that most of the outlines are different. To test the discriminative abilities between similar outlines, the Mercedes and BMW are included. Both of these badges have a circular silhouette with different insides. Figure 6.5 also demonstrates the dataset will contains logo from both the front and the rear of the vehicle, and these have been captured from various angles,which will add an additional layer of complexity.

For each manufacturer's class, 30 training samples and 20 testing images were chosen. The

training data is excluded from the testing data in order to test the true performance of the system. Manual ground-truth locates the vertical centre of the logo in the image in order to define the logo patches.

For each image, the logo's RoI was located by the NP position, and subdivided into patches, which results in 18 patches per RoI. Therefore, there were 2700 training and 1800 testing samples. The combined feature vector for each sample, before dimensional reduction, is 2248 components.

6.5 Experimental Results

6.5.1 Logo Localisation Analysis

		Ground Truth	
		Logo	Background
Predicted	Logo	86%	14%
	Background	9%	91%

Table 6.1: Confusion Matrix of Logo and Background Patch Classification Results

The accuracy of the vehicle logo localisation was validated against manually labelled, ground truth data. The results in Table 6.1 show that the method involved in this project was able to achieve 86% accuracy for classifying logo patches and 91% for correct background patches. If only unsupervised PCA is used, the results decrease to 80% for logo and 83% for background patches.

The 86% correct logo classification actually means that 97% of all testing samples would have at least one correctly predicted logo patch that could be forwarded to the logo manufacturer classification stage.

		Ground Truth				
		Nissan	BMW	Merc	Audi	PG
Predicted	Nissan	80%	5%	0 %	0%	15%
	BMW	0%	100%	0 %	0%	0%
	Merc	5%	21%	69%	5%	0%
	Audi	0%	0%	0%	100%	0%
	PG	0%	5%	0%	5%	90%

Table 6.2: Confusion Matrix for Logo Classification Results Using Badge Patch. The results shows an 85.56% classification rate based on 5 vehicle logo classes

6.5.2 Logo Classification

The system was trained using only the ground truth logo patches and the trained model was tested by previously unseen classified logo patches from the local localisation stage. Table 6.2 shows the confusion matrix of the logo classification results using the correctly classified logo patches only (lpo). The main diagonal shows the percentage of correctly classified manufacturer class. When all of the predicted logo patches (plp) are used, including the background patches incorrectly classed as logo patches, the overall classification rate of 85.56% is obtained, as shown in Table 6.3. The value of 85.56% indicates the performance of the system in a real life environment.

Table 6.2 illustrates the challenges when there is a variation of the level of confusion when classifying badges with similar shapes. The Mercedes classification result demonstrates a high level of confusion, however, the BMW badge results have zero confusion. A possible explanation might be due to the BMW badges having a constant filled centre, irrespective of the vehicle colour. With very limited variation, compared to the case of Mercedes badge, where the colour to fill the centre differs depending on its location or the colour of the vehicle, the effect will result in different variations in its representation as the edge detection techniques are used to extract the features.

The addition of the Local-Fisher learned metric to the PCA feature space, significantly improves the performance of the system, as demonstrated in Table 6.3, comparing with the principal components of the original feature vector.

Type	HOG		pHOG		HOG + pHOG	
	lpo	plp	lpo	plp	lpo	plp
PCA	70.40%	69.70%	68.04%	68.04%	70.41%	69.70%
PCA + LF	87.62%	85.67%	80.11%	79.80%	87.62%	85.67%

Table 6.3: Logo Classification Success Rate With Different Features

Table 6.3 also shows that identical performance is achieved between HOG only and the combination of both HOG and pHOG. To further validate this finding, an additional cross-validation experiment is done, the results are shown in Figure 6.6.

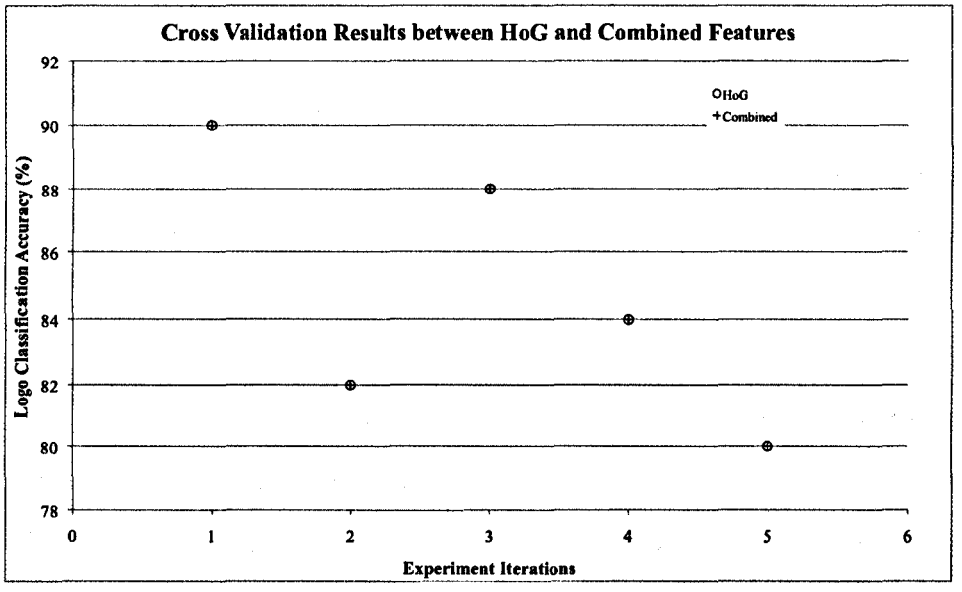


Figure 6.6: Cross-validation experiment, where 40 training and 10 testing samples for each class is chosen which is repeated for 5 iterations. For this experiment, only the groud truth logo patch is used. This removes the uncertainty that may be introduced by the Logo Localisation stage.

Figure 6.6 shows 5 iterations of a cross validation experiment. The results demonstrate that no extra advantage is gained when a combined feature is used, compared to RGB and HSV features in Chapter 3, where the feature vectors is created from two separate representations of the input. The HOG and pHOG gather the same feature representation which is output from the Sobel edge detection. The difference between the HOG and pHOG is actually how the data is gathered into their respective histograms. This experiment illustrates that the granularity of the patched HoG process means that all the key information that pHOG would have captured has already been acquired, therefore no extra advantage is gained when HOG and pHOG are combined.

6.6 Summary

This chapter has presented a novel vehicle logo localisation and classification process, with aid of features composed of local histograms of gradients and the use of Local Fisher Discrimination Analysis, to obtain a more effective learning metric. The proposed method relieves the reliance of needing to locate dominant features, such as the grille, in order to localise the logo. Thus, the approach can be applied to locate vehicle logos on both the front and rear of a vehicle.

The results achieved by the process are not directly comparable to those from recently published techniques, as those only concentrate on locating the logo on the front of the vehicle and the data used are captured in controlled environments, such as in the studies of Wang et al. (2007) and Psyllos et al. (2010). As such, the results achieved in this project provide a benchmark for techniques of logo recognition on medium-view CCTV data in a video surveillance environment.

Although Chapter 3 has shown the advantages of concatenating similar representations of the features to improve the performance, experiment results in this chapter have confirmed that the choice of features is important. If two representations of the same feature space, i.e. Sobel edge detected image, then the two features might not contribute enough independent information. As a result of this, no extra advantages could be gained. However, the method developed in this project has revealed that decision fusion techniques can be used to assist with video surveillance challenges.

Chapter 7

Fusion Models and Evaluation Framework

7.1 Introduction

As stated in Chapter 1, there are many challenges within Automated Video Surveillance system. There are two main problems that any system needs to overcome. The first is the uncertainty that exists throughout a particular automated video surveillance system's pipeline, as demonstrated in Chapter 5. The second is the identification of a suitable method to fuse the information from different sources in order to conduct various types of objectives, rather than a single objective, as outlined in Chapter 4.

Information Fusion is the technique that can be used to overcome these problems. It can result in more accurate inference than a single sensor does (Hall and McMullen, 2004). The inferences made with these techniques range from simple estimates of the identity of certain entities to complex inferences about current or future relationships between multiple entities and the events involved.

As the types of information in this chapter are incommensurate, the data must be fused at a *decision level* (Hall and Llinas, 1997). Decision-level fusion, as outlined in Section 2.3.3, consists of merging information at a higher level of abstraction and combining the results from multiple algorithms to yield a final 'fused' decision (Dong et al., 2009). The data is therefore

processed separately by multiple algorithms, e.g. to identify and classify observed entities and events, based on their different features. The resulting information is combined with a chosen set of decision rules to obtain an overall inference. Due to the uncertainties created at every step of the processing pipeline, the result could be expressed in a probabilistic form. Two well-known types of decision rules, which allow the fusion of probabilistic outcomes, are Bayesian inference and the Dempster-Shafer (DS) theory. These can both be applied to the designed fusion scenario to combine multiple cues extracted from surveillance videos.

The two main outcomes of this chapter are, firstly, a theoretical investigation on the appropriate formulation of these two fusion models for the target scenario. Secondly, as suggested in Section 2.4.4, it seems that any previous proposals could not conduct a general performance evaluation for these two fusion methods. Thus, this chapter will create a generic evaluation framework to evaluate Bayesian estimates and be extended to accommodate the DS methodology.

The remainder of the chapter is organised as follows. In Section 7.2 and 7.3, a detailed design with Bayesian and Dempster-Shafer theory is given, including a theoretical evaluation. In Section 7.4 the generalised evaluation framework is introduced. Some experiments are conducted in Section 7.5 and the discussion will be presented in Section 7.6.

7.2 Statistical Parametric Fusion Methods: Bayesian Inference

7.2.1 Bayesian Theory

In Bayes' theorem Bayes et al. (1984), it is assumed that h_i is a hypothesis about a state, taking values in the set of hypotheses $H = h_1, ...h_n$, exactly one of them is 'true', and the remainder 'false'. The prior probabilities, $P(h_i), i = 1, ..., n$ constitute the prior probability

mass function of the hypotheses h_i:

$$0 \leq P(h_i) \leq 1 \text{ and } \sum_{i=1}^{n} P(h_i) = 1 \tag{7.1}$$

Normally, the hypothesis with the highest prior probability will be assumed as the 'true' one. However, a more accurate estimation of the state can be made by incorporating some relevant 'posterior' evidence x. It is assumed that the h_i are distributed according to the class-conditional probability distribution function $P(x|h_i)$ (Jensen, 1996). Therefore, given the prior probability and the class conditional probability, the posterior probability can be calculated by Bayes' formula:

$$P(h_i|x) = \frac{P(h_i)P(x|h_i)}{\sum_{j=1}^{n}(P(h_j)P(x|h_j))} \tag{7.2}$$

The denominator is the 'evidence factor' that normalises the posterior probabilities so that they will sum to one.

The Bayes formula has gained popularity as a fusion method because it provides a direct and easily applicable means for combining the prior information with the current observation.

7.2.2 Graph Theory - Bayesian Network

To create complex Bayesian systems, Jensen (1996) introduced the ideal of combining Bayesian theory with Graph Theory to produce the notion of Bayesian Networks. According to Druzdzel and Van Der Gaag (1995), there are two distinctive parts to a Bayesian Network: Qualitative and Quantitative.

7.2.2.1 Bayesian Network: Qualitative Analysis

The qualitative part is defined by the structure of the Network, which is represented by a set of random variables and their conditional dependencies via a directed acyclic graph, as modelled in Figure 7.1.

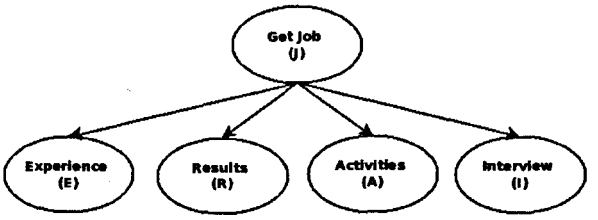


Figure 7.1: Example of Bayesian Network Formulation For Getting a Job, where the root node of the query under investigation and child nodes are the evidence in support of the query.

A Bayesian Network is created by different nodes and linked by edges. The edges represent the relationship between the nodes. There are three types of nodes:

1. Root nodes, representing the queries that the system has been constructed to answer. Edges are only directed *away* from root nodes.

2. Child nodes, representing the evidence, such as one extracted from surveillance sensors that may provide information about the queries. Child nodes only have edges directed *towards* them.

3. Parent nodes are the Root nodes of sub-networks in a complex network. Therefore, a parent node is a child node of a Root node and it is the Root node of child nodes in the sub-network. As such, it will have edges directed to and from it.

7.2.2.2 Bayesian Network: Quantitative Analysis

The quantitative part comprises the values of the variables of the child nodes and is recorded into the Conditional Probability Matrix (CPM). An example of a sensor's CPM is given in Section 7.2.3.1, where each column represents the class-conditional probability distribution function for the hypothesis.

These conditional probabilities could be generated by domain experts and/or data obtained directly from observations made about the environment. Furthermore, these can be updated over time to improve the accuracy of the model.

7.2.2.3 Design Assumptions

The evidence, represented by the child nodes, is included in the Conditional Probability Matrix to calculate the joint probability of the network. However, Equation 7.2 needs to be adapted for the calculation of joint evidence, as the class-conditional probability distribution function for a joint set of evidence $P(x_1, ...x_c|h_i)$,in general, does not have an analytic solution. It is impossible to numerically evaluate this for all instances of the evidence due to the extensive number of combinations.

To resolve these issues, it is assumed that the child nodes are conditionally independent. A child node is assumed conditionally independent if the knowledge of a child node does not change the belief of any other child nodes in the network. In addition, child nodes are conditionally independent if the state of the root node is known.

This assumption allows the child nodes' co-occurrences to be calculated as a simple multiplication. Equation 7.3 is now transformed into:

$$P(h_i|\cap_k^c x_k) = \frac{P(h_i)\prod_{i=1}^{k} P(x_k|h_i)}{\sum_{j=1}^{n}(P(h_j)\prod_{i=1}^{k} P(x_k|h_j))} \tag{7.3}$$

In addition, the inference of multiple sensors can be calculated sequentially, as discussed by Lewicki (2007), whereby the posterior probability provided by one sensor can be used as a prior probability for the following sensor's calculations.

7.2.3 Bayesian Network: Surveillance Scenario Design

In this section, the theoretical adaptation of Bayesian Network for the Surveillance Scenario under investigation will be conducted. The section firstly outlines a simple framework to show the uncertainty reduction capabilities of the Bayesian Network (BN) in the experiment. The framework design will be followed by the experiment's results. The framework will be further improved to show how it can be adopted with sensors for the investigation of a particular surveillance scenario. Finally the framework will be further modified to demonstrate its ability of inferences for different scenarios.

7.2.3.1 Uncertainty Reduction -Simple Framework

In this section, the vehicle-identification problem is used to aid the evaluation of uncertainty reduction. The Bayesian Network for this scenario is illustrated in Figure 7.2.

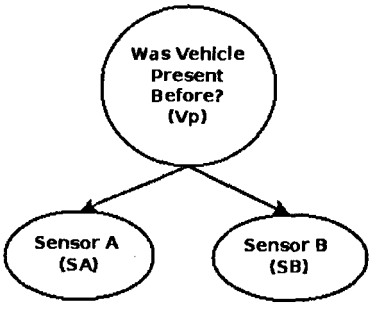


Figure 7.2: Simplified Bayesian Network Inference: "Was This Vehicle Present Before?". The root query is supported by two sensors as the evidence

In the simple network shown in Figure 7.2, the root query is: whether a particular 'probe' vehicle is present in a 'target' set. The prior probability of the root is given in Table 7.1.

	$P(V_p)$	$P(\overline{V_p})$
Seen Before	0.1	0.9

Table 7.1: Prior Probability of the Root Node of Figure 7.2

The network is informed by two sensors, Sensor A (SA) and Sensor B (SB). For each member of the 'target' set T_i, ($i = 1, 2, 3...n$, where n total members), each sensor provides an independent measurement, and outputs a positive identification ('Yes') if the 'probe' vehicle matches the target and a negative identification ('No') otherwise. Therefore, for each sensor, n independent measurements are obtained. In the case of a *perfect* sensor, when the vehicle *was present*, there should be exactly one 'Yes', and when the vehicle *was not* present, then all measurements should be 'No'.

However, when the sensors are imperfect, the Conditional Probability Matrix (CPM) of each sensor can be used in conjunction with Prior probability to calculate the outcome. The CPM probabilities can be estimated by the sensors' performance on test datasets by using known

probe and target sets, as described in Section 7.2.2.2. The CPM of near-perfect sensors are given in Table 7.2 and 7.3.

Sensor A	$P(SA\|V_p)$	$P(SA\|\overline{V_p})$
Yes (SA_1)	0.999	0.001
No (SA_2)	0.001	0.999

Table 7.2: Sensor A CPM, illustrating a sensor error rate of 0.1%

Sensor B	$P(SB\|V_p)$	$P(SB\|\overline{V_p})$
Yes (SB_1)	0.999	0.001
No (SB_2)	0.001	0.999

Table 7.3: Sensor B CPM, illustrating a sensor error rate of 0.1%

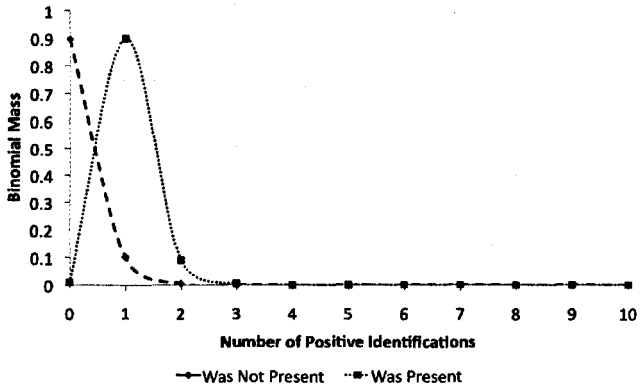
The imperfections in the sensors will result in a varying number of positive ('Yes') identification, represented by Z (where $Z = 0.....n$). The discrete probability of obtaining any specific value of Z from the n independent tests can be simulated by a Binomial Distribution using the information provided by the CPM tables. The binomial mass function determines whether the target vehicle is "not present" in the target set and is given by Equation 7.4.

$$B_{(\overline{V_p})}(Z, N, P(S_1|\overline{V_P})) = \binom{N}{Z} P(S_{A,B}|\overline{V_P})^Z (1 - P(S_1|\overline{V_P}))^{N-Z} \tag{7.4}$$

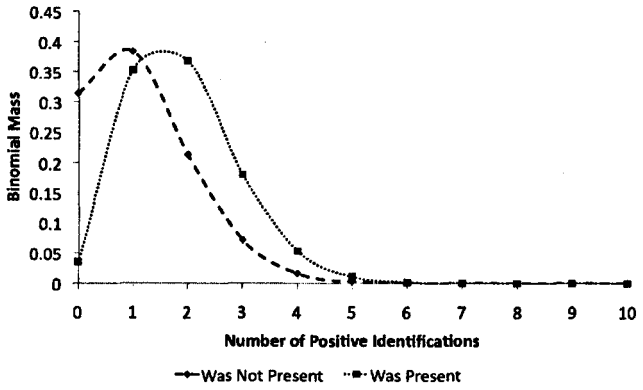
where Z is the variable representing the number of positive responses from the $N = n + 1$ total number of possible outcomes, and $P(S_1|\overline{V_P})$ could be either $P(SA_1|\overline{V_P})$ or $P(SB_1|\overline{V_P})$ depending on the sensor under investigation. The binomial mass function of getting Z positive identifications when it "is present" is given by Equation 7.5.

$$B_{(V_p)}(Z) = P(S_1|V_P)B_{(\overline{V_p})}(Z-1, N-1, P(S_1|\overline{V_P})) + P(S_2|V_p)B_{(\overline{V_p})}(Z, N-1, P(S_1|\overline{V_P})) \tag{7.5}$$

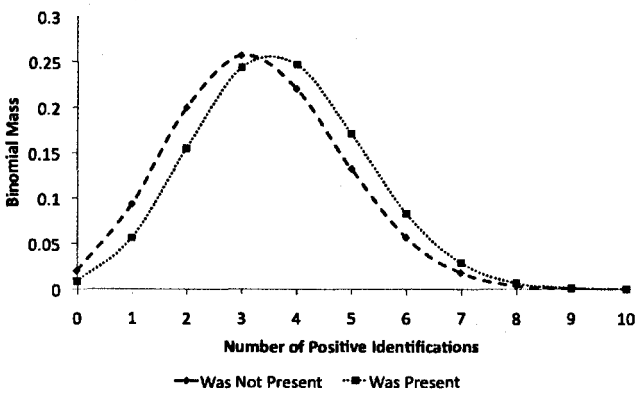
This equation takes into consideration that in the V_p scenario where one positive identification must exists within the n independent tests. An example of two distributions for one sensor and $n = 10$ is given by Figure 7.3.



(a)



(b)



(c)

Figure 7.3: Binomial Distribution of Equation 7.4and 7.5 for $n = 10$; (a) $P(S_1|\overline{V_P}) = 0.01$; (b) $P(S_1|\overline{V_P}) = 0.1$; (c) $P(S_1|\overline{V_P}) = 0.3$

As the error increases, the difference between the distribution becomes very small. As illustrated in the case of $Z = 2$ as the error rate is 30%, the probability for the two outcomes is almost equal (50%)

Figure 7.3 illustrates the changes in the discrete distribution with changes in the error rate

of the sensor. In this simulation, it is the difference between two discrete distributions that provides the information required for the Bayesian Network calculations. The Bayes equations are represented by Equation 7.6 and 7.7.

$$P\left(V_p|Z\right) = \frac{P(V_P)P(Z|V_P)}{P(V_P)P(Z|V_P) + P(\overline{V_P}))P(Z|\overline{V_P})} \tag{7.6}$$

$$P\left(\overline{V_P}|Z\right)\big) = \frac{P(\overline{V_P}))P(Z|\overline{V_P})}{P(V_P)P(Z|V_P) + P(\overline{V_P}))P(Z|\overline{V_P})} \tag{7.7}$$

where Z represents the discrete probability of getting Z a positive indication from the sensor.

7.2.3.2 Uncertainty Reduction - Experiment

To simulate a real-life scenario, the discrete distribution for the sensors will concentrate on the application of target set of 300, $n = 300$. A portion of the distribution is illustrated in Figure 7.4.

To measure the amount of information uncertainty, Shannon Entropy (Shannon and Weaver, 1949) is employed. Based on Shannon's equation, the prior entropy is calculated by Equation 7.8:

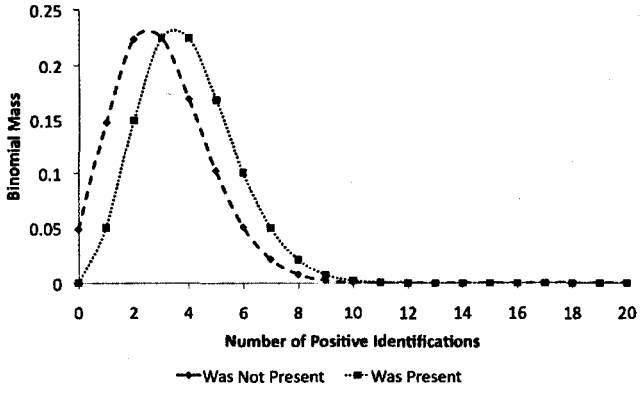
$$H_{prior} = -\sum_{i}^{m} p_i \log_{10}(p_i) = -P(V_p) \log_{10}(P(V_p)) - P(\overline{V_p}) \log(P(\overline{V_p})) \tag{7.8}$$

where p_i is the probability of the hypothesis i, and m is the total number of possible hypotheses in this case of $m = 2$. Based on the information provided in Table 7.1 and Equation 7.8, the prior uncertainty in the network is 0.148.

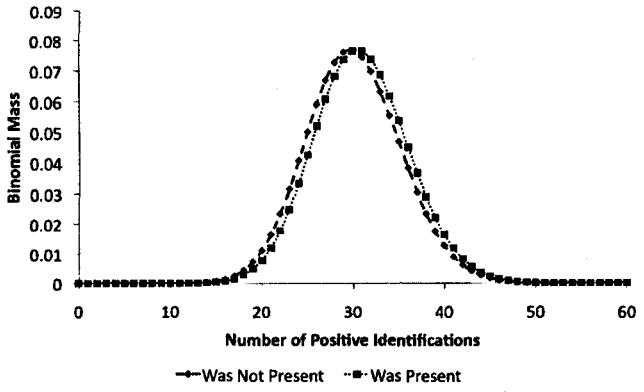
To evaluate the reduction in the entropy, the amount of Shannon information obtained with the evidence will be determined by Equation 7.9.

$$H_{posterior} = -\sum_{Z=0}^{n} P\left(V_p|P(Z)\right) \log_{10}\left(P\left(V_p|(Z)\right)\right) + P\left(\overline{V_p}|P(Z)\right) \log_{10}\left(P\left(\overline{V_p}|P(Z)\right)\right) \tag{7.9}$$

Based on the information provided in Section 7.2.3.1 and the Entropy calculation above, the



(a)



(b)

Figure 7.4: Binomial Distribution of $n = 300$ where (a) $P(S_1|\overline{V_P}) = 0.01$; (b) $P(S_1|\overline{V_P}) = 0.1$. In the $n = 300$ even at 10% error rate, the difference between two outcomes is very small, and larger values of Z would be required before the difference becomes useful.

effect of varying the prior probability listed in Table 7.1 and the sensor information listed in Table 7.2 and 7.3 are illustrated in the figures below. The figures demonstrated the variation in uncertainty by plotting the variation of Shannon entropy through the use of the line graph. The graphs also shows percentage reduction in the amount of uncertainty with varying numbers of sensor information, as the prior information changes through the use of the bar chart.

Figure 7.5 illustrates the effect of varying the prior probability, as listed in Table 7.1. It demonstrated that when additional information is provided, the amount of uncertainty (entropy) is reduced. The amount of uncertainty reduction of two sensors is considerably more than that of one single sensor.

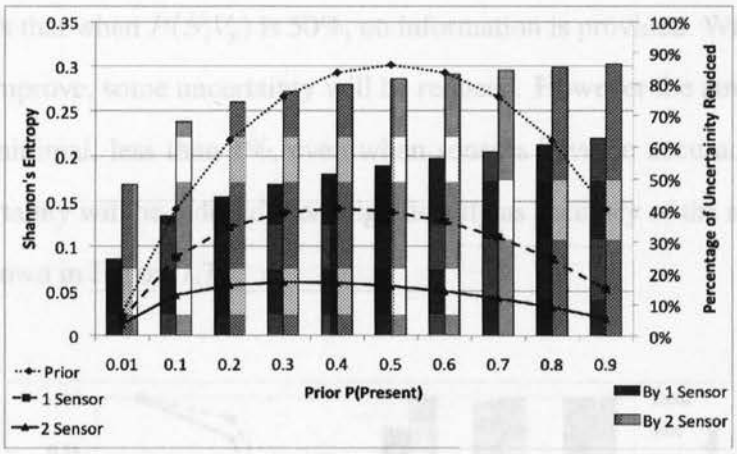


Figure 7.5: Variation of Entropy with Changes of the Prior Probability - The largest reduction of Shannon entropy between three scenarios can be observed when the Prior probability is at Maximum Entropy

When the Prior probability of the vehicle "was present" is low (0.1) but the sensor has the opposite belief, the amount of reduction is lower than one, when the sensor and prior have the same belief. As the prior probability moves toward the same belief as the sensors, the amount of entropy also reduces at a faster rate. The steeper decrease in the uncertainty is more visible when the two sensors are applied.

When the Prior probability is at maximum entropy, the effect of varying the sensor's error in the CPM Tables 7.2 and 7.3 is demonstrated in Figure 7.6.

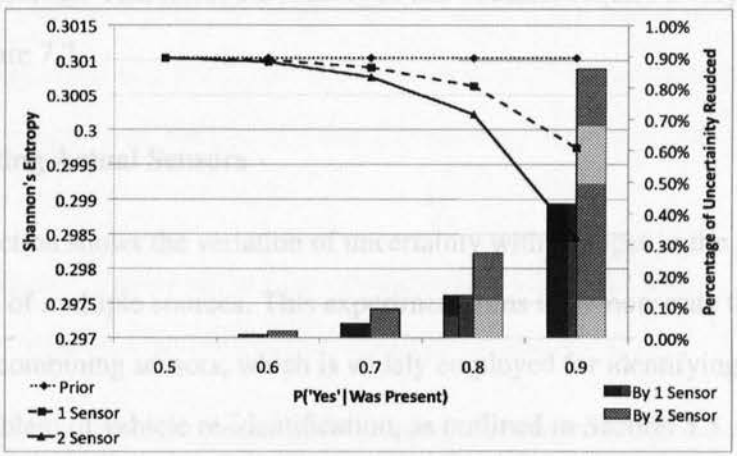


Figure 7.6: Variation of Entropy with Change of the Sensor's Accuracy - Even the sensor with accuracy of 90%, the amount of reduced entropy is still less than 1%

Figure 7.6 shows that when $P(S|V_p)$ is 50%, no information is provided. When the accuracy of the sensors improve, some uncertainty will be reduced. However the amount of reduced uncertainty is minimal, less than 1%, even when sensors have an accuracy of 90%. The amount of uncertainty will be reduced more significantly as accuracy of the sensors improves to 99.99% as shown in Figure 7.7.

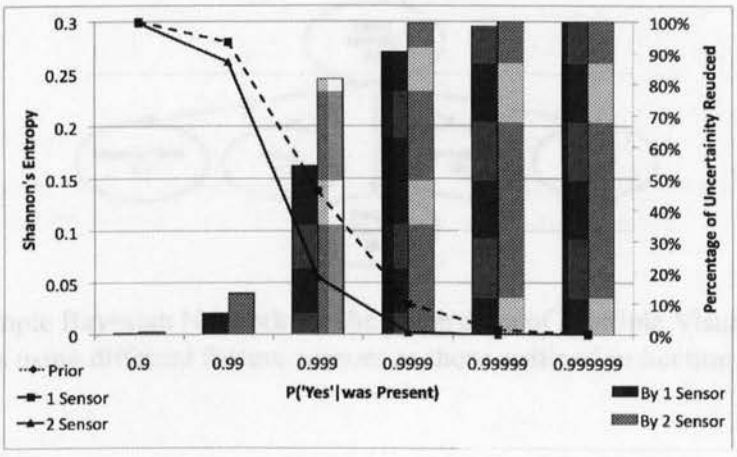


Figure 7.7: Variation of Entropy with Higher Accuracy Rates - The system can remove almost 100% of the uncertainty, only when two sensors with error rates are close to 0.01%.

In this scenario, there is an increased difficulty when inferencing compound events. Because the individual sensor errors accumulate as the number of positive identifications increase, the study aims to distinguish the very small differences between the probability distributions of the different outcomes. Therefore, the sensors in this scenario require a very small error rate as stated in Figure 7.7.

7.2.3.3 Adopting Actual Sensors

The previous section shows the variation of uncertainty with changes in the prior probability and the benefits of multiple sources. This experiment aims to demonstrate the improvement in accuracy, by combining sensors, which is widely employed for identifying a vehicle, when tackling the problem of vehicle re-identification, as outlined in Section 5.3.

All of the features outlined in Section 5.3 can be extracted from the vehicles within our test bed. Therefore, within a practical application, the input to the fusion framework outlined in

this section would be the output from each of the feature sensors, similar to those demonstrated in Chapter 6.

Within this theoretical investigation, the design of Bayesian Network for the problem of vehicle re-identification, could be conceptualised by Figure 7.8.

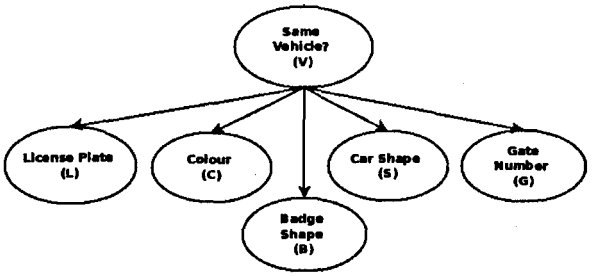


Figure 7.8: Simple Bayesian Network for the integration of Multiple Visual Surveillance Cues using different feature sensors as those outlined in Section 5.3.

For a given "probe" and "target", each sensor can give out the following information:

- License Plate - The number of character differences between the ANPR output of the probe against the target's NP string.
- Colour - Could be the RGB difference between the two vehicles measured by a distance metric, such as Euclidean.
- Logo Manufacturer Class - As addressed in Chapter 6, the output of the classifier for a probe image could be a confidence measure for each of the manufacturer. By knowing the target's logo class, the corresponding confidence measure of the target's class can be used as a metric.
- Vehicle Shape - Similar to the Logo manufacturer Class, the metric can be the corresponding confidence measure of the target vehicle shape class.
- Gate - Will output a binary argument to show if the target and probes are both used at the same gate or not.

According to Section 7.2.2.2, each child node will require a conditional probability matrix (CPM), which is used in the calculation of posterior probability. An example of a conditional probability matrix for each child node is given in the Tables below.

License Plate	$P(L\|V_r)$	$P(L\|\overline{V_r})$
0 Mistakes (L_1)	0.6	0.0
1 Mistakes (L_2)	0.2	0.2
2 Mistakes (L_3)	0.1	0.3
> 2 Mistakes (L_4)	0.1	0.5

Table 7.4: License Plate's Conditional Probability Matrix

Colour (Dist)	$P(C\|V_r)$	$P(C\|\overline{V_r})$
< 10 (C_1)	0.4	0.1
< 20 (C_2)	0.3	0.2
< 30 (C_3)	0.2	0.3
> 30 (C_3)	0.1	0.4

Table 7.5: Colour Difference's Conditional Probability Matrix

Vehicle Logo	$P(B\|V_r)$	$P(B\|\overline{V_r})$
>90%(B_1)	0.4	0.1
>75% (B_2)	0.2	0.4
>50%(B_3)	0.3	0.4
<50% (B_4)	0.1	0.1

Table 7.6: Vehicle Manufacture Logo's Conditional Probability Matrix

Body Shape	$P(S\|V_r)$	$P(S\|\overline{V_r})$
>90% (S_1)	0.3	0.2
>75% (S_2)	0.3	0.3
>50% (S_3)	0.2	0.3
<50% (S_4)	0.2	0.2

Table 7.7: Car Body Shape's Conditional Probability Matrix

Gate	$P(G\|V_r)$	$P(G\|\overline{V_r})$
Same (G_1)	0.78	0.5
Different (G_2)	0.22	0.5

Table 7.8: Gate's Conditional Probability Matrix

Experiments can be performed if the following results were acquired from each of the sensors:

1. Prior Probability: $P(V_r) = 0.5$ and $P(\overline{V_r})$
2. There is 1 mistake (L_2)
3. The colour difference is < 10 (C_1)

4. Probabilities of same logo is $> 75\%$ Other (B_2)

5. Probability of the same vehicle body shape $> 90\%$ (S_1);

6. The target vehicle uses different gates (G_2);

Based on the information above, the manual calculation is conducted in Appendix A, the probabilities distribution tables for each sensor, the posterior probability of same vehicle, $P(V|L_2, C_1, B_2, S_1, G_2) = 0.823$ and for 'not same vehicle' $P(\overline{V}|L_2, C_1, B_2, S_1, G_2) = 0.176$.

Even though the probability of the two vehicles being the same have improved when compared to the prior. There is a relatively small decrease in the entropy from 0.301 before fusion to 0.202 after fusion. This reveals that although entropy is a great measure of uncertainty, in some circumstances, the measure of performance should be further assessed by other appropriate strategy, such as when comparing fusion methods with non-fusion methods.

7.2.3.4 Expanding Inference Model

Figure 7.9 is developed with the base of the network depicted in Figure 7.8 to infer extra query of "Is the car park is full?".

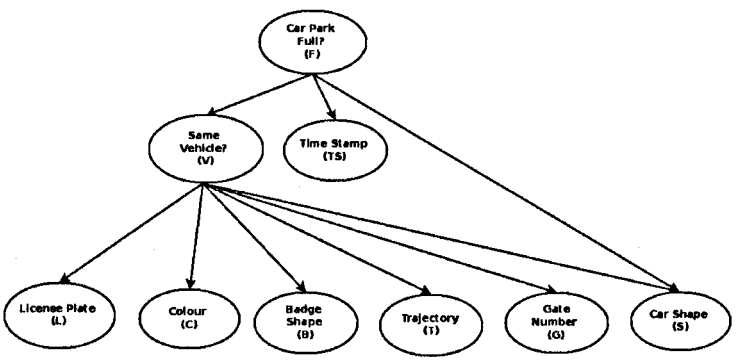


Figure 7.9: Simple Bayesian Network for integration of Multiple Visual Surveillance Cues.

This investigation assumes that visibility of the whole parking spaces is not available. The simplest measure for this inference is to count the number of cars in and out of the car

park during a period, guided by a maximum number of possible spaces with the following conditions:

1. All spaces are available
2. No overnight stays - Therefore, at the start of counting there is zero usage
3. Assuming all vehicles entering the car park are with the sole purpose of parking

In this project the entrance and exit are also used by delivery vehicles, and the number of possible parking places may be reduced without prior notice. In these situations, the fusing of information such as the time of observation and vehicle shape can help to infer this objective. Because:

1. An indicator of "the car park is full" might be if the vehicle only spends a short time (t_s) in the car park before exiting.
2. An indicator of a vehicle for delivery might be based on the vehicle shape and the time (t_d) spent in the car park, where the t_d time will be longer than t_s but smaller than average length of parking time.

7.2.4 Discussion

The framework is devised, jointly using Bayes and the extension to the Bayesian Network, by the introduction of graph theory. In order to adapt the Bayesian network for this test, a theoretical evaluation of Bayesian Network for reducing uncertainty was conducted.

The results of the experiment show that the model is capable of reducing uncertainty and the increase in the number of sensors would enhance the capability of reducing uncertainty. It also reveals that the prior probability plays major role in the effectiveness of the system at reducing uncertainty because the choice of the prior probability is important. The prior probability also relies on the accuracy of the sensors for reducing the system uncertainty.

The experimental process also shows how the network can improve the accuracy, as applied to the re-identification problem, by different vehicle sensors. The re-identification network

is further expanded to deduce other queries, thus demonstrating that the BN framework is very adaptable and is a suitable model for inferring a range of queries, meeting the aims as outlined in Section 7.1.

7.3 Statistic Parametric Fusion Methods: Dempster Shafer

7.3.1 Introduction

The discussion in Section 7.2.4 reveals the importance of both the prior probability and the sensor's conditional probabilities in the Bayesian model. An alternative model that eliminates the need for both of the probabilities in BN, but maintains the use of the subjective probability, is proposed below based on the Dempster Shafer Theory (DST).

7.3.2 Dempster-Shafer Theory

The Dempster-Shafer (DS) theory was introduced by Dempster (Dempster, 1967) and further developed by Shafer (Shafer, 1976). In the DS Theory, let $\Theta = \theta_1, .., \theta_n$ be a collection of mutually exclusive and exhaustive set of hypotheses to a given query, containing n elementary hypothesis, known as the frame of discernment. This is the same as the set of the hypothesis as outlined in Section 7.2.1.

There is also a set of general propositions developed with Boolean combinations, as the number of general propositions equates to (2^{n-1}). All possible states of the propositions are represented by Power Set (2^{Θ}), which contains all subsets of the elementary hypothesis and the empty set. An example is given by the Equation 7.10.

$$\begin{gathered} \Theta = \{\theta_1, \theta_2, \theta_3\} \\ 2^{\Theta} = \{\varnothing, \theta_1, \theta_2, \theta_3, \{\theta_1, \theta_2\}, \{\theta_1, \theta_3\}, \{\theta_2, \theta_3\}, \{\theta_1, \theta_2, \theta_3\}\} \end{gathered} \tag{7.10}$$

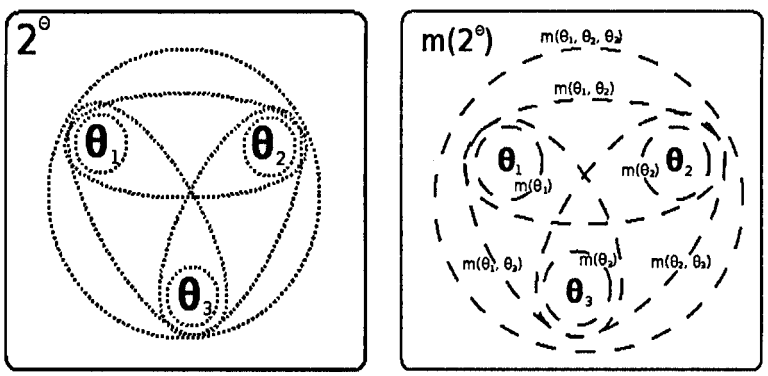
A Basic Belief Assignment (bba) is a function of Θ that assigns a mass of belief to each subset A of the power set 2^{Θ}, satisfying Equation 7.11.

$$0 \leq m(A) \leq 1$$

$$m(\emptyset) = 0 \tag{7.11}$$

$$\sum_{A \in 2^{\Theta} \neq \emptyset} m(A) = 1$$

The basic belief mass $m(A)$ represents a measure of the belief that is assigned to the subset $A \subseteq \Theta$, given the available evidence, and that cannot be committed to any strict subset of A. All of the assigned probabilities sum to unity, and there is no belief in the empty set ($\emptyset$). An illustration of the power set and the power set mass are given by Figure 7.10.



(a) (b)

Figure 7.10: Illustration of Power Set of Equation 7.10. (a) Power Set Representation, (b) Power Set Mass representation, showing all combinations of the Boolean combination in a 3 hypothesis scenario

To combine the different sources of information, a combination rule (the most widely used) is proposed by Dempster. The successful application of the Dempster's combination rule assumes that the different bba are independent pieces of evidence, and uses the orthogonal sum to combine the multiple belief structures. For two bba m_1 and m_2, the combination rule is as follows:

$$[m_1 \oplus m_2](\theta) = \frac{\sum_{A_i \cap B_j = \theta} m_1(A_i) m_2(B_j)}{1-K} \tag{7.12}$$

$$\text{where: } K = \sum_{A_k \cap B_m \neq \emptyset} m_1(A_k) m_2(B_m)$$

Dempster's rule of combination is both commutative and associative (Yager et al., 2008): these two properties mean that evidence could be combined ($\oplus$) iteratively by the Equation 7.12 and in any order of pair-wise methods, as depicted by Equation 7.13.

$$m_1 \oplus m_2 \oplus m_3 = (m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3) \tag{7.13}$$

Bayesian theory operates on the frame of discernment and offers a probability estimate of the hypothesis, the D-S approach operates on the power set and computes for each A with an evidential interval as described in Figure 7.11. The evidential interval is created by the probability mass functions and guided by two values of an uncertainty measure; firstly, the lower bound Belief measure, $Bel(A) = \sum_{A \subseteq B \neq \emptyset} m(B)$ represents the exact support for A, and secondly, the higher bound Plausibility measure $Pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$ represents the possible support for A.

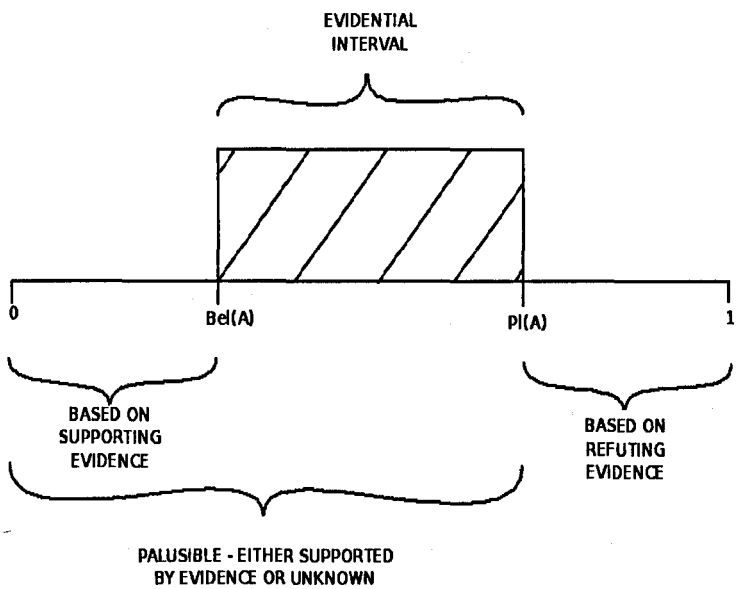


Figure 7.11: Evidential Interval and Uncertainty, showing the evidence interval which is bound by the lower bound Belief and the upper bound Support

For each hypothesis $\theta \in \Theta$, there are two measures of probability, rather than the single measure, provided by Bayesian analysis. The reduction is the evidential interval after combining sensors are used as a measure of uncertainty reduction in the DST model, when compared to the original sensor's evidential intervals.

7.3.3 Dempster-Shafer Surveillance Experiment

For the Surveillance scenario under investigation, this project adapts the Dempster-Shafer theory outlined in the previous section. The first DST model shows the design of the DST to reduce uncertainty under a case with two sensors. This is similar to the Bayesian experiment in Section 7.2.3.1. The second DST model illustrates an alternative design of the DS model with actual sensors, similar to the Bayesian design in Section 7.2.3.3.

7.3.3.1 Dempster-Shafer Model 1

The first formulation of DS model is illustrated by Figure 7.12. Unlike Section 7.2.3.1, in this case there is no prior probability. This is only true with the combination of the sensor information is used to calculate the probability of the hypothesis.

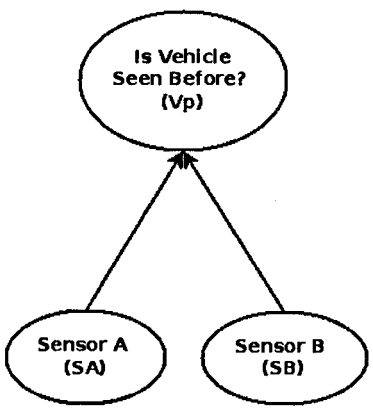


Figure 7.12: DS Inference: Has the Vehicle Been Present Before.

In this case, there are two possible hypotheses in their frames of discernment, and along with their corresponding power sets, they are illustrated in Equation 7.14 .

$$\Theta_{SA} = \{\theta_y, \theta_n\} \text{ and } \Theta_{SB} = \{\theta_y, \theta_n\}$$
$$2^{\Theta_{SA}} = \{\theta_y, \theta_n, \{\theta_y, \theta_n\}\} \text{ and } 2^{\Theta_{SB}} = \{\theta_y, \theta_n, \{\theta_y, \theta_n\}\} \tag{7.14}$$

I. Sensors' Agreement

In this case, given a probe vehicle and a target set, the sensors would output a probability that the probe vehicle exists in the target set. Therefore, the probability of the hypothesis θ_y, will be assigned to the $m(\theta_y)$. Since the sum of the power set mass is 1, the remaining mass is assigned to $m(\{\theta_y, \theta_n\})$ rather than $m(\theta_n)$. An example is illustrated in Equation 7.15.

$$m_1(2^{\Theta_{SA}}) = [0.8, 0, 0.2] \text{ and } m_2(2^{\Theta_{SB}}) = [0.9, 0, 0.1] \tag{7.15}$$

With the combination Equation 7.12, the mass after fusion and the corresponding belief and plausibility functions for each hypothesis are shown in Equation 7.19.

$$m_{12}(2^{\Theta}) = [0.98, 0, 0.02] \text{ and } \begin{bmatrix} Bel(\theta_i) & Pl(\theta_i) \\ 0.98 & 1 \\ 0 & 0.02 \\ 0.02 & 1 \end{bmatrix} \tag{7.16}$$

The results demonstrate that there is near total belief in proposition θ_y being 'true' and near total belief proposition θ_n being 'false'. In both cases, the evidential interval is minimal, compared with one before fusion. Therefore, after gaining information from the two sensors, the uncertainty in the hypothesis has been decreased.

Assuming the cases are independent, the reduction of uncertainty is also significant when the two sensors are unable to determine the results before fusion, as depicted in Equation 7.17.

$$m_1(2^{\Theta_{SA}}) = [0.5, 0, 0.5] \text{ and } m_2(2^{\Theta_{SB}}) = [0.5, 0, 0.5]$$

$$m_{12}(2^{\Theta}) = [0.75, 0, 0.25] \text{ and } \begin{bmatrix} Bel(\theta_i) & Pl(\theta_i) \\ 0.75 & 1 \\ 0 & 0.25 \\ 0.25 & 1 \end{bmatrix} \tag{7.17}$$

II. Sensors' Disagreement

In this case, given a probe vehicle and a target set, each sensor still outputs a probability that the probe vehicle exists in the target set. The sensors will, however, output two different results, as displayed in Equation 7.18.

$$m_1(2^{\Theta_{SA}}) = [0.8, 0, 0.2] \text{ and } m_2(2^{\Theta_{SB}}) = [0.1, 0, 0.9] \tag{7.18}$$

With the Dempster combination rule, the results are reflected in Equation

$$m_{12}(2^{\Theta}) = [0.82, 0, 0.18] \text{ and } \begin{bmatrix} Bel(\theta_i) & Pl(\theta_i) \\ 0.82 & 1 \\ 0 & 0.18 \\ 0.18 & 1 \end{bmatrix} \tag{7.19}$$

When two sensors disagree, although the uncertainty reduction is small, the benefits of sensor fusion still exists, due to the small additional information provided in the hypothesis $\{\theta_y, \theta_n\}$.

7.3.3.2 Dempster-Shafer Model 2

The model in Section 7.3.3.1 can be extended to identify the corresponding probe vehicle in a target set with the classification results output from the sensor illustrated in Figure 7.13. Given a vehicle target set, where the relevant features have been identified, as illustrated in

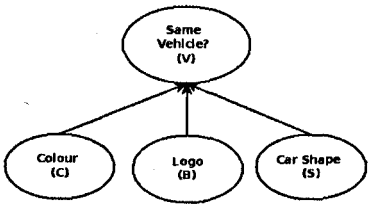


Figure 7.13: DS Inference: Identifying the Same Vehicle using the feature sensors, as outlined in Section 5.3

Table 7.9, for a given probe vehicle, each of the sensors could supply a probability for each of the possible categories of the feature. This is illustrated in sensor state tables below.

Vehicle ID	Colour	Badge	Shape
x_1	Blue	BMW	Car
x_2	Red	Audi	Van
x_3	Red	VW	Car

Table 7.9: Target Vehicle Knowledge Base

Colour Categories		
Red	Blue	Green
60%	80%	30%

Table 7.10: Colour Category Probability Table.

Logo Categories		
Audi	BWM	VW
40%	60%	70%

Table 7.11: Vehicle Logo Category Probability Table

Shape Categories		
Van	Car	Bus
67%	40%	8%

Table 7.12: Vehicle Shape Category Probability Table

With the information provided by the sensors, the probability of the probe vehicle (x_p) being the same for each member in the target set can be calculated. For example, the mass assignment of the sensors for the vehicle x_1 is:

- Colour Sensor - Given the x_1's colour with the 'red' and the probably of x_p with 'red', the mass assignment for the sensor is $m_C = \{0.8, 0, 0.2\}$
- Vehicle Logo Sensor - Given the x_1's logo with 'BMW' and the probably of x_p with 'BMW', the mass assignment for the sensor is $m_B = \{0.6, 0, 0.4\}$
- Vehicle Shape Sensor - Given the x_1's shape with 'Car' and the probably of x_p with 'BMW', the mass assignment for the sensor is $m_S = \{0.4, 0, 0.6\}$

With the information provided in the above list, the result of $m_C \oplus m_B \oplus m_S$ is given in Equation 7.20.

$$m_{C \oplus B \oplus S}(2^{\Theta}) = [0.952, 0, 0.048] \text{ and } \begin{bmatrix} Bel(\theta_i) & Pl(\theta_i) \\ 0.952 & 1 \\ 0 & 0.048 \\ 0.048 & 1 \end{bmatrix} \tag{7.20}$$

The above process is repeated for every member in the target set. In order to make a decision for the member with the closest similarity of the probe, only the the masses of θ_y, its corresponding $Bel(\theta_y)$ and $Pl(\theta_y)$ are taken for comparison. The results for every member in the target set are illustrated in Table 7.13.

Vehicle ID	$m_{C \oplus B \oplus S}(\theta_y)$	$Bel(\theta_y)$	$Pl(\theta_y)$	Evidential Interval
x_1	0.952	0.952	1	0.048
x_2	0.9215	0.9215	1	0.0785
x_3	0.9280	0.9280	1	0.0720

Table 7.13: Target Vehicle Knowledge Base

Based on the results in Table 7.13, the decision of the most likely member of the target set would be determined by one of the following rules:

1. $\max(Bel_i(\theta_y))$ – The outcome with the maximum belief function is chosen.

2. $\max(Pl_i(\theta_y))$ – The outcome with the maximum plausibility function is chosen.

3. Rational Rule – The outcome with the $\max(Bel_i(\theta_y))$ function and lowest evidential interval. As the evidence interval reflects the degree of uncertainty of the hypothesis, the shorter the length of the interval, the more certain the hypothesis is.

In this case, all hypotheses have the same plausibility, so the most suitable rules selected would involve belief function. The most likely hypothesis would be the vehicle with ID x_1.

7.3.4 Discussion

As mentioned above, the techniques of the Dempster Shafer Theory for uncertainty reduction do not need any prior probability. The DST method is selected for a case, able to fuse sensor outputs from the video surveillance scenario under investigation.

The theoretical deduction shows that uncertainty reduction can be achieved with information from multiple sources, even in situations when sensor results contradict. The measurement of uncertainty reduction in the DST model is conducted by measuring the Evidential Interval of the hypothesis. In addition, when DST is applied for a 'goal-oriented' environment, the 'true' hypothesis is chosen, based on its corresponding belief and plausibility, as expressed in Section 7.3.3.2.

In some situations, the model illustrated in Section 7.3.3.2 may be unable to find an outright winner because the evidential intervals could be very similar, such as the difference $|x_2 - x_3| = 0.0065$ in Table 7.13. Therefore, the robustness of the sensor's output could dramatically reduce the effectiveness of the DST fusion model, similar to the effect of Conditional Probability Matrix in the Bayesian model.

7.4 A Generalised Evaluation Approach

7.4.1 Introduction

Although the two decision models in Section 7.2 and 7.3 are all based on the subjective probability and appear to be different, DST is sometimes considered as a generalisation of the Bayesian inference, as outlined in previous section.

To state the generalisation, a modification to the cases in Section 7.3.3 is made. Instead of assigning the remaining sensor mass $(1 - m(\theta_y))$ to the hypothesis $m_1(\{\theta_y, \theta_n\})$, the hypothesis θ_n is assigned. When the two sensors are combined, as shown in Equation 7.21, they would have a total disbelief in the hypothesis $\{\theta_y, \theta_n\}$.

$$m_1(2^{\Theta_{SA}}) = [0.8, 0.2, 0] \text{ and } m_2(2^{\Theta_{SB}}) = [0.9, 0.1, 0]$$

$$m_{12}(2^{\Theta}) = [0.97, 0.03, 0] \text{ and } \begin{bmatrix} Bel(\theta_i) & Pl(\theta_i) \\ 0.97 & 0.97 \\ 0.03 & 0.03 \\ 0 & 1 \end{bmatrix} \tag{7.21}$$

In the case illustrated by Equation 7.21, all masses have been assigned to hypothesis in the frame of discernment (Θ). The DST model has been generalised into a special case of the Bayesian model.

Although the two models have a very close link, the DST model would still offer two additional uncertainty measures compared a single one in the Bayesian model. This restricts the effective evaluation of the uncertainty reduction capabilities of the two models. Due to this restriction, a generalised evaluation framework to accommodate the extra information offered by the DST is developed.

7.4.2 Kelly Betting Strategy

For a discrete set of outputs, Bayesian models can be evaluated by a Kelly Betting Criterion (Kelly, 1956). This consists of placing a nominal "stake" on each possible output in proportion of the odds estimated by the available observations. The pay-off can be defined by the prior (fair) odds. The doubling rate is proportional to mean information gain; the average amount of information provided by the observations can be inferred from the outcome of the betting strategy.

In section 7.2.3.1, it was discussed how the information provided by the posterior estimates are used to reduce the overall uncertainty for the state. The effectiveness (or accuracy) of any given Bayesian model can be evaluated by measuring the reduction of uncertainty relative to a prior model. This is equivalent to measuring the *side information* that the measurements provide for the system. These probabilistic measurements can be combined in many different ways, assuming independence or some model of co-dependency, for example, of parametric or explicit models. In all cases, however, the Bayesian model outputs an overall probability, per hypothesis, and the accuracy of any given model can be evaluated by measuring the expected log probability or entropy of the correct hypothesis.

$$\begin{aligned} H(X) &= \langle -\log p(x) \rangle \\ &\approx \frac{-1}{n} \sum_{i=1}^{n} \log p(x_i) \end{aligned} \tag{7.22}$$

The information gain is proportional to the mean doubling rate $\bar{W}$:

$$\begin{aligned} \bar{W} &= \log \left\langle \frac{1}{p(x)} \right\rangle \\ &= \langle -\log p(x) \rangle \\ &\approx \frac{-1}{n} \sum_{i=1}^{n} \log p(x_i) \end{aligned} \tag{7.23}$$

It is not straightforward to apply this Bayesian information-theoretic evaluation method to a DS model. That type of model contains two scalar quantities for each hypothesis; the belief and the plausibility contribute the reduction in uncertainty. Neither of these quantities directly relate to a Bayesian probability, so it is not clear how to apply the various information-theoretic results noted above. Nevertheless, the below describes a context where the log optimal doubling rate can be used to evaluate additional meaning expressed by the belief and plausibility provided by the DS model.

The important characteristic distinguishing the {belief, plausibility} pair from the {probability} singleton is that the former pair can be encoded by their difference, as an expression about the uncertainty of the estimate. If in a certain case the DS model provides a pair {0.05,

0.95}, it will be questioned how this can be distinguished from the pair {0.45, 0.55}, and can be evaluated against a Bayesian predicated estimate of 0.5.

It is proposed that the contribution of these extra indications provided by DS can be quantified by an appropriate generalization of Equation. 7.23. This represents the standard Bayesian evaluation with the expectation of the log posterior over a sample set, when the contribution of each element in the sample set is implicitly scaled to one.

The proposed generalization to accommodate the extra indication output of DS is to assign a weight α_i to each sample, subject to the constraint that $\langle \alpha_i \rangle = 1$. Thus, the evaluation metric is written as:

$$\bar{W} = \frac{-1}{n} \sum_{i=1}^{n} \alpha_i \log p(x_i) \tag{7.24}$$

If these weights are given random values, e.g. uniformly in the interval between 0 and 2, it can be shown that the measurements obtained from Equation. 7.24 are unchanged from those obtained from Equation. 7.23. However, doing so will set these weights up to be interpreted as a 'degree of confidence in the estimate'. For those samples considered as the model estimated as 'more accurate', the intention is to assign a larger scaling weight for those estimates with a greater degree of uncertainty.The intention is also to assign a smaller weight, thus fulfilling the overall constraint on the weights that their expectation is unity. This creates an opportunity to define an evaluation protocol used for both Bayesian and DS.

7.5 Experiments

7.5.1 Toy Example

The proposed evaluation procedure is applied to a toy example for an estimated model of a two-horse race with information provided by two sources including measurement of the Horses' attributes and measurement of the Jockeys' attributes. The evaluation metric is the

mean percentage winnings (or losses) per race, following a Kelly Betting strategy using the estimated model. This has a direct relation with the informative capacity of the model. Fundamentally, this percentage will depend on the relationship between three probabilistic models. The first model is the real (actual) probabilities that determine the outcome of the race. The second determines the bookmakers' odds, which will be used to calculate the pay-off after the outcome of each race. The third is the estimating model representing a subjective understanding of the likely outcome of each race with the two sources.

7.5.2 Kelly Betting with DS-Dependent Stake

The capability of the estimated model can be measured by treating it with a Kelly betting strategy, named as the log optimal strategy. The stake for each outcome is placed in proportion to the model prediction (estimated probability). Conventionally, the sum of these bets (i.e. the total stake) for each race can be fixed at an arbitrary quantity; the *total stake* for each race can be fixed at an arbitrary value, e.g. 1, and the accumulated winnings are logged. However, to accommodate the extra information provided by DS, this total stake is varied depending on the interval between the plausibility and support. Over the evaluation sample, the expected (mean) stake is constrained to be equal to the stake for the simple evaluation.

As a starting point, let all three models be identical. The expected outcome of both the fixed-stake and DS-dependent strategies is to "break-even", both with standard Kelly betting (fixed stake size) and the generalised Kelly strategy, where the total stake of each race is allowed to vary.

The above outcome is observed for any joint distribution between sources, i.e. for both correlated and anti-correlated distributions of *Horse* and *Jockey* measurements. However, the DS analysis does treat these two cases differently; divergent estimates between the two sources will result in a larger "unknown state", this states equates to the $\{\theta_y, \theta_n\}$ in Equation 7.14, rather than the case that they are in agreement.

7.5.3 Perturbation of the Prior Estimate

In the perturbation of the prior estimate, the prior information provided to the real model and bookmakers' model is the same, but the prior information from the sensors to the estimated model is perturbed from the real model, in order to simulate some imperfection in the available information. The perturbation takes the form of a percentage change to the estimated 'difference between means' that forms the model for generating each sensor measurement. The sign of the change is also generated randomly with equal probability.

Since the real odds and the bookmakers' odds are still identical, Kelly betting, using the estimated model, will always result in losses. However, more successful fusion strategies will reduce these losses, and the extent of the reduction can be used to evaluate the efficiency of the fusion strategy by using this generalised Kelly betting process, in which, a variable total stake is allowed for each case.

The DS fusion strategy provides a rationale for varying the total stake: when there is a large "unknown state", the total stake can be reduced; and conversely when there is a small "unknown state", a comparatively large total stake can be used. This strategy is repeated over 15,000 samples, at each level of estimated model perturbation, to compare the mean percentage loss from the DS strategy against the default (fixed total stake) alternative.

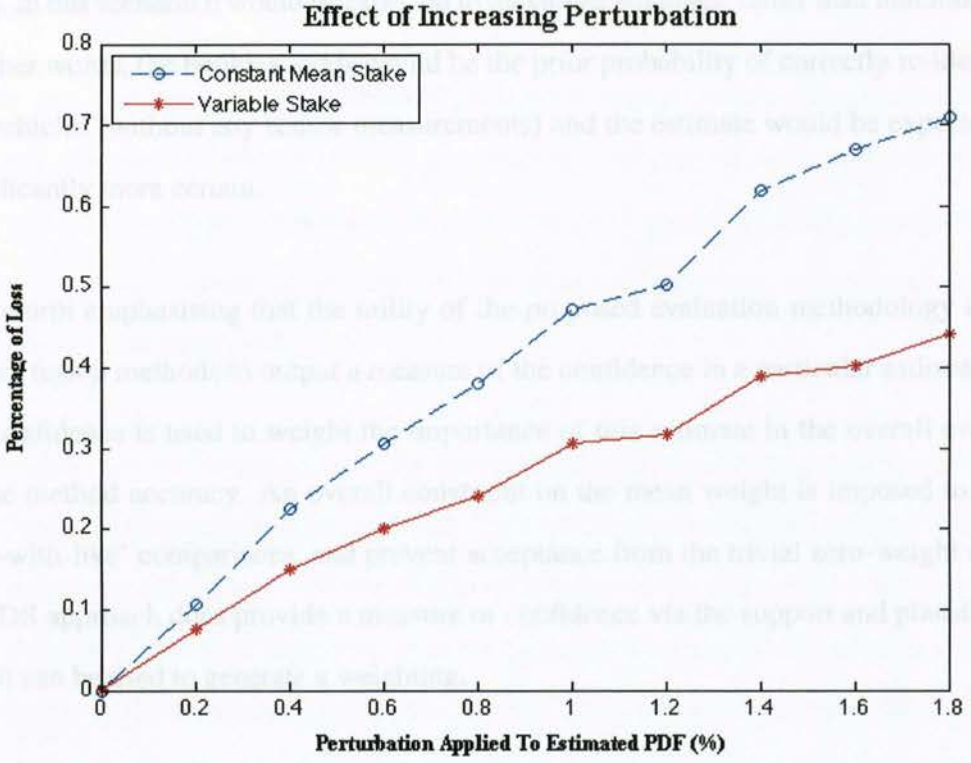


Figure 7.14: Effect on Variation of the Amount of Perturbation, which shows 33% reductions in the amount of loss when a variable stake strategy is employed.

The results of this simulation are plotted in Figure 7.14. It shows a clear advantage from the use of the DS fusion strategy. It is clear that approximately one-third of the adverse effects of the model perturbation are removed as a consequence of using DS outputs to determine the stake size. One explanation for this effect is that cases which source estimates from disagreement are more likely to have been significantly affected by the perturbation, so the consequential reduction in the total stake reduces the effect of more substantial inaccuracy in probability.

7.5.4 Application to Surveillance Fusion

The above methods can be adapted to simulate the information fusion process demanded for visual surveillance scenarios. One of the key capabilities is the re-identification of vehicles from a pair of vehicle observations, as outlined in Section 7.3.3.2 and 7.2.3.3. The above methodology can be used to evaluate the benefit of the fusion method that exploits the DS outputs, using probabilistic sensor measurements of a vehicle's type, make, colour and number

plate. In this scenario it would be expected to maximize winnings, rather than minimize losses. In other words, the bookies' odds would be the prior probability of correctly re-identifying the vehicles (without any sensor measurements) and the estimate would be expected to be significantly more certain.

It is worth emphasising that the utility of the proposed evaluation methodology is that it allows fusion methods to output a measure of the confidence in a particular estimate. Then, the confidence is used to weight the importance of this estimate in the overall evaluation of the method accuracy. An overall constraint on the mean weight is imposed to enforce 'like-with-like' comparisons, and prevent acceptance from the trivial zero-weight solution. The DS approach does provide a measure of confidence via the support and plausibility so that it can be used to generate a weighting.

It is important to examine the significance and utility of the proposal in context of the evaluation in automated video analytic systems. A frequent criticism of these systems is whether they are unable to indicate the point that they are 'not sure'. Hence, this proposal fits well into that context. By requiring that a system also outputs a weight to calculate the evaluation, the indication of certainty is incorporated and, in a straightforward manner, consistent with standard information-theoretic evaluation of 'side-information'. Furthermore, the proposed strategy is identical to the standard information-theoretic evaluation, in the limiting cases, in that, each weight is constant and equal.

Nevertheless, there are still several tasks outstanding. Since there are various ways that the DS output could be transformed into a single weight, it is not yet clear which of these would be the most appropriate. One specific aspect is the mechanism to enforce a fixed mean weight over the test set ensemble. Another task is a more comprehensive evaluation over the range of possible perturbations to verify whether the proposed approach works in this range. It may be possible to obtain some theoretical results for this general case as well.

7.6 Summary

Under this investigation, two statistical parametric fusion models using multiple vehicle features , are used to improve estimate precision in achieving visual surveillance objectives. The models have shown extendibility to infer a range of queries by fusing a range of incommensurate visual features of the vehicle rather than a singular one, which is restrictive and suboptimal in the final result.

An evaluation metric, based on the Kelly betting strategy for the direct comparison of Bayesian and Dempster Shafer models, is produced. This metric accommodates the extra information provided by the DS model.It was shown that by using a simple example of 3 broad conditions, the DS model provides an improvement in the mean log winnings. This is a fundamental information metric of the standard Bayesian evaluation and is proportional to the side-information provided by the observation. It also described how this simple example can be adapted for the surveillance scenario. Furthermore, it is also able to be applied for the general case of fusion problems under investigation.

Chapter 8

Conclusions

8.1 Introduction

With the principle aim of demising a feasible framework that could reduce the uncertainty within a video analytic system in mind, this thesis has made the following knowledge contributions:

- Demonstrated the benefits of using new higher resolution datasets for the problem of pedestrian re-identification under various scenarios including occlusion.
- Devised an approach, using the Fisher Discriminative classifier and decision fusion techniques, to identify and classify logos. The approach relieves the reliance of needing to locate dominant features, thus allowing it to classify vehicle logos on both the front and rear of a vehicle.
- Theoretically investigated the feasibility of two fusion frameworks, based on the probabilistic Bayesian and Evidential Dempster-Shafer techniques, to get the inference of multiple objectives and to reduce uncertainty by combining multiple techniques.
- Theoretically evaluated the developed evaluation framework based on the Kelly Betting Strategy to effectively accommodate the additional information offered by the Dempster-Shafer ,allowing it to be compared with the single probabilistic output from a Bayesian framework.

To support the above contribution, an in-depth analysis of the aims of video analytic systems was conducted which enabled the creation of the experimental test bed. The data and supporting redundancy reduction systems were used to create two videos used during international TV programmes. An understanding of the types of uncertainty and where it might reside within our experimental test-bed was also investigated. These findings help the creation of the fusion frameworks.

The following section will summarise the findings within each chapter of this thesis, where an appropriate comparison to literature is given, along with a future direction of research following the discoveries. As surveillance video contains personal data, some ethical considerations should follow. The chapter concludes by analysing the limitations in the proposed methods, with details of future work that should be carried out.

8.2 Summary

As mentioned above, the main aim of this thesis is to investigate the feasibility of devising a framework that could reduce the uncertainty of a video analytic system, so as to improve the accuracy of the outcome of video surveillance challenges, aided by various video analytic approaches. A suitable candidate to locate the target is the Information Fusion technique. The literature review conducted in Chapter 2 demonstrated that, though they are limited compared to other research fields, fusion techniques have been adopted in various video and image analytics. Reviews of the Information Fusion models allowed the categorisation of the techniques into two main sections:

1. Low Level – In video analytics, the inputs are the descriptions of the object. The descriptor can be thought of as a sensor that outputs an object's feature. Therefore, in low level fusion, a number of an object's features are combined into a single feature that is used as the input to the system.

2. High Level – Combined decisions become the outputs from a video analytic system to achieve a more accurate result. There are a range of techniques available:

   (a) Reasoning – Exemplified by a range of voting techniques. Generally, those

techniques require the output to be of the same format, in order to conduct the decision making process. e.g. A decision whether the hypothesis is true or not, and whether the hypothesis with the most votes is the winner or not.

(b) Evidence – These techniques work with the subjective probability of the outcome to reduce the uncertainty of the outcome. Unlike reasoning techniques, the outcome is not necessarily in the same format, but the output must be with a format that can be interpreted by the fusion model.

Compared to some of the Information Fusion review papers, the techniques reviewed in Chapter 2 are only a snippet of what it is widely available. This is because the use of information fusion techniques have mainly concentrated on physical sensors, rather the different information that can be extracted from a single physical sensor. Therefore, the research conducted in this thesis would be a useful illustration of the links between information fusion techniques and traditional video surveillance techniques. This link is intended to allow a number of researchers to adopt these techniques in the future to increase its popularity within the video analytics field.

In Chapter 3 an investigation was conducted on the advantageous use of high resolution images, to improve accuracy in pedestrian re-identification challenges. As the literature on the use of higher resolution data to tackle video surveillance challenges was very limited, their direct benefits were never explored. By combining the current state-of-the-art techniques for varying resolutions of the input data in Chapter 3, a direct comparison of the benefits were explored. The results illustrated that using higher resolution images and smaller overlapping patches enables the capturing of finer details of images and is beneficial to improving accuracy. In addition, the accuracy can also be improved, depending on the viewpoints and whether a fused feature vector is involved.

As occlusion is also a common factor in many pedestrian re-identification datasets. Chapter 3 also explored the challenges surrounding the treatment of occlusions. In a lot of literature, the occlusions are treated as a source of uncertainty which contribute to the effectiveness of their proposed systems and the best treatment for occlusion data is not explicitly explored. Through

the simulation of various scenarios in Chapter 3, it was concluded that the best treatment for occlusion was to deploy a system which can detect its existence within the dataset before deploying the most appropriate classifier.

To support the aim of developing an extendible model that allows inferences of a range of surveillance objectives, a suitable dataset needs to be selected. In Chapter 4 a discussion on the different video surveillance objectives and an examination of publicly available video surveillance datasets were conducted. The discussion concludes that the video surveillance challenges mainly concentrate on two different objects of interest: People and Vehicles. The currently available datasets are mainly used to solve particular surveillance challenges, as they only capture snippets of the monitored environments. Based on the discussions, the proposed dataset is designed for the use of high definition cameras, which has increased in popularity for newer CCTV networks.

To extract key information within the captured data and to reduce the redundant data, a simple background subtraction technique was used in Chapter 4. The implemented techniques were successful during offline processing, as false-positives can be removed manually. If the same techniques were applied to online processing, however, this may cause issues. The volume of these false-positive rates also increases when the lighting intensity decreases. The conducted experiments have re-confirmed the issues related to implemented techniques, as those demonstrated in the literature. For the purpose of this thesis, the technique was combined with simple tracking algorithms to create some demonstration videos, which were used in two international video programs to demonstrates how pedestrian tracking can be performed.

As the main aim is to reduce the uncertainty within the video surveillance systems, an analysis of the factors that contribute to the uncertainty in a video surveillance analytic system was conducted in Chapter 5. The chapter illustrated that uncertainties are introduced as soon as physical signals are converted to the digital signals by the video cameras. These uncertainties are propagated through the video analytic pipeline, and the characteristics of each of the processing blocks will introduce their own uncertainties. Chapter 5 also included a detailed

discussion of the uncertainties in video analytic systems, with different features related to vehicles. The analysis conducted in Chapter 5 shows that there is a variety of uncertainty in the video analytic pipeline that effects the accuracy of the system, therefore, an uncertainty reduction model should be developed. The review process also pointed to the fact that within the video analytic research community, the evaluation process is very goal oriented. Therefore, when comparing techniques, the results might be subjectively dependant on the evaluation techniques used. As such, an objective evaluation metric, based on a measure of uncertainty reduction, might allow a more systematic approach to evaluate and compare different methods.

The uses of higher resolution data and feature fusion are further explored in Chapter 6, when they are combined with high level reasoning techniques to develop a new technique for the localisation and classification of vehicle manufacturer logos. The techniques assessed which patches in a Region of Interest are likely to be the logo patches. Only the logo patches are passed to a multi-class classifier to group the logo into its respective category. As there might be multiple logo patches, the determination is conducted with a voting technique, and a subsequent weighted voting process is also implemented when an outright winner cannot be found. A multi-class classifier, rather than a one-against-all classifier, is used in order to avoid the need of training different classifiers for different logo manufacturer categories.

Although the developed system still requires a reference point to validate the search co-ordinates of the logo, it eliminates the need of a dominant feature, such as the grille. Compared to the other proposed techniques, this allows a degree of freedom in the localisation of badges that are not located on the grille. In addition, it allows the classification of logos from both the front and rear of the vehicle. The illustrated process would also allow future researchers, a standard processing pipeline to effectively evaluate different classification techniques.

Chapter 6 also demonstrated that the choice of features involved is important in the fusion process. If two representations of the same feature space are used, the two features might not contribute enough independent information. As a result, no extra advantages could be gained. However, it is believed that feature fusion would still improve the accuracy of analytic

systems when the correct features are selected, as demonstrated in Chapter 3.

Based on the investigation conducted in the previous chapters, the output from different visual analytical systems can be a probability in relation to the accuracy of the results. A fusion framework, based on probability or evidence is developed in Chapter 7. The literature review in Chapter 2 has suggested that evidential reasoning techniques have been adopted, in some challenges, to fuse visual sensors with other physical sensors. The fusion of different descriptors from the same physical visual sensor, however, has not fully been investigated. For this reason, the fusion framework in Chapter 7 adapted two popular evidential reasoning techniques from other research communities, for the challenge of uncertainty reduction in video analytics, using a single physical sensor.

The developed decision fusion framework is based on Bayesian and Dempster-Shafer Theories, where the DST removes some restrictions on the need for prior probability in the Bayesian scenario. Through a range of theoretical modelling and experiments, both theories have shown their ability to reduce the uncertainty with aid from a range of information sources, even when different independent sensors disagree. It is difficult, however, to evaluate which model is more efficient at reducing uncertainty because each hypothesis Bayesian outputs a singular probability measure and DST outputs an additional uncertainty measure in the form of belief and plausibility for each hypothesis.

Although DST can be generalised to a special case of the Bayesian model, the standard uncertainty metrics could not accommodate the extra uncertainty measure offered by DST. Thus, it becomes necessary to develop an evaluation strategy that could accommodate the extra information, based on the Kelly Betting system. A simple experiment for evaluating DST and the Bayesian model shows that, under certain conditions, the DST model is more efficient at reducing uncertainty than the Bayesian model. In addition, it is believed that the developed evaluation model can be extended to evaluate other fusion models.

As mentioned before, the proposed decision fusion methods were adopted from other research domains and have been adapted to accommodate the information that might be supplied from a traditional video surveillance system. The framework proposed will offer an alternative solution to the help future researchers in the video surveillance domain. The framework can also be used to construct new video surveillance systems, as they have exhibited the capabilities of extending models to incorporate inferencing in a range of different queries.

8.2.1 Ethical Considerations

Due to the nature of the monitored environment, a range of personal information is gathered, such as the Number Plate of a car and daily behaviours of vehicle owners. For a successful proceeding of this project, a range of ethical implications were evaluated and presented to the Kingston University Ethical Committee for approval.

A range of controls were also constructed to protect the privacy of the personal information and the data integrity. These controls include:

1. Notices that advised users of the car park that they are under monitoring, and informing them that, if they wanted, they can be removed from the captured data.
2. Secure storage of the captured footage in separate network data storage facilities that are put into a secure room, where only authorised individuals can access to the data and the room.
3. Any publication of any personal data, such as license plates number, will either be blocked or prior consent is acquired.

8.3 Future Work

There is significant potential for improving and extending the strategies proposed in this thesis. In this section, the limitations within each of the chapters are explored and the future research direction is outlined.

In order to investigate the effectiveness of the higher resolution data and occlusion, as proposed in Chapter 3, some of the data was synthetically created, therefore, future work might be needed to validate the findings using non-modified data. The investigation only looked at one re-identification technique, consequently, a comparison between different techniques is required to identify the best re-identification method in evaluating the use of high resolution information. The chapter also highlighted issues related to the domain specificity and that the effort needed to identify a globally effective training set, capable of avoiding these issues, should be investigated. The incorporation of occlusion detection systems is needed to effectively choose the most suitable model for the re-identification process.

As stated in Chapter 4 the proposed background subtraction techniques have various limitations, as a result, this component within the overall proposed process pipeline can only be used for offline processing. Although BS is an active research area, the researchers are often restricted in improving the performance to one or two limited areas. For this reason, future research should investigate a framework that combines the various techniques to tackle all challenges as a whole. To assist in this investigation, the data created in this chapter could be used, as their meta-data is available and can assist in the manual partitioning and categorisation of the data into various categories, such as illumination and weather. This partitioned data can also help to test the effectiveness of any proposed surveillance system under various known conditions.

To improve the system performance for the vehicle manufacturer classification, it is important to investigate the application of additional features which provide uncorrelated indications of a logo's identity. As the current size of the Region of Interest has been predefined to accommodate the expected range of scales, the method is tolerant to changes in scale, limiting its effectiveness to a mid-range view of the vehicle. This means that a further improvement could be to scale the ROI by the license plate size. However making the proposed method scale invariant will require a calibration of the camera and its measured space, thus restricting the flexibility of the approach.

The Local Fisher classification methods, proposed in Chapter 6 are computationally expensive when searching for the optimal subspace to conduct the classification, thus restricting its use for real-time applications. However, if an optimal space can be found to satisfy the various changes in the monitored environment, then this method can be adopted. The experiments conducted use a relatively small test and training sample, a great quantity of testing data would be necessary for further evaluating the potential for learning. In addition, the experiments only used 5 categorises, therefore the potential classification techniques need to be further explored. To test the effectiveness of the system under various environmental conditions, a partitioned dataset could be used.

The features used in Chapter 6 also failed to demonstrate the advantage of feature fusions, as the features chosen represented the same feature space. In order to build on this, a different edge detector such as those mentioned in Maini and Aggarwal (2009) can be used to represent the HOG or pHOG vectors. Alternatively, instead of fusing the feature, a classification result using each of the features can be applied as the input to a probabilistic fusion model, such as those created in Chapter 7. The developed decision fusion model and the theoretical investigation in Chapter 7 have both demonstrated that they are significantly promising in reducing uncertainty for a range of visual surveillance objectives. The investigation has only been theoretical, therefore further investigation, which combines actual results from different analytic systems, would be needed. At the moment, the DST model has indicated potential for reducing uncertainty for particular challenges. This needs to be further developed in scalability, in order to infer a range of different queries similar to those that have been developed for the Bayesian model.

Currently only the Dempster combination rules have been used. As Sentz and Ferson (2002) stated that there is a range of different combination rules available, an investigation on determining the most useful rule for the framework for this study would also be helpful.

Although the evaluation framework has been analysed theoretically and have shown the ability to create an environment for evaluating the DST and Bayesian, further investigation

on the evaluation of other techniques is still needed to examine the robustness of the fusion frameworks.

Appendix A

Section 7.2.3.3 Bayesian Network Calculations

A.1 Problem Information

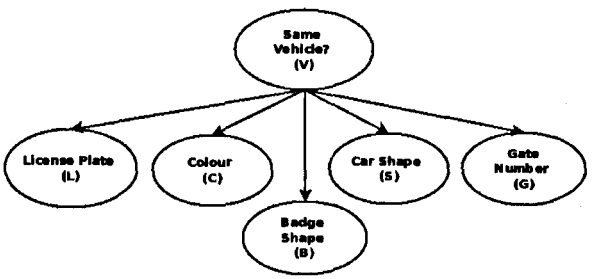


Figure A.1: Simple Bayesian Network for integration of multiple visual surveillance cues.

For a given single "probe" and a single "target", each sensor can give out the following information:

- License Plate - the number of character differences between the ANPR output of the probe against the target's LP string.
- Colour - could be the RGB difference between the two vehicles measured with a distance metric, such as Euclidean.
- Logo Manufactures Class - as discussed in Chapter 6, the output of the classifier of a probe image could be a confident measure for each of the manufactures. By knowing

the target's logo class, the corresponding confident measure of the target's class can be used as a metric.

- Vehicle Shape - the metric can be the corresponding confident measure of the target's vehicle class.
- Gate - will output a binary argument to show if the target and probe's both are used at the same gate or not.

License Plate	$P(L\|V_r)$	$P(L\|\overline{V_r})$
0 Mistakes (L_1)	0.6	0.0
1 Mistakes (L_2)	0.2	0.2
2 Mistakes (L_3)	0.1	0.3
> 2 Mistakes (L_4)	0.1	0.5

Table A.1: License Plate CPM

Colour (Dist)	$P(C\|V_r)$	$P(C\|\overline{V_r})$
< 10 (C_1)	0.4	0.1
< 20 (C_2)	0.3	0.2
< 30 (C_3)	0.2	0.3
> 30 (C_3)	0.1	0.4

Table A.2: Colour Difference CPM

Vehicle Logo	$P(B\|V_r)$	$P(B\|\overline{V_r})$
>90%(B_1)	0.4	0.1
>75% (B_2)	0.2	0.4
>50%(B_3)	0.3	0.4
<50% (B_4)	0.1	0.1

Table A.3: Vehicle Manufacture Logo CPM

Body Shape	$P(S\|V_r)$	$P(S\|\overline{V_r})$
>90% (S_1)	0.3	0.2
>75% (S_2)	0.3	0.3
>50% (S_3)	0.2	0.3
<50% (S_4)	0.2	0.2

Table A.4: Car Body Shape CPM

Gate	$P(G\|V_r)$	$P(G\|\overline{V_r})$
Same (G_1)	0.78	0.5
Different (G_2)	0.22	0.5

Table A.5: Gate CPM

Experiments can be performed if the following results were acquired from each of the sensors:

1. Prior Probability: $P(V_r) = 0.5$ and $P(\overline{V_r})$

2. There is 1 mistake (L_2)
3. The colour difference is < 10 (C_1)
4. Probabilities of same logo is $> 75\%$ Other (B_2)
5. Probability of the same vehicle body shape $> 90\%$ (S_1);
6. The target vehicle used different gates (G_2);

A.2 Calculating Probability of Same

Since the $P(L_2|V_r) = P(L_2|\overline{V_r})$ no additional information was provided to change the prior belief's, as shown in Equation equ:app:SAME.

$$\begin{aligned} P(V_r|L_2) &= \frac{P(L_2|V_r)P(V_r)}{P(L_2)} \\ &= \frac{P(L_2|V_r)P(V_r)}{P(L_2|V_r)P(V_r) + P(L_2|\overline{V_r})P(\overline{V_r})} \\ &= \frac{0.2 * 0.5}{0.2 * 0.5 + 0.2 * 0.5} = 0.5 \end{aligned} \tag{A.1}$$

$$\begin{aligned} P(V_r|L_2, C_1) &= \frac{P(C_1|V_r)P(V_r|L_2)}{P(C_1|V_r)P(V_r|L_2) + P(C_1|\overline{V_r})(1 - P(V_r|L_2))} \\ &= \frac{0.4 * 0.5}{0.4 * 0.5 + 0.1 * (1 - 0.5)} = 0.8 \end{aligned} \tag{A.2}$$

$$\begin{aligned} P(V_r|L_2, C_1, B_2) &= \frac{P(B_2|V_r)P(V_r|L_2, C_1)}{P(B_2|V_r)P(V_r|L_2, C_1) + P(B_2|\overline{V_r})(1 - P(V_r|L_2, C_1))} \\ &= \frac{0.2 * 0.8}{0.2 * 0.8 + 0.4 * (1 - 0.8)} = 0.667 \end{aligned} \tag{A.3}$$

$$\begin{aligned} P(V_r|L_2, C_1, B_2, S_1) &= \frac{P(S_1|V_r)P(V_r|L_2, C_1, B_2)}{P(S_1|V_r)P(V_r|L_2, C_1, B_2) + P(S_1|\overline{V_r})(1 - P(V_r|L_2, C_1, B_2))} \\ &= \frac{0.3 * 0.667}{0.3 * 0.667 + 0.2 * (1 - 0.667)} = 0.75 \end{aligned} \tag{A.4}$$

$$P(V_r|L_2,C_1,B_2,S_1,G_1) = \frac{P(G_1|V_r)P(V_r|L_2,C_1,B_2,S_1)}{P(G_1|V_r)P(V_r|L_2,C_1,B_2,S_1) + P(G_1|\overline{V_r})(1-P(V_r|L_2,C_1,B_2,S_1))}$$
$$= \frac{0.78*0.75}{0.78*0.75+0.5*(1-0.75)} = 0.824 \tag{A.5}$$

A.3 Calculating Probability of Different

Since the $P(L_2|V_r) = P(L_2|\overline{V_r})$ no additional information was provided to change the prior belief's, as shown in Equation A.6.

$$P(\overline{V_r}|L_2) = \frac{P(L_2|\overline{V_r})P(\overline{V_r})}{P(L_2)}$$
$$= \frac{P(L_2|V_r)P(V_r)}{P(L_2|V_r)P(V_r)+P(L_2|\overline{V_r})P(\overline{V_r})} \tag{A.6}$$
$$= \frac{0.2*0.5}{0.2*0.5+0.2*0.5} = 0.5$$

$$P(\overline{V_r}|L_2,C_1) = \frac{P(C_1|\overline{V_r})P(\overline{V_r}|L_2)}{P(C_1|V_r)(1-P(\overline{V_r}|L_2))+P(C_1|\overline{V_r})P(\overline{V_r}|L_2)}$$
$$= \frac{0.1*0.5}{0.4*(1-0.5)+0.1*0.5} = 0.2 \tag{A.7}$$

$$P(\overline{V_r}|L_2,C_1,B_2) = \frac{P(B_2|\overline{V_r})P(\overline{V_r}|L_2,C_1)}{P(B_2|V_r)(1-P(\overline{V_r}|L_2,C_1))+P(B_2|\overline{V_r})P(\overline{V_r}|L_2,C_1)}$$
$$= \frac{0.4*0.2}{0.2*(1-0.2)+0.4*0.2} = 0.333 \tag{A.8}$$

$$P(\overline{V_r}|L_2,C_1,B_2,S_1) = \frac{P(S_1|\overline{V_r})P(\overline{V_r}|L_2,C_1,B_2)}{(1-P(S_1|V_r)P(V_r|L_2,C_1,B_2))+P(S_1|\overline{V_r})P(\overline{V_r}|L_2,C_1,B_2)}$$
$$= \frac{0.2*0.333}{0.3*(1-0.333)+0.2*0.333} = 0.25 \tag{A.9}$$

$$P(\overline{V_r}|L_2, C_1, B_2, S_1, G_1) = \frac{P(G_1|V_r)P(V_r|L_2, C_1, B_2, S_1)}{P(G_1|V_r)(1 - P(\overline{V_r}|L_2, C_1, B_2, S_1)) + P(G_1|\overline{V_r})P(\overline{V_r}|L_2, C_1, B_2, S_1))}$$
$$= \frac{0.5 * 0.25}{0.78 * (1 - 0.25) + 0.5 * 0.25} = 0.176 \tag{A.10}$$

References

Alahi, A., Vandergheynst, P., Bierlaire, M., and Kunt, M. (2010). Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640. 14

Atrey, P. K., Kankanhalli, M. S., and Jain, R. (2006). Information assimilation framework for event detection in multimedia surveillance systems. *Multimedia systems*, 12(3):239–253. 19

Barnich, O. and Van Droogenbroeck, M. (2011). Vibe: A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on*, 20(6):1709–1724. 54

Basir, O. and Yuan, X. (2007). Engine fault diagnosis based on multi-sensor information fusion using dempster–shafer evidence theory. *Information Fusion*, 8(4):379–386. 16

Bayes, T., Bunn, D. W., Raiffa, H., Schlaifer, R., and Von Winterfeldt, D. (1984). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53. 94

Benezeth, Y., Jodoin, P.-M., Emile, B., Laurent, H., and Rosenberger, C. (2008). Review and evaluation of commonly-implemented background subtraction algorithms. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE. 54

Bevington, P. R. and Robinson, D. K. (1969). *Data reduction and error analysis for the physical sciences*, volume 2. McGraw-Hill New York. 63

Bhanu, B. and Kafai, M. (2012). Dynamic Bayesian Networks For Vehicle Classification in Video. WO Patent 2,012,159,109. 79

Bishop, C. M. (1996). *Neural Networks for Pattern Recognition.* Oxford University Press, USA. 80

Blasch, E. P. and Plano, S. (2002). Jdl level 5 fusion model: user refinement issues and applications in group tracking. In *AeroSense 2002*, pages 270–279. International Society for Optics and Photonics. 9

Bloch, I. (1994). Information combination operators for data fusion: a comparative review with classification. In *Satellite Remote Sensing*, pages 148–159. International Society for Optics and Photonics. 16

Blum, R., Xue, Z., and Zhang, Z. (2005). An overview of image fusion. *Multi-Sensor Image Fusion and Its Applications, Signal and Image Processing*, pages 1–36. 9, 10

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM. 79, 83

Boström, H., Andler, S. F., Brohede, M., Johansson, R., Karlsson, A., van Laere, J., Niklasson, L., Nilsson, M., Persson, A., and Ziemke, T. (2007). *On the definition of information fusion as a field of research.* Institutionen för kommunikation och information. 8

Breitenstein, M., Reichlin, F., Leibe, B., Koller-Meier, E., and Van Gool, L. (2009). Robust tracking-by-detection using a detector confidence particle filter. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1515–1522. IEEE. 25

Brutzer, S., Hoferlin, B., and Heidemann, G. (2011). Evaluation of background subtraction techniques for video surveillance. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1937–1944. IEEE. 54

Burkhard, T., Minich, A., and Li, C. (2011). Vehicle Logo Recognition and Classification: Feature Descriptors vs. Shape Descriptors. *Stanford University, EE368 Final Project.* 79

Candes, E., Demanet, L., Donoho, D., and Ying, L. (2006). Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899. 79

Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, 2013. 16, 19

Chen, C., Yao, Y., Page, D., Abidi, B., Koschan, A., and Abidi, M. (2008). Comparison of image compression methods using objective measures towards machine recognition. *International Journal of Information Analysis and Processing, IJIAP*, 1(2):63–74. 61

Choudhury, T., Rehg, J. M., Pavlovic, V., and Pentland, A. (2002). Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 3, pages 789–794. IEEE. 20

Comaniciu, D., Ramesh, V., and Meer, P. (2000). Real-time tracking of non-rigid objects using mean shift. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 142–149. IEEE. 73

Constantinos, S., Pattichis, M. S., and Micheli-Tzanakou, E. (2001). Medical imaging fusion applications: An overview. In *Signals, Systems and Computers, 2001. Conference Record of the Thirty-Fifth Asilomar Conference on*, volume 2, pages 1263–1267. IEEE. 10

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1:886–893. 13

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176. 68

Das, A., Manyam, O. K., and Tapaswi, M. (2008). Multi-feature audio-visual person recognition. In *Machine Learning for Signal Processing, 2008. MLSP 2008. IEEE Workshop on*, pages 227–232. IEEE. 11

Dasarathy, B. V. (1997). Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE*, 85(1):24–38. 9

Dempster, A. P. (1967). A generalization of Bayesian inference. Technical report, DTIC Document. 20, 109

Dikmen, M., Akbas, E., Huang, T., and Ahuja, N. (2011). Pedestrian Recognition with a Learned Metric. *Computer Vision–ACCV 2010*, pages 501–512. 13, 14, 26, 28, 32, 34

Dlagnekov, L. and Belongie, S. (2005). Recognizing cars. *University of California San Diego, Tech. Rep. CS2005-0833*. 77, 78, 79

Donald, F. M. (2010). A model of cctv surveillance operator performance. *Ergonomics SA: Journal of the Ergonomics Society of South Africa*, 22(2):2–13. 1

Dong, J., Zhuang, D., Huang, Y., and Fu, J. (2009). Advances in multi-sensor data fusion: algorithms and applications. *Sensors*, 9(10):7771–7784. 11, 93

Dong, X. L. and Naumann, F. (2009). Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, 2(2):1654–1655. 3

Druzdzel, M. J. and Van Der Gaag, L. C. (1995). Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 141–148. Morgan Kaufmann Publishers Inc. 95

Du, S., Ibrahim, M., Shehata, M., and Badawy, W. (2013). Automatic license plate recognition (alpr): A state of the art review. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2):311–325. 67

Durrant-Whyte, H. F. (1988). Sensor models and multisensor integration. *The International Journal of Robotics Research*, 7(6):97–113. 8

Ekenel, H. K. and Stiefelhagen, R. (2005). Local appearance based face recognition using discrete cosine transform. In *13th European Signal Processing Conference (EUSIPCO 2005), Antalya, Turkey*. 12

Elgammal, A., Harwood, D., and Davis, L. (2000). Non-parametric model for background subtraction. In *Computer Visionï£¡ECCV 2000*, pages 751–767. Springer. 53

Ess, A., Leibe, B., and Van Gool, L. (2007). Depth and appearance for mobile scene analysis. 47

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188. 85, 86

Florea, M. C., Jousselme, A.-L., and Bosse, E. (2007). Fusion of imperfect information in the unified framework of random sets theory: Application to target identification. Technical report, DTIC Document. 3

Foresti, G. L. and Snidaro, L. (2002). A distributed sensor network for video surveillance of outdoor environments. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–525. IEEE. 12, 18

Freund, Y. and Schapire, R. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.*, 55(1):119–139. 80

Freund, Y. and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer. 78

Giacinto, G., Roli, F., and Didaci, L. (2003). Fusion of multiple classifiers for intrusion detection in computer networks. *Pattern Recognition Letters*, 24(12):1795–1803. 16

Goodman, I. R. (1997). *Mathematics of data fusion*, volume 37. Springer. 9

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, page 5. 13, 29, 32, 47, 65

Gray, D. and Tao, H. (2008). Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Computer Vision–ECCV 2008*, pages 262–275. 14, 28

Hall, D. D. L. and McMullen, S. A. H. (2004). *Mathematical Techniques in Multisensor Data Fusion 2nd Ed.* Artech House Publishers. 93

Hall, D. L. and Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, 85(1):6–23. 8, 16, 20, 93

Hamdoun, O., Moutarde, F., Stanciulescu, B., and Steux, B. (2008). Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. 47

Hamming, R. W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160. 67

Han, J., Kamber, M., and Pei, J. (2006). *Data mining: concepts and techniques*. Morgan kaufmann. 18

Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). Person re-identification by descriptive and discriminative classification. In *Proc. Scandinavian Conf. on Image Analysis*. The original publication is available at www.springerlink.com. 47

Hossain, M. A., Atrey, P. K., and Saddik, A. E. (2011). Modeling and assessing quality of information in multisensor multimedia monitoring systems. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 7(1):3. 22

Huang, T. and Russell, S. (1998). Object identification: A bayesian analysis with application to traffic surveillance. *Artificial Intelligence*, 103(1):77–93. 19

Iqbal, U., Zamir, S., Shahid, M., Parwaiz, K., Yasin, M., and Sarfraz, M. (2010). Image based vehicle type identification. In *International Conference on Information and Emerging Technologies (ICIET), 2010*, pages 1–5. IEEE. 79

Jensen, F. V. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London. 95

Joachims, T. (1999). Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. MIT Press. 80

Jolliffe, I. (2005). *Principal Component Analysis*. John Wiley Sons, Ltd. 29, 84, 85

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45. 73

Kanwal, Y., Arta, I., and Ali, J. (2013). Comparative analysis of automatic vehicle classification techniques: A survey. *International Journal of Image, Graphics and Signal Processing*, 4(9):52–59. 72

Kastinger, M., Hirzer, M., Wohlhart, P., Roth, P., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2295. IEEE. 81

Kelly, J. L. (1956). A new interpretation of information rate. *Information Theory, IRE Transactions on*, 2(3):185–189. 118

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464. 63, 75

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(6):1153–1160. 34

Khaleghi, B., Khamis, A., Karray, F. O., and Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion*, 14(1):28–44. 16, 22

Kokar, M. M., Tomasik, J. A., and Weyman, J. (2004). Formalizing classes of information fusion systems. *Information Fusion*, 5(3):189–202. 9

Koller, D., Daniilidis, K., Thorhallson, T., and Nagel, H.-H. (1992). Model-based object tracking in traffic scenes. In *Computer Visionï£¡ECCV'92*, pages 437–452. Springer. xii, 71, 72

Krell, G., Glodek, M., Panning, A., Siegert, I., Michaelis, B., Wendemuth, A., and Schwenker, F. (2013). Fusion of fragmentary classifier decisions for affective state recognition. In *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, pages 116–130. Springer. 3

Ku, H. (1969). Notes on the use of propagation of error formulas. *Precision Measurement and Calibration, NBS SP 3D0*, 1:331–341. 75

Kumar, P., Mittal, A., and Kumar, P. (2010). Addressing uncertainty in multi-modal fusion for improved object detection in dynamic environment. *Information Fusion*, 11(4):311–324. 23

Lee, H. (2006). Neural network approach to identify model of vehicles. *Advances in Neural Networks-ISNN 2006*, pages 66–72. 77, 79

Lee, S. H. and Chen, W. (2009). A comparative study of uncertainty propagation methods for black-box-type problems. *Structural and Multidisciplinary Optimization*, 37(3):239–253. 75

Lei, B., Thing, V. L., Chen, Y., and Lim, W.-Y. (2012). Logo classification with edge-based daisy descriptor. In *Multimedia (ISM), 2012 IEEE International Symposium on*, pages 222–228. IEEE. 68

Leon, F. P. and Kammel, S. (2003). Image fusion techniques for robust inspection of specular surfaces. In *AeroSense 2003*, pages 77–86. International Society for Optics and Photonics. 10

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707. 68

Lewicki, M. (Mar 27, 2007). Bayesian networks i. In *lecture notes distributed in Artificial Intelligence Class 15-381 at Carnegie Mellon University*. http://www.cs.cmu.edu/afs/cs/academic/class/15381-s07/www/slides/032707bayesNets1.pdf. 97

Li, W. and Li, L. (2009). A Novel Approach for Vehicle-logo Location Based on Edge Detection and Morphological Filter. In *Second International Symposium on Electronic Commerce and Security, 2009. ISECS'09*, volume 1, pages 343–345. IEEE. 77, 78

Li, X., Dick, A., Shen, C., Zhang, Z., van den Hengel, A., and Wang, H. (2013). Visual tracking with spatio-temporal dempster-shafer information fusion. *IEEE Transactions on Image Processing*, 22(8):3028–3040. 21

Lira, I. (2002). *Evaluating the measurement uncertainty: fundamentals and practical guidance*. CRC Press. 75

Liu, Y. and Li, S. (2011). A vehicle-logo location approach based on edge detection and projection. In *Vehicular Electronics and Safety (ICVES), 2011 IEEE International Conference on*, pages 165–168. IEEE. 77, 78

Liu, Y. and Rong, J. (2006). An efficient algorithm for local distance metric learning. In *Proceedings of AAAI*. 80

Llinas, J., Bowman, C., Rogova, G., Steinberg, A., et al. (2004). Revisiting the jdl data fusion model ii. In *In P. Svensson and J. Schubert (Eds.), Proceedings of the Seventh International Conference on Information Fusion (FUSION 2004*. Citeseer. 9

Lo, B. and Velastin, S. (2001). Automatic congestion detection system for underground platforms. In *Intelligent Multimedia, Video and Speech Processing, 2001. Proceedings of 2001 International Symposium on*, pages 158–161. IEEE. 53

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee. 78

Loy, C. C., Xiang, T., and Gong, S. (2010). Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129. 47

Lu, W., Zhang, H., Lan, K., and Guo, J. (2010). Detection of vehicle manufacture logos using contextual information. *Computer Vision–ACCV 2009*, pages 546–555. 77, 78

Luo, R. and Kay, M. (1992). Data fusion and sensor integration: State-of-the-art 1990s. *Data Fusion in Robotics and Machine Intelligence*, pages 7–135. 9

Ma, Q., Fosty, B., Crispim-Junior, C. F., Bremond, F., et al. (2013). Fusion framework for video event recognition. In *The 10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications*. 21

Maguire, B. and Desai, S. (2012). Dempster-shafer fusion for personnel detection: Application of dempster-shafer theory with ultrasonic micro-doppler and pir sensors. In *Information Fusion (FUSION), 2012 15th International Conference on*, pages 2209–2214. IEEE. 21

Maini, R. and Aggarwal, H. (2009). Study and comparison of various image edge detection techniques. *International journal of image processing (IJIP)*, 3(1):1–11. 134

Masashi, S. (2006). Local Fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of 23rd International Conference on Machine Learning*, pages 905–912. ACM. 84

Matthies, H. G. (2007). Quantifying uncertainty: modern computational representation of probability and applications. In *Extreme Man-Made and Natural Hazards in Dynamics of Structures*, pages 105–135. Springer. 75

Metternich, M., Worring, M., and Smeulders, A. (2010). Color based tracing in real-life surveillance data. *Transactions on data hiding and multimedia security V*, pages 18–33. 13

Morbee, M., Tessens, L., Aghajan, H., and Philips, W. (2010). Dempster-shafer based multi-view occupancy maps. *Electronics letters*, 46(5):341–343. 21

Morris, T. (2004). *Computer vision and image processing*. Palgrave Macmillan. 60

Mukundan, R. (2005). Radial tchebichef invariants for pattern recognition. In *TENCON 2005 2005 IEEE Region 10*, pages 1–6. IEEE. 79

Nakamura, E. F., Loureiro, A. A., and Frery, A. C. (2007). Information fusion for wireless sensor networks: Methods, models, and classifications. *ACM Computing Surveys (CSUR)*, 39(3):9. 16, 18

Network, E. R. (2011). Global video surveillance market to reach us $37.7 billion by 2015. *http://www.electronics.ca/presscenter/articles/1391/1/Global-Video-Surveillance-Market-to-reach-US-377-billion-By-2015/Page1.html*. Accessed: 2014-11-16. 2

Norris, C. and Armstrong, G. (1999). *The maximum surveillance society: The rise of CCTV*. Berg Publishers. 1

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3153–3160. IEEE. 48

Oliveira, L., Nunes, U., and Peixoto, P. (2010). On exploration of classifier ensemble synergism in pedestrian detection. *Intelligent Transportation Systems, IEEE Transactions on*, 11(1):16–27. 17

Oliver, N. M., Rosario, B., and Pentland, A. P. (2000). A bayesian computer vision system for modeling human interactions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):831–843. 53

Park, U., Jain, A., Kitahara, I., Kogure, K., and Hagita, N. (2006). Vise: Visual search engine using multiple networked cameras. *Pattern Recognition*, 3:1204–1207. 13

Pedagadi, S., Orwell, J., Velastin, S., and Boghossian, B. (June 23-27 2013). Local Fisher Discriminant Analysis for Pedestrian Re-identification. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE. 81, 84, 85

Petrovic, V. and Cootes, T. (2004). Analysis of Features for Rigid Structure Vehicle Type Recognition. In *British Machine Vision Conference*, volume 2, pages 587–596. 77, 79

PETS (2000). *PETS 2000 Dataset.* The original publication is available at http://www.stats.ox.ac.uk/ wauthier/tracker/pets2000.html. 47

PETS (2001). *PETS 2001 Dataset.* The original publication is available at http://www.anc.ed.ac.uk/demos/tracker/pets2001.html. 47

Piccardi, M. (2004). Background subtraction techniques: a review. In *Systems, man and cybernetics, 2004 IEEE international conference on*, volume 4, pages 3099–3104. IEEE. 54

Pohl, C. and Van Genderen, J. (1998). Review article multisensor image fusion in remote sensing: concepts, methods and applications. *International journal of remote sensing*, 19(5):823–854. 11, 16

Prince, S. J. (2012). *Computer vision: models, learning, and inference*. Cambridge University Press. 63

Psyllos, A., Anagnostopoulos, C., and Kayafas, E. (2011a). Vehicle model recognition from frontal view image measurements. *Computer Standards & Interfaces*, 33(2):142–151. 66

Psyllos, A., Anagnostopoulos, C., and Kayafas, E. (2011b). Vehicle model recognition from frontal view image measurements. *Computer Standards & Interfaces*, 33(2):142–151. 77, 78, 79

Psyllos, A. P., Anagnostopoulos, C.-N., and Kayafas, E. (2010). Vehicle logo recognition using a sift-based enhanced matching scheme. *IEEE Transactions on Intelligent Transportation Systems*, 11(2):322–328. 70, 77, 78, 79, 92

Rhead, M., Gurney, R., Ramalingam, S., and Cohen, N. (2012). Accuracy of automatic number plate recognition (anpr) and real world uk number plate problems. In *Procs 46th IEEE Int Carnahan Conf on Security Technology*. IEEE. 66, 67

Richardson, I. E. (2004). *H. 264 and MPEG-4 video compression: video coding for next-generation multimedia*. Wiley. com. 61

Rodriguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1619–1630. 79

Rosasco, L., Vito, E., Caponnetto, A., Piana, M., and Verri, A. (2004). Are loss functions all the same? *Neural Computation*, 16(5):1063–1076. 30

Ruta, D. and Gabrys, B. (2000). An overview of classifier fusion methods. *Computing and Information systems*, 7(1):1–10. 17

Sam, K. and Tian, X. (2012). Vehicle Logo Recognition Using Modest AdaBoost and Radial Tchebichef Moments. In *International Conference on Machine Learning and Computing*. 78, 79

Sanin, A., Sanderson, C., and Lovell, B. C. (2012). Shadow detection: A survey and comparative evaluation of recent methods. *Pattern recognition*, 45(4):1684–1695. 62

Seki, M., Wada, T., Fujiwara, H., and Sumi, K. (2003). Background subtraction based on cooccurrence of image variations. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–65. IEEE. 53

Sentz, K. and Ferson, S. (2002). *Combination of evidence in Dempster-Shafer theory*, volume 4015. Citeseer. 134

Shafer, G. (1976). *A mathematical theory of evidence*, volume 1. Princeton University Press. 20, 109

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press. 2, 101

Singh, V. P. (2013). *Entropy theory and its application in environmental and water engineering*. John Wiley & Sons. 3

Sivic, J. and Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. 13

Smets, P. (1983). Information content of an evidence. *International Journal of Man-Machine Studies*, 19(1):33–43. 3

Snidaro, L., Foresti, G. L., Niu, R., and Varshney, P. K. (2004). Sensor fusion for video surveillance. *Electrical Engineering and Computer Science*, Paper 108. 15

Snoek, C. G., Worring, M., and Smeulders, A. W. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM. 11

Sonka, M., Hlavac, V., Boyle, R., et al. (1999). *Image processing, analysis, and machine vision*. PWS Pub. Pacific Grove. 60

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE. 53, 54

Steinberg, A. N., Bowman, C. L., and White, F. E. (1999). Revisions to the jdl data fusion model. In *AeroSense'99*, pages 430–441. International Society for Optics and Photonics. 9

Stolkin, R., Rees, D., Talha, M., and Florescu, I. (2012). Bayesian fusion of thermal and visible spectra camera data for region based tracking with rapid background adaptation. In *Multisensor Fusion and Integration for Intelligent Systems (MFI), 2012 IEEE Conference on*, pages 192–199. IEEE. 19

Sumalee, A., Wang, J., Jedwanna, K., and Suwansawat, S. (2012). Probabilistic fusion of vehicle features for reidentification and travel time estimation using video image data. *Transportation Research Record: Journal of the Transportation Research Board*, 2308(1):73–82. 23, 24, 65

Sun, C. C., Arr, G. S., Ramachandran, R. P., and Ritchie, S. G. (2004). Vehicle reidentification using multidetector fusion. *IEEE Transactions on Intelligent Transportation Systems*, 5(3):155–164. 24

Swain, M. and Ballard, D. (1990). Indexing via color histograms. In *Third International Conference on Computer Vision, 1990. Proceedings*, pages 390–393. 13

Tan, X. and Triggs, B. (2010). Enhanced local texture feature sets for face recognition under difficult lighting conditions. *Image Processing, IEEE Transactions on*, 19(6):1635–1650. 12

Taylor, B. N. (2009). *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results (rev.* DIANE Publishing. 75

Taylor, J. R. (1997). *An introduction to error analysis: the study of uncertainties in physical measurements*. University science books. 61

Tetko, I. V., Livingstone, D. J., and Luik, A. I. (1995). Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833. 63

Torabi, A., Massé, G., and Bilodeau, G.-A. (2012). An iterative integrated framework for

thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications. *Computer Vision and Image Understanding*, 116(2):210–221. 24

Town, C. (2007). Multi-sensory and multi-modal fusion for sentient computing. *International Journal of Computer Vision*, 71(2):235–253. 20, 22

Toyama, K. and Horvitz, E. (2000). Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *FOURTH ASIAN CONFERENCE ON COMPUTER VISION (ACCV)*. 20

Turing-Institute and Siebert, J. (1987). *Vehicle Recognition Using Rule Based Methods*. TIRM–87-018. Turing Institute. 48

Tuzel, O., Porikli, F., and Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. *Computer Vision–ECCV 2006*, pages 589–600. 13

UK-Home-Office (2008). i-lids multiple camera tracking scenario definition. 47

Vezzan, R. (2010). Video surveillance online repository - stopped vehicle. The original publication is available at http://imagelab.ing.unimore.it/visor/. 47

Wang, J., Kankanhalli, M. S., Yan, W., and Jain, R. (2003). Experiential sampling for video surveillance. In *First ACM SIGMM international workshop on Video surveillance*, pages 77–86. ACM. 12

Wang, Y., Liu, Z., and Xiao, F. (2007). A fast coarse-to-fine vehicle logo detection and recognition method. In *IEEE International Conference on Robotics and Biomimetics, ROBIO 2007*, pages 691–696. IEEE. 77, 78, 92

Weinberger, K., Blitzer, J., and Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. In *In NIPS*. Citeseer. 29

Weinberger, K. and Saul, L. (2009). Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244. 80

Weinhaus, F. (2013). Feather mask. *http://www.fmwconcepts.com/imagemagick/feather/*. Accessed: 2014-11-19. 30

White, F. E. (1987). Data fusion lexicon. *Joint Directors of Laboratories, Technical Panel for C , Data Fusion Sub-Panel.* 9

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785. 53

Wu, Y., Chang, E. Y., Chang, K. C.-C., and Smith, J. R. (2004). Optimal multimodal fusion for multimedia data analysis. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 572–579. ACM. 15

Xiaofei, H. (2004). Locality preserving projections. In *Neural information processing systems*, volume 16, page 153. MIT. 85

Xing, E., Ng, A., Jordan, M., and Russell, S. (2003). Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15. 80

Xu, H. and Chua, T.-S. (2006). Fusion of av features and external information sources for event detection in team sports video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2(1):44–67. 15

Xu, L., Krzyzak, A., and Suen, C. Y. (1992). Methods of combining multiple classifiers and their applications to handwriting recognition. *Systems, Man and Cybernetics, IEEE Transactions on*, 22(3):418–435. 17

Yager, R. R., Liu, L., et al. (2008). *Classic works of the Dempster-Shafer theory of belief functions*, volume 219. Springer. 111

Yang, W., Yang, G., Zheng, X., and Zhang, L. (2012). An improved vehicle-logo localization algorithm based on texture analysis. In *International Conference on Computer Science and Information Processing (CSIP), 2012*, pages 648–651,. IEEE. 77, 78

Ying, Y., Huang, K., and Campbell, C. (2009). Sparse Metric Learning via Smooth Optimization. In *NIPS*, pages 2214–2222. 80

Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, 8(3):338–353. 21

Zelnik-Manor, L. and Perona, P. (2004). Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press. 85

Zhang, B. and Zhou, Y. (2012). Vehicle Type and Make Recognition by Combined Features and Rotation Forest Ensemble. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(03). 78, 79

Zhang, D., Lu, G., et al. (2001). A comparative study on shape retrieval using fourier descriptors with different shape signatures. In *Proc. of international conference on intelligent multimedia and distance education (ICIMADE01)*, pages 1–9. 79

Zhao, W., Fang, T., and Jiang, Y. (2007). Data fusion using improved dempster-shafer evidence theory for vehicle detection. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, volume 1, pages 487–491. IEEE. 21

Zheng, W., Gong, S., and Xiang, T. (2011). Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition*. 14, 28

Re-identification of Pedestrians with Variable Occlusion and Scale

Simi Wang, Michal Lewandowski, James Annesley, James Orwell
Digital Imaging Research Center
Kingston University
simi.wang@kingston.ac.uk

Abstract

This paper presents results of experiments designed to measure the accuracy with which people can be re-identified using multiple visual surveillance observations. Two public data sets are used: VIPeR and a new public data set, V-47. The re-identification method is a Large Margin Nearest Neighbour classifier using feature vectors constructed from overlapping block histograms. The experiments investigate the performance with respect to the level of occlusion, the training regime, specificity of the domain and the resolution of the observations. A method is proposed that reduces the adverse impact of occlusions, when present; and increases the beneficial impact of higher resolution data, when available.

1. Introduction

Pedestrian re-identification is an important component of visual surveillance analysis in public spaces. The ability to assign a single correct identifier to multiple observations of an individual improves the semantic coherence of the analysis. This, in turn, is useful to construct descriptions of behaviour, and to facilitate the retrieval of data relevant to a given individual.

The difficulty of the problem is partly determined by the extent of time and space over which these 'multiple observations' are recorded. At one extreme, it is part of the pedestrian tracking problem; this is particularly important in more crowded or occluded scenes. Similarity of appearance is used [3] alongside spatio-temporal measurements to estimate their trajectories. In these cases, the appearance and pose of the observations are relatively similar. The problem is more difficult using observations from multiple cameras and discontinuous trajectories: not only pose, viewpoint and illumination vary, but the number of possible candidates is most likely increased. In the extreme case, re-identifying a pedestrian at some arbitrary location and future point in time is indeed a challenging problem: some aspects of the appearance such as clothing and hair may have changed,

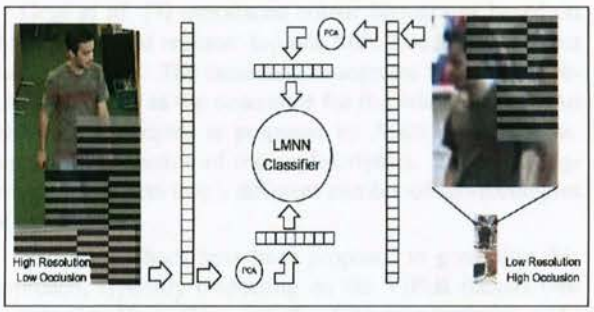


Figure 1. The proposed method extracts histograms from overlapping letterbox-shape regions and uses their principal components to train a Large Margin Nearest Neighbour classifier. This paper contributes enhancements to the technique, which improves performance for occluded (right) and higher resolution (left) observations. It is demonstrated on the VIPeR and the (new) V-47 public datasets.

and the less changeable aspects such as face and gait are more difficult to analyse (assuming the pedestrian is not actively co-operating and that the environment is relatively uncontrolled). Furthermore, re-identifying people 'in the wild' will require robust processing of partial and incomplete observations, due to crowds and clutter.

An increasingly common place formulation of the problem [7, 5] presents a single 'probe' image, together with a 'target' set containing exactly one additional observation of the probe. The output is a ranked list of elements of the target set, which can be aggregated over a test set into a cumulative match characteristic (CMC) curve. The normalised area under this curve is a straightforward and intuitive performance measure.

The work presented in this paper ('this work') validates and extends recent work [5] that applies the Large Margin Nearest Neighbour classifier to this problem. That method currently provides the best published ('benchmark') performance on the VIPeR dataset and problem formulation. This work examines how this impressive performance can be retained and improved under varying experimental con-

Table 1. Pedestrian Re-identification Datasets

Name	media	resolution	#people	#images	pose	occlusion
VIPeR	Still	128 by 48	316	632	various	No
CAVIAR	Video	384 by 288	?	?	various	Yes
ETHZ1	Video	64 by 32	83	4857	various	No
ETHZ2	Video	64 by 32	35	1936	various	No
ETHZ3	Video	64 by 32	28	1762	various	No
MCTS	Still	128 by 64	119	476	various	Yes
V-47	Video	576 by 720	47	752	various	Yes

ditions.

The two principle variations are the resolution of the dataset and the presence of occlusions in the probe and target observations. The experiments show that an alternative construction of the feature vector leads to improved performance on higher resolution data, yet retaining the same performance of the benchmark experiments. Similarly, experiments show that by using a training set that includes occluded pedestrians, performance in these conditions can also be improved. Two secondary investigations are the dependence on pose; and on the size and specificity of the training set. To accompany this paper a new public dataset is presented along with all the source code for the benchmark and additional experiments.

This paper is organized as follows. In the following sections the previous approaches and available available datasets are described. The Large Margin Nearest Neighbour classifier is described in some detail since it is used in the subsequent sections and experiments. In section 5 the adaptation of this approach is described, to accommodate observations with part-occlusion and to exploit higher resolution, when available.

2. Previous Work

There are many categories of person re-identification methods, this paper will concentrate on the use of colour histogram as a tool for generating descriptions. The seminal work [14] demonstrated colour indexing for retrieval; histograms were compared using an 'intersection' operator that is similar to the L1 norm. The histogram is a non-parametric, quantised representation of the accumulated values. One alternative is the parameteric family of representations of *e.g.* second order statistics [10], possibly with mixture estimation [15]. Another alternative is to find and represent multiple salient points in the observation, *e.g.* using SIFT features [13].

Park *et al.* [12] extend the histogram-based representation by dividing the region of interest into horizontal partitions, to form histograms which are then concatenated together into the feature vector. This is a special case of a more general set of robust computer vision methods in which overlapping regions are used to achieve spatial selectively with spatially tolerant accumulators, *e.g.* HOG for pedestrian detection [4].

Gray *et al.* [7] introduced colour histograms based on three predefined regions: top one fifth; middle and bottom each two-fifths. The combined histograms for all three regions are used as the descriptor for the whole image. An improved descriptor is proposed by Alahi *et al.* [1], using a grid collection of region descriptors. Each grid segments the objects into a different number of sub-rectangles of equal sizes.

Various methods have been proposed to generalise this approach, typically evaluating on the VIPeR dataset (see section 3.) First, Gray and Tao [8] use a boosting technique to optimise a set of histogram features from a large combinatorial space. Zheng *et al.* [17] construct histograms for each of over twenty types of feature, for six horizontal stripes across the bounding box. Finally, Dikmen *et al.* [5] extends the Large Margin Nearest Neighbour classifier [16] and applies on an array of histogram responses extracted from overlapping regions. This is regarded as the current benchmark method since it provides the best performance when evaluated on an unseen VIPeR subset.

3. Datasets for Re-identification

There are several public datasets appropriate for this problem, such as VIPeR [7], CAVIAR [9], and ETHZ sets [6]. Also, the i-LIDS Multiple Camera Tracking Dataset [11] is intended to evaluate systems that would use pedestrian re-identification components.

These datasets provide variations in pose, viewpoint, lighting variation and occlusion. One issue with these datasets is their modest common resolution: pedestrians are 128 image lines in height. This effectively excludes the use of high spatial frequency features since they are not represented with sufficient resolution. One motivation for the proposed dataset is to introduce a higher resolution dataset for the pedestrian re-identification problem. The dataset also includes occlusion that blocks part of the body, this is useful to solving re-identification problems in crowded environments where only parts of the person could be observed.

3.1. The V-47 dataset

The V-47 dataset comprises video of 47 participants walking in both directions through an indoor route, observed by two progressive scan DV cameras. The scene had both artificial and natural lighting, which varied thought the duration of the activity. There are 4 video sequences for each participant (two cameras, two directions). Four still images are extracted for each view. More information is available at the dataset website [2]

4. Re-identification method using LMNN-R

In this section, we briefly describe the use [5] of the Large Margin Nearest Neighbour classifier with Rejection (LMNN-R) to learn a metric suitable for pedestrian re-identification. This is the state-of-the-art approach for person re-identification on the benchmark VIPeR dataset.

First, images are vectorized using 8-bin histograms for each channel of RGB and HSV colour space. These histograms are extracted from 8x24 rectangular regions, which in turn are densely collected from a regular grid with 4 pixel spacing in vertical and 12 pixel spacing in horizontal direction. This step size is equal to half the width and length of the rectangles resulting in an overlapping representation. A feature vector $\vec{x}$ is obtained by concatenation of all histograms. Finally, a PCA is applied to obtain a smaller space for the metric learning.

The objective is to learn a linear transformation $L : \Re^d \rightarrow \Re^d$ that minimises a distance between each training low dimensional point and its K nearest similarly labelled neighbours, while maximizing the distance between all differently labelled points according to a constant margin. As a consequence, a pairwise similarity of feature vectors is measured by following weighted squared distance:

$$D(\overrightarrow{x_i}, \overrightarrow{x_j}) = \|L(\overrightarrow{x_i} - \overrightarrow{x_j}\|^2 \tag{1}$$

which can be reformulated to the equivalent Mahalanobis metric:

$$D(\overrightarrow{x_i}, \overrightarrow{x_j}) = (\overrightarrow{x_i} - \overrightarrow{x_j})^T M (\overrightarrow{x_i} - \overrightarrow{x_j}) \tag{2}$$

assuming that M is a symmetric positive-semidefinite matrix, so it can be factorised into real-valued matrices as $M = L^T L$. The objective function over the distance metrics parameterised by eq. 1 or eq. 2 has two competing terms: $\epsilon(M) = \epsilon_1(M) + \epsilon_2(M)$. The first term penalizes large distances between each point i and its neighbours j according to Euclidean norm:

$$\epsilon_1(M) = \sum_{i,j} D(\overrightarrow{x_i}, \overrightarrow{x_j}) \tag{3}$$

while the second term penalizes small distances between each point and all other differently labeled ones:

$$\epsilon_2(M) = \sum_{i,k} (1 - \delta_{i,k})[1 + \frac{1}{NK} \sum_{m,l} D(\overrightarrow{x_m}, \overrightarrow{x_l})) - D(\overrightarrow{x_i}, \overrightarrow{x_k})]_+ \tag{4}$$

Here, $\delta_{i,k}$ is an indicator variable which is 1 if and only if x_i and x_j belong to the same class, 0 otherwise. The x_k for which $\delta_{i,k} = 0$ are so called impostors for x_i. The closest impostors of a training point are forced to be at least a certain distance away from the considered point x_i. This distance is computed by the average distance of all K nearest neighbour pairs (m, l) in the training set and it is only marginally affected by its own K nearest neighbours. The expression $[z]_+ = \max(z, 0)$ denotes the standard hinge loss. This optimisation process can be be solved as an instance of semi-definite program when distance D is given by non-quadratic equation 2.

5. Generalising over occlusions and scales

5.1. Observations with Part Occlusions

CCTV observations of pedestrians are often partly occluded due to crowded environments and obstacles: it is important for re-identification methods, e.g. LMNN-R, to perform robustly in these cases. However, all members of the VIPeR dataset are fully visible. To investigate the re-identification performance in the presence of occlusions, a set of occluded pedestrians was synthesised from the VIPeR dataset, and an occluded subset of the V-47 dataset was also used. The former dataset was synthesised by overlaying another randomly selected pedestrian (from the same dataset) on top of the target pedestrian, using a feathered elliptical mask. The placement was varied stochastically, with a mean occlusion level of 50%, to simulate typical observations taken from crowded scenes. In the V-47 dataset, the layout of the scene entailed that a specific subset of the pedestrian observations included real occlusions, from approximately the waist down, and so these formed the second occluded dataset to work with. A few examples of occluded images are depicted in Figure 5.1. Such occluded images can then be directly fed into training/testing procedure of LMNN-R classifier as described in section 4.

First, an experiment can be designed to investigate if a classifier trained on occluded examples will improve the performance, as measured on an occluded test set. However the performance of that classifier may deteriorate (compared with the benchmark) when tested on the original (unoccluded) test set. To balance these opposing factors one must then construct a second experiment using a mixed test set of (for example) 50% each of occluded and non-occluded data.

Two strategies for constructing a classifier to work on this mixed dataset are considered. First, a single hybrid

Table 2. Comparison of Performance with & without occlusions

notation	Training set	Test set	Performance	
TrTe	no occlusions	no occlusions	95 %	*benchmark result*
*TrTe***C**	no occlusions	with occlusions	80 %	
*Tr***C***Te*	with occlusions	no occlusions	74 %	
*Tr***C***Te***C**	with occlusions	with occlusions	87 %	

Table 3. Comparison of Performance on mixed test set

Strategy 1: hybrid classifier	83%	*benchmark result*
Strategy 2: joint classifiers (perfect occlusion detection):	91 %	
Strategy 2: joint classifiers (random occlusion detection):	84 %	

classifier, trained with a mixed training set. Second, a joint system with two classifiers, trained on occluded / non-occluded data respectively. For each test input, the former classifier is used only if an occlusion is detected in either probe or target observation. The second strategy requires an additional component to detect if either probe or target observation is occluded. Even so, an upper (and lower) bound can be estimated for its performance on the mixed dataset, by using simulating a perfect (and random) occlusion detector. This will provide a preliminary indication of relative performance between these strategies.

(a)

(b)

Figure 2. Examples of occluded observations: (a) synthesised from VIPeR dataset and (b) real from V-47 dataset.

5.2. Variable Resolution

In the V-47 dataset, the pedestrian height (in image scanlines) varies from 140 to 480. The method outlined in Section 4 was designed and tested on a vertical resolution of 128 lines. Higher resolution input can always be down sampled as necessary to fit this expected input size but this will discard higher spatial frequency signals which may contribute to the discrimination task. Operating directly on the original resolution signal gives the opportunity to preserve this information. The operation to accumulate RGB (or HSV) values from overlapping boxes into histograms can generalised to accept input of any size, by scaling the size of these input boxes accordingly. This aggregation of pixel values into histograms also discards information. However, it is hypothesised that by including smaller scale boxes, this effect is mitigated and better use is made of the higher resolution input.

6. Experimental Results

Four different experiments are presented, investigating the effects of occlusion, change in resolution, any dependency on pose, and the similarity of training set to test set. Where possible reference is made to the benchmark experiment provided in [5]. In all experiments, the subjects used in the test sets were never included in training sets. Random splits between training and testing were used to generate enough results to be statistically significant.

6.1. On occluded & non-occluded data

As discussed in Section 5.1, this experiment measures the re-identification performance on occluded observations. The first step is to measure the performance on datasets with/without occlusions, using classifiers trained with/without occlusions. From analysis of the benchmark results [5], the following configuration is adopted: original RGB+HSV image space, retaining first 60 principal components, averaged over 10 random splits.

The following classifiers were trained and evaluated, in each case reporting the Normalised Area Under CMC (%):

From the results in Table 2 it is clear that the occluded test set (with results underlined) is a more challenging problem, and that better performance is achieved by a classifier that is trained on occluded data. However it is equally clear that this classifier performs worst of all (74 %) on the unoccluded data.

We now turn to the problem of how to achieve best possible performance on mixed (occluded and unoccluded) data. Three results are presented: the hybrid strategy, the joint strategy (upper bound) and the joint strategy (lower bound). Recall that the upper bound simulates a perfect decision

between occluded and unoccluded input, while the lower bound simulates a random decision.

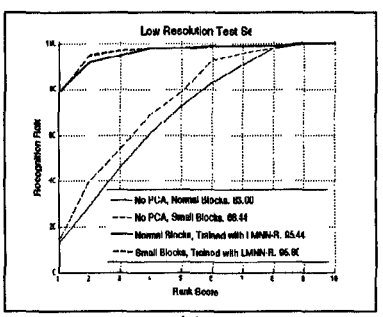
The results presented in Table 3 suggest that it is worthwhile to pursue a strategy of training specific classifiers for occluded and non-occluded observations, rather than use a single classifier, trained on both types of data.

6.2. Higher Resolution Observations

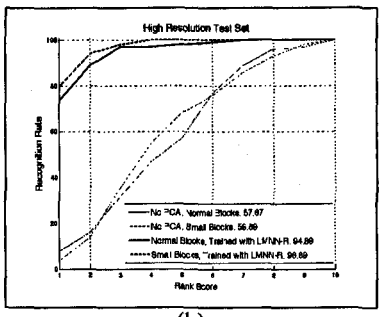
The feature vector defined in the benchmark method [5] uses 38 vertical and 4 horizontal overlapping bins to generate the block histograms, as the original VIPeR image dimensions are 128x48. In comparison to the VIPeR set, the V-47 dataset has a higher resolution, which may improve performance. To investigate this, bi-cubic interpolation was used to render the extracted images to a common highest denominator of resolution (480x264). Also, a low resolution (128x48) version of the V-47 dataset was produced to allow direct comparison with the VIPeR dataset.

Experiments were conducted to compare the normal block size [5] with the smaller block size proposed in Section 5.2. These alternate schemes were tested on both low and high resolution test sets, and plotted in Figure 6.2.The numbers in the legend describe the normalised percentage area under the CMC curve. The training set consisted of 37 individuals, with the remaining 10 used in the test. This was repeated ten times with random splits.

Firstly it is worth noting that the results obtained on the 'raw' feature vectors (57-68 %) are significantly inferior to those obtained from the trained classifier (95-97 %) Comparison of the two block sizes suggests that the smaller block size produces better results: for the high resolution data, approximately two-fifths of the error is removed by using it. In addition, for the training process, the high resolution data using the smaller block size also converges quicker.



(a)



(b)

Figure 3. CMC Curves For Changes in Feature Vector; (a) Low Resolution V-47 Data (b) High Resolution V-47 Data

6.3. Dependency on Pose

Experiments were conducted to investigate any dependency on the pose of the subjects in the probe and target sets. These experiments use the higher resolution V-47 data, adopting the increased vertical feature vectors; the performance is listed in Table 6.3. Similar results were achieved when at least one (target or probe) set of images in the training uses a front pose image. Increased accuracy rate is observed when only the back pose images are used. The explanation for this difference is that the colour pattern in the front of the subjects' clothing has more variance (less consistency) compared to the back of the clothing.

Pose Variation (Target Vs Probe)		
FP Vs FP	BP Vs BP	FP Vs BP or BP Vs FP
96.56 ± 2.79	99.00 ± 1.26	96.89 ± 1.63

Table 4. Area under the CMC curve for different poses: FP = Front Pose, BP = Back Pose.

6.4. Domain Specificity

This experiment evaluates the performance of three classifiers on 10 unseen pairs of high resolution (V-47) observations. The first classifier was trained on 316 pairs from the VIPeR dataset. The second classifier was trained on 333 pairs taken from 37 individuals, observed from *different* camera views from those used in the test set. The final classifier has was trained on 333 pairs taken from 37 individuals, observed from the *same* camera views from those used in the test set. The process is repeated over ten random splits to generate sufficient results, which are plotted in Figure 4.

There is a significant dependence on the type of training data: the best results are obtained when a similar view is used to provide training data for the classifier. This suggests that both the VIPeR and V-47 training sets exhibit some degree of exclusive properties that are not shared.

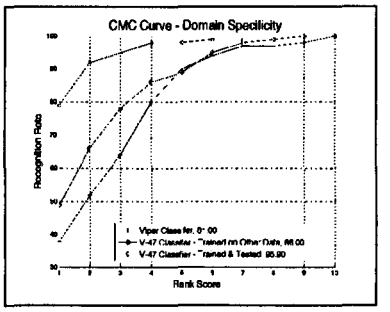


Figure 4. CMC Curves Domain Specificity

6.5. Use of Additional Training Samples

This experiment is designed to investigate whether a given performance can be improved by adding more observations of the same individuals used to construct the test set. For example, if 3 observations of each individual are available in 2 cameras, then there are 9 possible intra-camera pairs that could be used for training. The area under the CMC curve is plotted at several stages of the classifier construction: Prior to the PCA step, *i.e.* the raw feature vector; after the PCA but prior to training; and the final classifier result. The results are plotted in Figure 5. As in the previous experiments, there is a 37/10 training/test split, randomly repeated 10 times. The results suggest that a small improvement in the performance can be obtained by using the additional training samples. The difference is clear prior to training: the PCA subspace obtained by using more samples appears to be a more effective space in which to categorise previously unseen pairs of observations. This small improvement is carried through the training stage, where it then represents a significant reduction in the error rate.

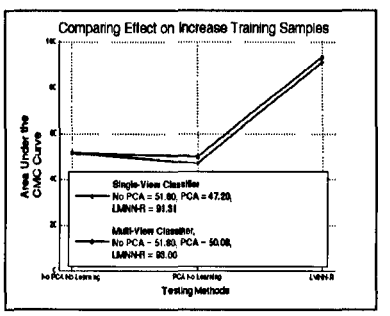


Figure 5. Performance of two classifiers, trained on 37 individuals, at several stages of the classifier construction. One was trained on 37 pairs of observations, while the other was trained on 333 pairs of observations.

7. Conclusion

This work has investigated the use of the Large Margin Nearest Neighbour classifier to re-identify pedestrians viewed from different cameras and at various resolutions and levels of occlusion. As far as the authors know this is the best-performing method, as evaluated on the VIPeR dataset. The experiments described in this work allow the following conclusions to be drawn. Firstly, for potentially occluded observations, the best strategy is to attempt the detection of an occlusion and then deploy the appropriate classifier. This achieves better results than training a classifier on a mix of occluded and non-occluded data. Second, the use of a smaller block size (than the benchmark) allows improved results when higher resolution data is available. Furthermore, best performance is obtained from the rear view of both probe and target; and (more significantly) when training data is from the same domain as the test set.

Future work will concentrate on the identification of a globally effective training set, the incorporation of occlusion information into the re-identification process, and the investigation of further methods to make better use of high resolution information.

References

[1] A. Alahi, P. Vandergheynst, M. Bierlaire, and M. Kunt. Cascade of descriptors to detect and track objects across any network of cameras. *Computer Vision and Image Understanding*, 114(6):624–640, 2010. 2

[2] Anon. Temporary website for v-47 dataset: https://docs.google.com/leaf?id=0b2qwvl6muriztqwmwiwzdgtyte4yy00mzhllwiynjutngfizjrjmje0odk4, 2011. 3

[3] M. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool. Robust tracking-by-detection using a detector confidence particle filter. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1515–1522. IEEE, 2009. 1

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, 1:886–893, 2005. 2

[5] M. Dikmen, E. Akbas, T. Huang, and N. Ahuja. Pedestrian Recognition with a Learned Metric. *Computer Vision–ACCV 2010*, pages 501–512, 2011. 1, 2, 3, 4, 5

[6] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. 2007. 2

[7] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, page 5, 2007. 1, 2

[8] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. *Computer Vision–ECCV 2008*, pages 262–275, 2008. 2

[9] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. 2008. 2

[10] M. Metternich, M. Worring, and A. Smeulders. Color based tracing in real-life surveillance data. *Transactions on data hiding and multimedia security V*, pages 18–33, 2010. 2

[11] U. H. Office. i-lids multiple camera tracking scenario definition, 2008. 2

[12] U. Park, A. Jain, I. Kitahara, K. Kogure, and N. Hagita. Vise: Visual search engine using multiple networked cameras. *Pattern Recognition*, 3:1204–1207, 2006. 2

[13] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. 2003. 2

[14] M. Swain and D. Ballard. Indexing via color histograms. In *Third International Conference on Computer Vision, 1990. Proceedings*, pages 390–393, 1990. 2

[15] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *Computer Vision–ECCV 2006*, pages 589–600, 2006. 2

[16] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *In NIPS*. Citeseer, 2006. 2

[17] W. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

Vehicle Logo Recognition Using Local Fisher Discriminant Analysis

Simi Wang, Sateesh Pedagadi, James Orwell and Gordon Hunter

Digital Imaging Research Centre, Kingston University, U.K.

Keywords: vehicle logo recognition metric learning

Abstract

This paper presents a method for localising and recognising vehicle manufacturer logos in both the front and rear views. The method assumes that the vehicle registration plate is visible and an estimate of its location is available. Features are constructed out of local histograms of gradients, in both conventional and hierarchical arrangements. The dimensionality of these vectors is then reduced using unsupervised PCA and, subsequently a supervised method based on Local Fisher Discriminant Analysis. This then provides a suitable metric for logo detection, localisation and multi-class classification. On a test set of data captured from a medium range CCTV camera, with five different manufacturers' logos, the proposed method provided a correct logo localisation rate of 97 % and a correct logo classification rate of 87.6%.

1 Introduction

The correct identification of vehicles is important in a range of surveillance situations. Currently, the completion of this task relies predominantly on the correct identification of the characters on the vehicle licence plate (LP). However, the fact that the LP can be obscured, altered or replaced means that secondary key attributes of the car are needed to identify the vehicle in these situations. This has given rise to the need for an accurate recognition and classification of a vehicle's manufacturer logo, as these cannot be easily altered.

Resent research has relied on the texture information of the vehicle's grille to find a coarse Region of Interest (RoI) in which the logo could then be finely located, as traditionally the logo is located in the middle of the vehicle's grille. However, with modern vehicle designs the logo could be placed on top of the bonnet, where grille information becomes less relevant. This is even more apparent when locating the logo on vehicles viewed from the rear, where the grille information is unavailable. This has given motivation for this paper, to break the traditional research path in this field of using only front views of the vehicles captured in controlled environmental conditions.

In this paper, we will introduce a challenging dataset, that simulates vehicles in a real-world situation, where the logos are of varying sizes, viewed from both the front and rear, and in varying lighting and environmental conditions. We propose a novel way to localise the logo's region of interest and then utilise this localised region to classify the observed logos into different categories using a Fisher Discriminative multi-class classifier. The vehicle logo recognition process can be divided up into two stages: Logo Localisation and Logo Classification.

The rest of paper is organised as follows: Section 2 presents a review of the previous work that has been conducted in this field. Section 3 introduces the logo localisation and classification methods. In section 4, some experimental results are present and discussed. Our conclusion and proposal for future work are given in section 5.

2 Related Work

2.1 Vehicle Logo Localisation

The localisation of the logo, within the view of the vehicle, is an essential step before its accurate classification can be achieved. The majority of localisation research relies heavily on some prior knowledge such as the position of the licence plate (LP). Once the position of the LP is located, researchers then define a Region of Interest (RoI) relative to it . The authors of [14, 15, 28] assumed the RoI for the vehicle logo to be a patch above the LP, of a size relative to the size of the extracted LP. The researchers in [5, 13, 21, 20, 22, 25] defined the RoI as an area of the front of the car relative to the size and location of the LP, incorporating the LP, grille, lights, etc. The approach used by Lu et al. [17] adopts three reference points to define a RoI containing the vehicle logo and the grille. Their three reference points are the LP, and the left and right headlights.

From a large RoI, Lee [13] extracted a smaller area of interest which incorporates the grille and logo by using texture information. The texture information of the grille is also used by the authors of [14, 15, 17, 28] and, combined with edge information obtained using different detectors and filters, the authors were able to reduce the size of the RoI to incorporate the logo only. In Wang et al. [25], the authors used the peaks in the edges' vertical direction projection as the initial location to start a symmetry search in order to locate the logo. A completely different method was adopted by Psyllos et al. [21, 22], who used the Phase Congruency Feature Map and its derivatives to divide the RoI into smaller areas, such as left/right light, grille and logo.

An attempt to remove the dependency on the LP is described by Sam and Tian [23], who utilised the Modest Adaboost algorithm for searching for vehicle logos, represented by extended Haar-like feature. However, the gradually sliding window used in the search makes the method sensitive to the

usually complicated background, thereby limiting its application. Zhang and Zhou's [31] method uses the frontal images of the vehicle and adopts a bilateral symmetry detection based on a set of SIFT features. Although this method has reported a localisation accuracy of 98.91%, its reliance on the grille information makes the method unsuitable rear-view logo localisation.

2.2 Vehicle Logo Classification

Previous research into the the recognition of a vehicle's manufacturer logo is limited, even though this is an attribute that could be used in vehicle identification systems. Early work by Dlagnekov et al. [5] uses SIFT features to recognise the vehicle's manufacturer using images of the whole rear-view of the vehicle, not just the logo, attaining 89.5% recognition rate. Psyllos et al. [21, 22] elaborated on the work of Dlagnekov et al. [5] by proposing a SIFT-based enhanced matching scheme, which concentrated on the logo only. The scheme boosted the recognition accuracy compared to the standard SIFT- based feature-matching method. Wang et al. [25] presented a method for logo recognition using template matching and an edge orientation histogram of the logo. The methods proposed by Wang et al. [25] and Psyllos et al. [21, 22] all face the issue of reduced robustness with variation of the environmental conditions, such as lighting. A solution was proposed by Burkhard et al. [3], whose method uses Fourier descriptors as they are not as sensitive to distortion due to environmental lighting effects. However, this method is highly dependent on the logo segmentation as it requires the logo to be the dominant element in the image. Recently, Sam and Tian [23] applied the invariance property of radial Tchebichef moments to recognise high resolution segmented vehicle logos, achieving a recognition rate of 92% on their own dataset.

Although there has previously been limited focus on the recognition and classification of vehicle logos, some research based on a RoI within a frontal or rear view of the vehicle which includes the logo has been conducted. Lee [13] used a set of 16 texture descriptors of the RoI taken from the front view of the vehicle as the input to a 3-layer back propagation multi-layer perception neural network. This method was used to classify vehicles into 24 different classes and achieved a recognition rate of 94%. Petrovic and Cootes [20] also used the RoI from the front of the vehicle in an approach based on edge oriented gradients and a match refinement algorithm for vehicle model recognition. Petrovic and Cootes reported a recognition rate of over 93% on parked cars containing 77 classes. Zhang et al. [31] proposed the use of a Rotation Forest Ensemble method for vehicle classification. Their classification method uses the features from a Fast Discrete Curvelet Transform and the Pyramid Histograms of Orientated Gradients (pHoG) of the RoI from the top view of the vehicle. Similar to the approach of Dlagnekov et al. [5], Bhanu and Kafai [1] tried to classify vehicles using the rear view of the vehicle, the vehicle being categorised into classes of vehicle type, such as van, car or truck etc., rather than make or model.

Iqbal et al. [9] conducted a comparison of previous techniques used for vehicle model classification using both environmentally controlled and uncontrolled datasets. They noted that techniques such as SIFT are sufficient in controlled environments, where there is little variation of illumination, viewing angle and scale. Their research also concluded that for make or model recognition, the RoI from rear-view images performed better than the RoI of the frontal view.

2.3 Metric Learning for Classification

Classification methods can be broadly categorized into feature-based and learning-based methods. Feature-based methods rely on the discriminative ability of the feature alone, while learning-based methods estimate a discriminative model by analyzing the training data which is representative of the collected data.

SVMs[10], Boosting[7] and Neural networks[2] have been successfully applied for learning two class classifiers in various vision related problems e.g.: Pedestrian detection, Face detection etc.. Multi-class classification has been addressed successfully in learning methods [26], [27], [29] that mainly focused on metric learning which requires a Mahalanobis distance metric to be estimated in the feature space. The feature space is often non-linear in nature and thus requires a transformed feature space in which the Euclidean distance between data samples maintains the neighbourhood characteristics of data. Metric learning has been considered as a data association problem when multiple classes are involved. The Mahalanobis metric is consistent with a positive semi-definite matrix, and the general set of such metric matrices all of which are positive semi-definite, can be considered to be the interior and surface of a cone with apex at the origin. Other methods such as LDM[16], LMNN[26] and that of Xing et al. [27] estimate this metric by modelling the solution as an optimization problem where strategies like gradient descent approaches are employed. However, scalability with increasing feature dimensions tends to be problematic with such approaches due to computationally expensive and time consuming iterative steps involved. A few very recent methods, KISS-ME[12], LF[19] have modelled the solution for estimating the metric in eigenspaces. The solution in these cases can be easily computed by solving a formulated eigenvalue problem. Recently, Local Fisher LF [19] was shown to produce better discrimination amongst sub space methods by using relatively simple features. The proposed method employs LF's ability to include two different feature types to maintain dual representation of data in feature space in order to learn a discriminative transformation space.

3 Vehicle Logo Localisation and Detection

3.1 Feature Extraction

The region of interest of most common logos in both the frontal and rear views, is at a position some distance above the LP. Therefore, to extract the RoI the position of the LP, which can be detected very accurately using Automated Number Plate

Recognition (ANPR) modules, is used. The relationship between the coordinates and the logo's RoI is shown in Figure 1, and is expressed in equation 1.

$$\begin{cases} \mathbf{r}_{2x} - \mathbf{r}_{1x} = R_x \\ \frac{\mathbf{r}_{2x}+\mathbf{r}_{1x}}{2} = \frac{\mathbf{p}_{2x}+\mathbf{p}_{1x}}{2} \\ \mathbf{p}_{1y} - \mathbf{r}_{1y} = R_y \end{cases} \quad (1)$$

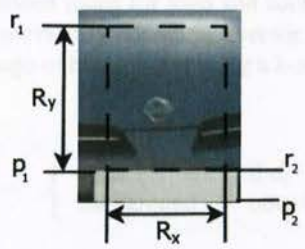


Figure 1: Logo RoI Extraction

Referring to equation 1 $\mathbf{p}_1$ and $\mathbf{p}_2$ are the respective top-left and bottom-right coordinates of the located LP $\mathbf{r}_1$ and $\mathbf{r}_2$ are the respective top-left and bottom-right coordinates of the located RoI. R_x and R_y denote the width and height of the RoI which 128 and 152 pixels respectively.

Once a logo's RoI is extracted, the region is sub-divided into patches. Each patch size is 128 by 64 pixels, and the first patch is at the top left of the RoI and successive patches are created by moving down 5 pixels each time. The patches were classed into two categories, logos (l) and background (b). A logo patch is defined as a patch that contains a least 50% of the logo and all other patches are defined as background patches.

The patches are then converted to feature vectors which consist of two edge histogram vectors, $\mathbf{u}$ and $\mathbf{v}$. $\mathbf{u}$ is an concatenation of multiple overlapping bounding boxes each with a normalised 8 bin Histogram of Orientated Gradient (HoG), as described in Wang et al. [24]. $\mathbf{v}$ is a three level 8 bin Pyramid-HoG as described in Chen et al. [4]. The use of different types of edge information is useful for estimating a reliable embedding space in the subsequent stages. It is common for the descriptor vectors $\mathbf{u}$ and $\mathbf{v}$ to be high dimensional and also the accumulation of descriptors from a dense grid is likely to introduce noise.

3.2 Local Fisher Discrimination

Local Fisher (LF) [19] explores the idea of projecting feature data into two successive sub-spaces. The first sub-space is estimated by employing the dimensionality reduction technique Principal Components Analysis (PCA)[11] and the second sub space by the application of a supervised dimensionality reduction method Local Fisher Discriminant Analysis (LFDA)[18] on the PCA projected feature data. A brief review of LF is presented here.

A low dimensional embedding space is obtained from the high dimensional feature space by first estimating a PCA transformation separately for each of the two input feature vector types, $\mathbf{u}$ and $\mathbf{v}$.

Principal Component Analysis enables the dimensionality of the data to be reduced, while also preserving a high proportion of variation in the input signal [11]. For an input vector $\mathbf{u}_i$, the data projected into the low dimensional manifold estimated by PCA is written $\mathbf{u}'_i = D_u \mathbf{u}_i$, where D_u is the embedding transformation matrix corresponding to the eigenvectors derived from PCA. Similarly, $\mathbf{v}'_i = D_v \mathbf{v}_i$. It is experimentally demonstrated in LF [19] that separate estimation and use of D_v and D_u retains information more effectively. The overall output $\mathbf{x}_i$ from the first stage is the concatenation of the two sets of separately PCA projected histogram vectors: $\mathbf{x}_i = \{\mathbf{u}'_i | \mathbf{v}'_i\}$.

LF combines the neighbourhood preserving property of Locality Preserving Projection (LLP) [8] with the traditional Fisher Discriminant Analysis (FDA)[6]. It is very common for a multi-class dataset to be multi-modal in nature, i.e to show a significant variation within class samples. LF captures this multi-modality within classes by constructing an affinity matrix A which estimates the neighbourhood characteristics of the dataset. A local scaling method [30] is used for the estimation of A, by choosing the n-th nearest neighbor and assigning individual scaling factors for samples from the same class.

The width in class, S^W, and between class, S^B, scatter matrices in traditional FDA [6] are weighted with the affinity matrix A such that the far apart in-class samples do not contribute to the estimation.

$$S^W = \frac{1}{2} \sum_{i,j=1}^{n} A^w_{i,j} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^t \quad (2)$$

$$S^B = \frac{1}{2} \sum_{i,j=1}^{n} A^b_{i,j} (\mathbf{x}_i - x_j)(\mathbf{x}_i - \mathbf{x}_j)^t \quad (3)$$

where

$$A^w_{i,j} = \begin{cases} A_{i,j}/n_c & \text{if } y_i = y_j = c \\ 0 & \text{if } y_i \neq y_j \end{cases} \quad (4)$$

$$A^b_{i,j} = \begin{cases} A_{i,j}\left(\frac{1}{n} - \frac{1}{n_c}\right) & \text{if } y_i = y_j = c \\ \frac{1}{n} & \text{if } y_i \neq y_j \end{cases} \quad (5)$$

Here, n_c is the number of samples in class c and n is the total number of samples. The transformation matrix T_{lfda} can then be defined as

$$T_{lfda} = \arg\max tr\left(\left(T^t S^W T\right)^{-1} T^t S^B T \right) \quad (6)$$

where $T \in \mathbb{R}^d \times \mathbb{R}^m$. Similar to FDA[6], the estimation of T_{lfda} is achieved by representing the above as a generalized eigenvalue problem, $S^B \varphi = \lambda S^W \varphi$, where $\{\varphi_i\}$ and $\{\lambda_i\}$ are the eigenvectors and eigenvalues of this system. The final projection into the embedding space characterized by LFDA can be written as

$$\mathbf{z}_i = T^t_{lfda} \mathbf{x}_i \quad (7)$$

The similarity measure between any two observations i and j is then given by the Euclidean distance between the LFDA transformed vectors of each observation

$$D(i,j) = |\mathbf{z}_i - \mathbf{z}_j| \quad (8)$$

3.3 Logo Localisation Using LF

To localise the logo patches, the training set of logo (l) and background (b) patches is used, as defined in section 3.1. At this stage the vehicle class label is not used. The training data is used to estimate the matrix T^t_{lfda} that transforms the feature vectors ($\mathbf{x}_i$) to their representation in the embedded space, $\mathbf{z}_i$. Let $\{\mathbf{z}^l_i\}$ and $\{\mathbf{z}^b_i\}$ be the sets of $n(l)$ and $n(b)$ training vectors in the embedded space for logo and background patches, respectively. A previously unseen test vector, $\mathbf{z}^*$, will be classified as either logo or background using a k-nearest neighbour basis:

$$\mathbf{z}^* = \begin{cases} \text{logo} & \text{if } d_l < d_b \\ \text{background} & \text{otherwise} \end{cases} \quad (9)$$

where

$$d_{l,b} = \min_{\mathbf{z}_i \epsilon (\{\mathbf{z}^l_i\}, \{\mathbf{z}^b_i\})} (|\mathbf{z}_i - \mathbf{z}^*|) \quad (10)$$

Using the above method on the ensemble of patches within an RoI, each of these patches categorised as a 'logo' is then classified into one of the N manufacturer logos, which are then input into a voting process to provide a final estimate for this vehicle. This is described in the next section.

3.4 Logo Classification Using LF

Hence, for a given RoI, n patches are categorised as 'logo'. If $n > 0$, a LF-based multi-class classifier is used in an analogous manner to assign a predicated class y_i to each logo patch, where $y_i \in \{1, \ldots, N\}$, where there are N categories corresponding to the different vehicle manufacturers. (Otherwise, if $n = 0$, no suitable patches are available and the classification cannot proceed.) The overall manufacturer class assigned to the RoI, $y_{\max}$, is the class to which the largest number of individual logo patches belongs, as follows:

$$y_{\max} = \max_{1 \le j \le N} \left(\sum_{i=1}^{n} \delta(y_i, j) \right) \quad (11)$$

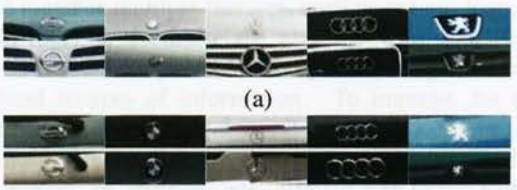
$$\delta(\alpha, \beta) = \begin{cases} 1 & \text{if } \alpha = \beta \\ 0 & \text{Otherwise} \end{cases} \quad (12)$$

If the voting process does not result in an outright winner, then a second voting procedure is adopted. For each logo patch there is a logo confidence measure which is represented by the value of d_l. For all the equal top ranking classes, from the first vote, their corresponding patches' d_l values are summed together, with the class which has lowest cumulative value becoming the overall winner.

4 Experimental Results and Analysis

4.1 Dataset

The data used was captured for a period of three month using full 1080p High Definition colour video camera, positioned at the gateway of a closed loop car park, at 30 fps. The footage captured vehicles entering and exiting the car park at different velocities, trajectories and under varying environmental conditions, such as lighting. The footage were segmented to extract the objects of interest, which was then processed to acquire the LP coordinates for the vehicles in this study.



(a)

(b)

Figure 2: Examples of Vehicle Logo Patches. a) Logo Frontal View, and, b) Logo Rear-View

Five common classes of vehicles, namely: Nissan, BMW, Mercedes (Merc), Audi, and Peugeot (PG) (example of the logos are given in Figure 2) were selected for experimentation. For each class, we choose 30 training images and 20 testing images. The training data is excluded from the testing data in order to test the true performance of the system. For each image the logo's RoI was located by using the LP position, then the RoI was subdivided into patches, which results in 18 patches per RoI. Therefore there were 2700 training and 1800 testing samples. The combined feature vector for each sample, before dimensional reduction, is 2248 components.

4.2 Logo Localisation Analysis

		Ground Truth	
		Logo	Background
Predicted	Logo	86%	14%
	Background	9%	91%

Table 1: Confusion Matrix of Logo and Background Patch Classification Results

The accuracy of the vehicle logo localisation was validated against manually labelled ground truth data. The results in Table 1 shows our method was able to achieve 86% accuracy for classifying logo patches and 91% for correct background patches. If only unsupervised PCA is used then the results decrease to 80% for logo and 83% background patches.

The 86% correct logo classification actually means that 97% of all testing RoIs would have at least one correctly predicted logo patch which could be forwarded to the logo manufacturer classification stage.

		Ground Truth				
		Nissan	BMW	Merc	Audi	PG
Predicted	Nissan	80%	5%	0 %	0%	15%
	BMW	0%	100%	0 %	0%	0%
	Merc	5%	21%	69%	5%	0%
	Audi	0%	0%	0%	100%	0%
	PG	0%	5%	0%	5%	90%

Table 2: Confusion Matrix for Logo Classification Results Using Badge Patch

4.3 Logo Classification

The system was trained using only the ground truth logo patches, the trained model was tested using previously unseen classified logo patches. Table 2 shows the confusion matrix of the logo classification results using the correctly classified logo patches only (lpo). The main diagonal shows the percentage of correctly classified manufacturer class. When all of the predicted logo patches (plp) are used, including the background patches that have been incorrectly classed as logo patches, the overall classification rate of 85.56% is obtained, as shown in Table 3. This 85.56% value illustrates the performance of the system in a real life environment.

Table 3 further demonstrates that the use of the Local-Fisher learned metric significantly improves the performance of the system, compared to using only the principal components of the original feature vector.

Table 3 also shows that HoG is superior to pHoG in this case, However the identical performance of just using HoG and the combination of both HoG and pHoG is caused by the inter-dependence of two sets of features. In this situation pHoG is not contributing enough information to improve performance when using the combined features.

Type	HoG		pHoG		HoG + pHoG	
	lpo	plp	lpo	plp	lpo	plp
PCA	70.4%	69.7%	68.04%	68.04%	70.41%	69.7%
PCA+LF	87.62%	85.67%	80.11%	79.8%	87.62%	85.67%

Table 3: Logo Classification Success Rate Using Different Features

5 Conclusion

This paper has presented a novel vehicle logo localisation and recognition process, using features composed of local histograms of gradients and Local Fisher Discrimination Analysis to obtain a more effective metric. The proposed method has been demonstrated on a data set that exhibits characteristics of the logo recognition problem in a real world scenario where there is no heavy reliance on the vehicle's viewpoint. Thus the method could be applied to both frontal and rear vehicle views. The results achieved by the process are not directly comparable to those from recently published techniques as these only concentrate on the frontal views of the vehicle, and the data used are captured in controlled environments, such as in the studies of Wang et al. [25] and Psyllos et al. [22]. As such, our results provide a benchmark for techniques for logo recognition on medium-view CCTV data in a video surveillance environment.

This research forms part of an overall project which aims to fuse together multiple features, such as colour, logo, shape etc., in order to maximise the certainty with which vehicle re-identification can be achieved, by combining relatively uncorrelated sources of information. To improve the system performance, one area that could be investigated is the use of additional features that provide uncorrelated indications of logo identify. In addition, currently the size of the RoIs have been predefined to accommodate the expected range of scales. Therefore the method is tolerant to changes in scale. A further improvement could be to scale the image by the license plate size, therefore making the method scale invariant. To further test the potential for learning, a greater quantity of testing data would be required, as to date the size of the training dataset is relatively modest (one hundred and fifty example).

References

[1] B Bhanu and M Kafai. Dynamic Bayesian Networks For Vehicle Classification in Video, November 23 2012. WO Patent 2,012,159,109.

[2] C M Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, USA, January 1996.

[3] T Burkhard, AJ Minich, and C Li. Vehicle Logo Recognition and Classification: Feature Descriptors vs. Shape Descriptors. *Stanford University, EE368 Final Project,* 2011.

[4] Z Chen and T Ellis. Multi-shape descriptor vehicle classification for urban traffic. In *International Conference on Digital Image Computing Techniques and Applications (DICTA), 2011,* pages 456–461. IEEE, 2011.

[5] L Dlagnekov and S Belongie. Recognizing cars. *University of California San Diego, Tech. Rep. CS2005-0833,* 2005.

[6] R A Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics,* 7(2):179–188, 1936.

[7] Y Freund and RE Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.,* 55(1):119–139, 1997.

[8] X He and P Niyogi. Locality Preserving Projections. In *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, 2004.

[9] U Iqbal, SW Zamir, MH Shahid, K Parwaiz, M Yasin, and MS Sarfraz. Image based vehicle type identification. In *International Conference on Information and Emerging Technologies (ICIET), 2010,* pages 1–5. IEEE, 2010.

[10] T Joachims. Making large-scale support vector machine learning practical. In *Advances in kernel methods*, pages 169–184. MIT Press, 1999.

[11] I Jolliffe. *Principal Component Analysis*. John Wiley Sons, Ltd, 2005.

[12] M Kastinger, M Hirzer, P Wohlhart, PM Roth, and H Bischof. Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2288–2295. IEEE, 2012.

[13] H Lee. Neural network approach to identify model of vehicles. *Advances in Neural Networks-ISNN 2006*, pages 66–72, 2006.

[14] W Li and L Li. A Novel Approach for Vehicle-logo Location Based on Edge Detection and Morphological Filter. In *Second International Symposium on Electronic Commerce and Security, 2009. ISECS'09*, volume 1, pages 343–345. IEEE, 2009.

[15] Y Liu and S Li. A vehicle-logo location approach based on edge detection and projection. In *Vehicular Electronics and Safety (ICVES), 2011 IEEE International Conference on*, pages 165–168. IEEE, 2011.

[16] Y Liu and J Rong. An efficient algorithm for local distance metric learning. In *Proceedings of AAAI*, 2006.

[17] W Lu, H Zhang, K Lan, and J Guo. Detection of vehicle manufacture logos using contextual information. *Computer Vision–ACCV 2009*, pages 546–555, 2010.

[18] S Masashi. Local Fisher discriminant analysis for supervised dimensionality reduction. In *Proceedings of 23rd International Conference on Machine Learning*, pages 905–912. ACM, 2006.

[19] S Pedagadi, J Orwell, SA Velastin, and B Boghossian. Local Fisher Discriminant Analysis for Pedestrian Reidentification. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 23-27 2013.

[20] VS Petrovic and TF Cootes. Analysis of Features for Rigid Structure Vehicle Type Recognition. In *British Machine Vision Conference*, volume 2, pages 587–596, 2004.

[21] A Psyllos, CN Anagnostopoulos, and E Kayafas. Vehicle model recognition from frontal view image measurements. *Computer Standards & Interfaces*, 33(2):142–151, 2011.

[22] A P Psyllos, C-NE Anagnostopoulos, and E Kayafas. Vehicle logo recognition using a SIFT-based enhanced matching scheme. *IEEE Transactions on Intelligent Transportation Systems*, 11(2):322–328, 2010.

[23] KT Sam and XL Tian. Vehicle Logo Recognition Using Modest AdaBoost and Radial Tchebichef Moments. In *International Conference on Machine Learning and Computing.*, 2012.

[24] S Wang, M Lewandowski, J Annesley, and J Orwell. Reidentification of pedestrians with variable occlusion and scale. In *International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1876–1882. IEEE, 2011.

[25] Y Wang, Z Liu, and F Xiao. A fast coarse-to-fine vehicle logo detection and recognition method. In *IEEE International Conference on Robotics and Biomimetics, ROBIO 2007*, pages 691–696. IEEE, 2007.

[26] KQ Weinberger and LK Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

[27] E Xing, A Ng, M Jordan, and S Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, volume 15, 2003.

[28] W Yang, G Yang, X Zheng, and L Zhang. An improved vehicle-logo localization algorithm based on texture analysis. In *International Conference on Computer Science and Information Processing (CSIP), 2012*, pages 648–651,. IEEE, 2012.

[29] Y Ying, K Huang, and C Campbell. Sparse Metric Learning via Smooth Optimization. In *NIPS*, pages 2214–2222, 2009.

[30] L Zelnik-Manor and P Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2004.

[31] B Zhang and Y Zhou. Vehicle Type and Make Recognition by Combined Features and Rotation Forest Ensemble. *International Journal of Pattern Recognition and Artificial Intelligence*, 26(03), 2012.

Evaluation of Bayesian and Dempster-Shafer Approaches to Fusion of Video Surveillance Information

S. Wang, J. Orwell and G. Hunter
Digital Imaging Research Centre,
Kingston University
London, UK
Email: k1047774@kingston.ac.uk, james@kingston.ac.uk and G.Hunter@kingston.ac.uk

***Abstract*—This paper presents the application of fusion methods to a visual surveillance scenario. The range of relevant features for re-identifying vehicles is discussed, along with the methods for fusing probabilistic estimates derived from these estimates. In particular, two statistical parametric fusion methods are considered: Bayesian Networks and the Dempster Shafer approach. The main contribution of this paper is the development of a metric to allow direct comparison of the benefits of the two methods. This is achieved by generalising the Kelly betting strategy to accommodate a variable total stake for each sample, subject to a fixed expected (mean) stake. This metric provides a method to quantify the extra information provided by the Dempster-Shafer method, in comparison to a Bayesian Fusion approach.**

Keywords*—Dempster-Shafer; evaluation; vehicle; fusion; Bayesian;

I. Introduction

In recent years, the amount of surveillance video data has substantially increased, and this trend appears set to continue. Norris and Armstrong [1] estimate that in a typical urban area, a person could be observed from as many as 300 cameras each day. With such a significant quantity of surveillance data, some tasks are impractical to accomplish with purely manual analysis. This implies a need for processing systems which are capable of analysing video data in an automatic or semi-automatic way.

There are many challenges in Automated Video Surveillance (AVS); two main problems that any system needs to overcome are, firstly, the uncertainty associated with the measurement of the object, which is due to the large variability of factors such as illumination, pose, and in-class variations of shape. The second problem is the identification of a suitable method to fuse the information from different sources, in order to extract various types of inferences, depending on the situation or the user needs.

Multisensor Information Fusion (MIF) is the study of techniques to overcome the latter problem, to achieve more accurate inference, than could be achieved by the use of a single sensor alone [2]. The inferences that can be made, using these techniques, range from simple estimates the identity of a certain entity, through to complex inferences about current or future relationships between multiple entities and the events with which they are involved.

As the types of information in this report are noncommensurate, the data must be fused at a *decision level* [3]. Decision-level fusion consists of merging information at a higher level of abstraction, combining the results from multiple algorithms to yield a final 'fused' decision [4]. The data is therefore processed separately by multiple algorithms, e.g. to identify and classify observed entities and events, based on their different features. The resulting information is then combined, applying a chosen set of decision rules to obtain an overall inference. Due to the uncertain nature of the input data, the inference is often expressed in a probabilistic form. Two well-known types of decision rules are Bayesian inference and the Dempster-Shafer (DS) theory, and these could both be applied to a specific fusion scenario, e.g. combining multiple cues extracted from surveillance video.

To the best of the authors' knowledge, there are no previous proposals for how to conduct a *performance evaluation* for these two fusion methods. One reason for this absence is that the two methods express their results in different forms: while a Bayesian analysis uses probabilities alone, DS encompasses the additional concepts of *belief* and *plausibility*. A neutral basis must be established to compare the two. Thus, this paper investigates how the standard theory for evaluation of Bayesian estimates can be extended to accommodate the DS methodology, which is arguably a generalization of Bayesian analysis.

The remainder of the paper is organised as follows. In Section II a review of literature dealing with information fusion within the video surveillance domain is conducted. Section III introduces a surveillance scenario and the uncertainties in the extractable features from this scenario. In Section IV, an introduction to Bayesian and Dempster-Shafer theory is given and in Section V the generalised evaluation framework is introduced. Some experiments are conducted in Section VI with the conclusion presented and further works proposed in Section VII.

II. Literature Review

Unlike other research fields where MIF have been used extensively, such as fault diagnostics [5], computer intrusion detection [6] and a range of military applications [3], MIF is a relatively new techniques for automated video surveillance, with comparatively few publications.

Snidaro *et al* [7] weighted the confidence measure from different sensors monitoring same scene to produce more accurate pedestrian tracking results. Similarly, Kumar *et al* [8] applied Fuzzy Logic modelling to produce a 'belief mass' for each of the sensors, to address the pedestrian tracking issue in varied illumination conditions. Recently, Torabi et al. [9] advanced Kumar's idea by applying fusion techniques to track multiple people walking in close proximity. In this work, data from colour-based and thermal sensors are fused to achieve the task of tracking peoples in both indoor and outdoor environments in varied lighting conditions.

For identification of types of vehicles in surveillance data, Sun *et al* [10] used 'inductive loop signatures' system that was build with some specialised equipment and reported an accuracy of 90%, on a three-category problem, .Sumalee et al. [11] applied similar ideas and applied it to the problem of vehicle re-identification within video. The author also introduced other vehicle features such as colour, shape and size that are derived from video image data using different image processing techniques. These features are then fused together using a probabilistic fusion technique to provide a probabilistic measure for the re-identification decision. The overall re-identification accuracy was about 54%, which represents the current state-of-the-art.

III. Problem Introduction

The paper reports on aspects of an investigation using a surveillance test bed with high-definition CCTV cameras, monitoring vehicles entering and exiting a car park over an extended period of time. The investigation concerns the feasibility of automatic inference methods that require the fusion of multiple features, and also fusion of data from multiple segments of video, possibly from different cameras. Examples of this type of inference task are the following:

1) 'Has this vehicle been here before?'
2) 'Is the car park full?'

To make inferences such as these, a key capability is the capacity to re-identify a vehicle, possibly from the reverse angle. There will often be some degree of uncertainty associated with the estimation of identities of any observations. One aspect of the investigation is to quantify and minimise the extent of this uncertainty, in both the input and output of the fusion process. The following subsections review the relevant forms of input data, and then in section IV methods to fuse these uncertainties are introduced.

A. License Plate

A vehicle's license plates are its most discriminating features, however automatic recognition of the characters is a challenging task even when specialised equipment is used. There are many factors that affect the accuracy, which can be categorised into plate variation and environmental variations as outlined by Du *et al* [12]. In this review paper, they also outline the accuracy rates for the current state of the art techniques, which are typically between 90 - 97% depending on plate format. The majority of the technique described are designed for vehicles that are almost stationary, and the accuracy rate decreases when moving-image (video) data is used.

Uncertainties are also introduced when comparing two vehicle license plates. A popular method that is used when comparing two strings is the Hamming Distance [13], however this requires the strings to be of the same size, while real and estimated license plate strings can have variable lengths, so a more sophisticated metric is required. Alternatively, fuzzy string searching methods which have been adopted for DNA comparisons can be used. A popular metric is the Damerau - Levenshtein distance [14], which measures the number of edit operations needed to make the strings identical by allowing 4 different edit operations; insertion, substitution, deletion and transposition. Since all edit operations are assign the same weight, true-negative results could score identically with true positives results, thus a degree of uncertainty still exists in the measured metric.

B. Vehicle Manufacturer's Logo

Vehicle logo classification is a type of shape classification and the logo class can be used as feature to assist with the vehicle recognition problem. Typically, a reference point is required, such as position of the license plate, to find and extract the logo region and estimate the category. Another challenge is the uncontrolled environment: the size and orientation of the logo can vary. These factors contribute to the uncertainty associated with the results.

Some additional uncertainty is also introduced by the classification approach. Two techniques have been previously proposed: multi-class classification or one-against-all binary classification. The challenge with multi-class classification, as the number of possible logo categories increase, is the training or the definition of a space where clusters for the different categories are sufficiently far apart. Thus, the correct match would be the cluster with shortest distance between the query sample and the correct cluster. However due to the similarity between the different logos this is not always possible. The current best result for classify 5 different logo classes is achieved by Wang et al [15], namely about 84% success for 5 different categories. The challenge with one-against-all binary classifier is the requirement of having for n classifiers for the n categories, therefore for a given logo there will be multiple estimates that need to be resolved. For similar logos the estimate the certainty score associated with the output from the multiple classier might be very similar. The current best results using one-against-all methods were achieved by Psyllos et al. [16] averaging 91% overall classification success for 10 categories, but this bias for front views only in controlled lighting conditions.

C. Vehicle Colour

Colour is an important cue in the surveillance context, for example for the re-identification of pedestrians. The re-identification of vehicles, based on colour, is less extensively researched, for perhaps three reasons. First, colour alone

could not re-identify the vehicle as many many vehicles share the same colour. Second, its perceptibility degrades significantly in deteriorated lighting conditions. Finally, an appropriate colour model is not straightforward to define: sometimes a single value is useful, but in other situation colour can be described using several colour channels. The overall colour of a vehicle is difficult to summarise as windows and wheels have large contrasting colour schemes compared to the body. One possible solution is suggested by Psyllos *et al* [17], where the authors collected a RGB histogram for a range of patches on the vehicle and chose the peak of each of the components to represent the overall colour.

The variation of colour observations wuth respect to changes in environmental conditions can be mitigated to some extent by using the HSV space, as this is more tolerant [11] to such change. Considering these challenges, the measurements from two 'vehicle colour sensors' could be used to produce an estimate of the probability that these two observations refer to the same vehicle, or that the two observations refer to vehicles having the same colour model (which excludes the problem of estimating prior probabilities of the various colour models).

D. Vehicle Shape Classification

Vehicle shape classification has commanded significant attention from researchers, as reported by Kanwal *et al* [18], the review concentrated on the various software based vehicle classification techniques, which have classification accuracies between 82% – 95%. The authors concluded that the best classifier was a hybrid system based on a Dynamic Bayesian Network classifier, however, direct comparison is between approaches is inappropriate as definition of the vehicle classes are different between the methods reviewed. These results demonstrate that some uncertainty is a consequence of the lack of a global definition of the vehicle shapes or classes.

Like many shape recognition problems, vehicle shape classification is restricted in several ways. The two main restrictions are shadowing and viewing angle. Shadows can have the effect of increasing the apparent size of the vehicle, and effective shadow removal is necessary. For most techniques, the vehicle orientation (in relation to the camera) is critical for the shape to be classified correctly.

E. Spatio-temporal Information

Spatio-temporal information can be used to estimate the probability that two observations refer to the same vehicle. Three such cues are: the time of the observation, the gate the vehicle used to enter (or exit), and the characteristics of the trajectory associated with each driver, which may be distinguishable over repeated observations. The first two cues will have negligible measurement noise; the third is a more complex measurement process. .

IV. Statistical Parametric Fusion Methods

In this section we will cover the formulation of the statistical parametric fusion methods that is under investigation.

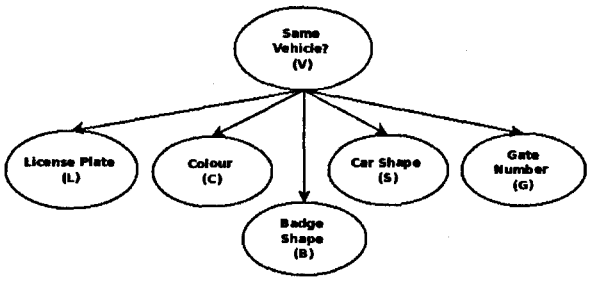


Fig. 1. Simple Bayesian Network for integration of multiple visual surveillance cues.

A. Bayesian Models

Bayes rule [19] lies at the heart of many data fusion methods. Using Bayes' theorem , we assume h_i is a hypothesis about a state, taking values in the set of hypotheses $H = h_1, ...h_n$, exactly one of which is 'true', and the remainder being 'false'. The prior probabilities, $P(h_i), i = 1, ..., n$ constitute the prior probability mass function of the hypotheses h_i:

$$0 \leq P(h_i) \leq 1 \text{ and } \sum_{i=1}^{n} P(h_i) = 1 \quad (1)$$

Often, the hypothesis with the highest prior probability will turn out to be the 'true' one. However, a more accurate estimation of the state can be made by incorporating some relevant 'posterior' evidence, x. It is assumed that the h_i are distributed according to the class-conditional probability distribution function $P(x|h_i)$ [20]. Therefore, given the prior probability and the class conditional probability, the posterior probability could be calculated using the Bayes' formula:

$$P(h_i|x) = \frac{P(h_i)P(x|h_i)}{\sum_{j=1}^{n}(P(h_j)P(x|h_j))} \quad (2)$$

The denominator is the "evidence factor", which normalises the posterior probabilities so they will sum to one.

An extension of Bayesian inference is the Bayesian Network [20], which is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph, as modelled in Figure 2. Each leaf node represents the evidence from each of the features supporting the querying root node, and the root node represents the query of which the system tries to establish the truth. Each leaf node will have a conditional probability matrix which is associated to the relationship linking the evidence to the root node. These conditional probabilities could be generated using data obtained directly from observations made about the environment. Furthermore, these can be updated over time to improve the accuracy of the model.

Equation 2 needs to be adopted to allow the calculation of joint evidence, as the class-conditional probability distribution function for a joint set of evidence, $P(x_1, ...x_c|h_i)$ in general does no have an analytic solution and it may not be possible to numerically evaluate it for all instances of the evidence. To overcome these issues, the assumption is made that the

leaf nodes are conditionally independent, therefore their co-occurrences can be calculated as a simple multiplication. Equation 2 is now transformed into:

$$P(h_i|\cap_k^c x_k) = \frac{P(h_i)\prod_{i=1}^{k} P(x_k|h_i)}{\sum_{j=1}^{n}(P(h_j)\prod_{i=1}^{k} P(x_k|h_j))} \quad (3)$$

B. Evaluation of Bayesian Models

For a discrete set of outputs, Bayesian models can be evaluated using a Kelly [21] betting criterion. This consists of placing a nominal "stake" on each possible output, in proportion to the odds estimated, using the available observations. The pay-off can be defined using the prior (fair) odds. The doubling rate is proportional to mean information gain: the average amount of information provided by the observations can be infered from the outcome of the betting strategy.

C. Dempster Shafer (DS) Models

The Dempster-Shafer theory was introduced by Dempster [22] and further developed by Shafer [23]. Like Bayesian probability theory, DS also deals with subjective probability. Therefore, DS could be seen as an extension of Bayesian approach to probability, as the latter does not explicitly model a lack of knowledge (or 'ignorance').

Let $\Theta = h_1, .., h_n$ be a collection of mutually exclusive set of hypotheses to a given query, called the frame of discernment. A basic belief assignment (bba) is a function of Θ from $2^\Theta \rightarrow [0,1]$ that assigns a mass of belief to each subset A of the frame of discernment Θ, satisfying

$$m(\emptyset) = 0 \text{ and } \sum_{A\subseteq\Theta} \quad (4)$$

The basic belief mass $m(A)$ represents a measure of the belief that is assigned to the subset $A \subseteq \Theta$, given the available evidence, and that cannot be committed to any strict subset of A. All the assigned probability sums to unity and there is no belief in the empty set ($\emptyset$).

There are two additional facets associated with each bba, which are all functions of $2^\Theta \rightarrow [0,1]$; firstly, the Belief measure, $Bel(A) = \sum_{B\subseteq A} m(B)$, which represent the exact support for A and secondly, the Plausibility measure, $Pl(A) = \sum_{A\cap B\neq\emptyset} m(B)$, represents the possible support for A. $[Bel(A).Pl(A)]$ constitute the the interval of support to A and forms the lower and upper bounds of the probability to which A is supported.

To combine the different sources of information, Dempster's combination rule is used. The successful application of this rule assumes that the different bba are independent pieces of evidence and uses the orthogonal sum to combine the multiple belief structures. For two bba m_1 and m_2, the combination rule is as follows:

$$[m_1 \oplus m_2](\theta) = \frac{\sum_{A_i\cap B_j=\theta} m_1(A_i)m_2(B_j)}{1-K} \quad (5)$$

where:

$$K = \sum_{A_k\cap B_m\neq\emptyset} m_1(A_k)m_2(B_m) \quad (6)$$

Demspter's rule of combination is both commutative and associative [24]: these two properties mean that evidence could be combined ($\oplus$) iteratively using the Equation 5 and in any order in pair-wise way .

$$m_1 \oplus m_2 \oplus ... \oplus m_k \quad (7)$$

D. Evaluation of DS Models

There is no single universally agreed procedure for the evaluation of a DS model. Where DS is applied within a 'goal-oriented' environment, the 'true' hypothesis is chosen based on its corresponding belief and plausibility. One method is to choose the hypothesis with the highest belief and plausibility. Alternatively the choice can be made by selecting the highest belief and the lowest interval of e.g. (plausibility - belief). However, these are all heuristics that do not appear to accommodate all the features of the DS representation. Furthermore, when the number of hypotheses is limited (≤ 2) then it may be difficult to find a outright winner, when using these methods.

V. A Generalised Evaluation Approach

In section IV-B it was discussed how the information provided by the posterior estimates acted to reduce the overall uncertainty about the state: the effectiveness (or accuracy) of any given Bayesian model can be evaluated by measuring the reduction in uncertainty relative to prior model. This is equivalent to measuring the *side information* that the measurements provide about the system. These probabilistic measurements can be combined in many different ways, assuming independence or some model of co-dependency, for example using parametric or explicit models. In all cases, however, the Bayesian model outputs an overall probability per hypothesis, and the accuracy of any given model can be evaluated by measuring the expected log probability or entropy of the correct hypothesis.

$$H(X) = \langle -\log p(x) \rangle \quad (8)$$

$$\approx \frac{-1}{n}\sum_{i=1}^{n} \log p(x_i) \quad (9)$$

The information gain is proportional to the mean doubling rate, $\bar{W}$:

$$\bar{W} = \log\left\langle \frac{1}{p(x)} \right\rangle \quad (10)$$

$$= \langle -\log p(x) \rangle \quad (11)$$

$$\approx \frac{-1}{n}\sum_{i=1}^{n} \log p(x_i) \quad (12)$$

It is not straightforward to apply this Bayesian, information-theoretic evaluation method to a DS model. That type of model contains two scalar quantities for each hypothesis: the belief and the plausibility. Neither of these quantities directly relate to a Bayesian probability, and so it is not clear how to apply the various information-theoretic results noted above. Nevertheless, the sections below describe a context where the log optimal doubling rate can be used to evaluate additional meaning that is expressed by the belief and plausibility, provided by the DS model.

The important characteristic that distinguishes the {belief, plausibility} pair, from the {probability} singleton, is that the former pair can encode, using their difference, an expression about the uncertainty of the estimate. Thus, if in a certain case the DS model provides a pair {0.05, 0.95}, how can this be distinguished from the pair {0.45, 0.55}, and can either of these be evaluated against a Bayesian predicated estimate of 0.5?

It is proposed that the contribution of these extra indications, provided by DS, can be quantified using an appropriate generalization of Eqn. 10. That equation represents the standard Bayesian evaluation, using the expectation of the log posterior over a sample set. There, the contribution of each element in the sample set is implicity scaled to one.

The proposed generalization, to accommodate the extra indication output of DS, is to assign a weight α_i to each sample, subject to the constraint that $\langle \alpha_i \rangle = 1$. Thus the evaluation metric is then written as this:

$$\bar{W} = \frac{-1}{n} \sum_{i=1}^{n} \alpha_i \log p(x_i) \tag{13}$$

If these weights are given random values, e.g. uniformly in the interval between 0 and 2, then it can be shown that the measurements obtained from Eqn. 13 are unchanged from those obtained from Equ. 10. However, allowing these weight to be interpreted as a 'degree of confidence in the estimate'. For those samples about which the model estimate may be considered 'more accurate', the intention is to assign a larger scaling weight, and for those estimates about which there is a greater degree of uncertainty, the intention is to assign a smaller weight, thus fulfilling the overall constraint on the weights that their expectation is unity.

This creates the opportunity to define an evaluation protocol that can be used for both Bayesian and DS.

VI. Experiments

A. Toy Example

The proposed evaluation procedure is applied to a toy example: an estimated model for a two-horse race with information provided by two sources: measurement of the Horses' attributes, and measurement of the Jockeys' attributes. In this scenario, the evaluation metric is mean percentage winnings (or losses) per race, following a Kelly Betting strategy, using the estimated model. This has a direct relation to the informative capacity of the model. Fundamentally, this percentage will depend on the relationship between three probabilistic models. The first model is the real (actual) probabilities that determine the outcome of the race. The second model determines the bookmakers' odds, which is used to calculate the pay-off after the outcome of each race. The third model is the estimated model, that represents a subjective understanding of likely outcome of each race, using the two sources.

B. Kelly Betting with DS-dependent stake

The usefulness of the estimated model can be measured by using it in a Kelly betting strategy, which is the log optimal strategy: the stake for each outcome is placed in proportion to the model prediction (estimated probability) for it. Conventionally, for each race, the sum of these bets (i.e. the total stake) can be fixed at an arbitrary quantity; the *total stake* for each race can be fixed at an arbitrary value, e.g. 1, and the log of the winnings accumulated. However, to accommodate the extra information provided by DS, this total stake is varied depending on the interval between the plausibility and support. Over the evaluation sample, the expected (mean) stake is constrained to be equal to the stake used for the simple evaluation.

As a starting point, let all three models be identical: in this scenario, the expected outcome of both the fixed-stake and DS-dependent strategies is to 'break-even', both with standard Kelly betting (fixed stake size) and the generalised Kelly strategy, where the total stake each race is allowed to vary.

The above outcome is observed for any joint distribution between sources, i.e. for both correlated and anti-correlated distributions of 'Horse' and 'Jockey' measurements. However, the DS analysis does treat these two cases differently: divergent estimates between the two sources will result in a larger 'unknown state' than the case in which they are in agreement.

C. Perturbing the Prior Estimate

In this experiment the prior information provided to the real and bookmakers' model is the same, but the prior information from the sensors to the estimated model is perturbed from the real model, to simulate some imperfection in the available information. The perturbation takes the form of a percentage change to the estimated 'difference between means' that forms the model for generating each sensor measurement. The sign of the change is also generated randomly with equal probability.

Since the real and bookmakers' odds are still identical, Kelly betting using the estimated model will always result in losses; however, more succesful fusion strategies will reduce these losses, and the extent of the reduction can be used to evaluate the efficiacy of the fusion strategy, using this generalised Kelly betting process, in which a variable total stake is allowed for each case.

The DS fusion strategy provides a rationale for varying the total stake: when there is a large 'unknown state', the total stake can be reduced, and conversely when there is a small unknown state, a comparatively large total stake can be used. This strategy is repeated over 15,000 samples, at each level of estimated model perturbation, to compare the mean percentage loss from the DS strategy against the default (fixed total stake) alternative.

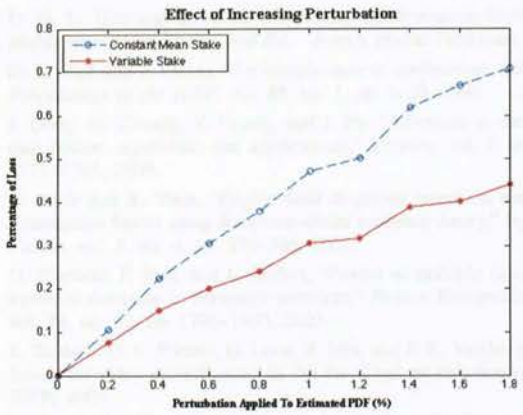


Fig. 2. Effect on Variation the amount of perturbation

The results of this simulation are plotted in Figure 2: this shows a clear advantage from the use of the DS fusion strategy, in that approximately one-third of the adverse effect of the model perturbation is removed, as a consequence of using DS outputs to determine the stake size. One explanation for this effect is that cases in which source estimates were disagreement are more likely to have been significantly effected by the perturbation, and so the consequential reduction in total stake reduced the effect of that more substantial inaccuracy in probability.

D. Application to Surveillance Fusion

The above scenario can be adapted to simulate the information fusion process that is required for visual surveillance scenarios, as outlined in Section III. For example, one of the key capabilities is the re-identification of vehicles, from a pairs of observation pair of vehicle. The above methodology can be used to evaluate the benefit of using the fusion method that exploits the DS outputs, using probabilistic sensor measurements of e.g. vehicle type, make, colour and license plate. In this scenario we would be expecting to maximize winnings, rather than minimize losses. In other words, the bookies odds would be the prior probability of correctly re-identifying the vehicles (without any sensor measurements) and the estimate would be expected to be significantly more certain.

It is worth emphasising that the utility of the proposed evaluation methodology is that it allows fusion methods to output a measure of the confidence in a particular estimate, and this confidence is then used to weight the importance of this estimate in the overall evaluation of the method accuracy. An overall constraint on the mean weight is imposed, to enforce like-with-like comparisons, and prevent acceptance of the the trivial zero-weight solution. The DS approach does provide a measure of confidence, via the support and plausibility, and so this can be used to generate a weighting.

It is important to examine the significance and utility of the proposal in the context of evaluation of automated video analytics systems. A frequent criticism of these systems is that they are unable to indicate when they are 'not sure'. Hence, this proposal fits well into that context: by requiring that a system also output a weight that is used to calculate the evaluation, the indication of certainty is incorporated, and in a straightforward manner, consistent with standard information-theoretic evaluation of 'side-information'. Furthermore the proposed strategy is identical to the standard information-theoretic evaluation, in the limiting case of when each weight is constant and equal.

Nevertheless, there are still several tasks that remain to be completed. There are various ways in which the DS output could be transformed into a single weight, and it is not yet clear which would be the most appropriate. One specific aspect requiring attention is the mechanism to enforce a fixed mean weight, over the test set ensemble. Another task is a more comprehensive evaluation over the range of possible perturbations, to verify that the proposed approach works in this range. It may be possible to obtain some theoretical results for this general case, too.

VII. Conclusion

This paper has examined the features, and methods for fusing them, that can be used for vehicle re-identification. This task is one of the fundamental building blocks that allows inferences about other, more complex and extended, queries to be attempted. Instead of using any single visual feature of the vehicle, which is restrictive and suboptimal in the final result, the aim is to use a statistical parametric fusion of multiple vehicle features to provide a more precise estimate. The main contribution of this paper is the development of an evaluation metric, based on the Kelly betting strategy to allow the direct comparison of Bayesian and Dempster Shafer methods for fusion. The metric accommodates the extra information provided by the DS model, by allowing a variation in the per-outcome stake, subject to a constraint on the expected overall stake. This accommodates fusion methods that provide an indication of the uncertainty of the fused estimate, such as the Dempster Shafer approach. It was shown, using a simple example, that under certain (broad) conditions the DS model provided an improvement in the mean log winnings, which are a fundamental information metric of the standard Bayesian evaluation, being proportional to the side-information provided by the observation. We also described how this simple example can be adapted to the surveillance scenario. Furthermore, it seems capable of being applied to the general case of fusion problems.

This paper has introduced the idea of adapting information fusion methods to the domain of video analytical and has outlined the potential benefits. Further work is required to fully integrate all of the features into the two fusion framework defined in this paper. Once the fundamental block of re-identifying the vehicle has been formalised, the framework would be extended to allow inferences of higher value surveillance information.

Acknowledgment

The authors would like to thank the U.K. Engineering and Physical Sciences Research Council and BAe Systems Ltd. for funding this programme of research.

References

[1] C. Norris and G. Armstrong, *The maximum surveillance society: The rise of CCTV*. Berg Publishers, 1999.

[2] D. D. L. Hall and S. A. H. McMullen, *Mathematical Techniques in Multisensor Data Fusion 2nd Ed.* Artech House Publishers, 2004.

[3] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[4] J. Dong, D. Zhuang, Y. Huang, and J. Fu, "Advances in multi-sensor data fusion: algorithms and applications," *Sensors*, vol. 9, no. 10, pp. 7771–7784, 2009.

[5] O. Basir and X. Yuan, "Engine fault diagnosis based on multi-sensor information fusion using dempster–shafer evidence theory," *Information Fusion*, vol. 8, no. 4, pp. 379–386, 2007.

[6] G. Giacinto, F. Roli, and L. Didaci, "Fusion of multiple classifiers for intrusion detection in computer networks," *Pattern Recognition Letters*, vol. 24, no. 12, pp. 1795–1803, 2003.

[7] L. Snidaro, G. L. Foresti, G. Luca, R. Niu, and P. K. Varshney, "Sensor fusion for video surveillance," in *7th Int. Conf. on Information Fusion*. IEEE, 2004.

[8] P. Kumar, A. Mittal, and P. Kumar, "Addressing uncertainty in multi-modal fusion for improved object detection in dynamic environment," *Information Fusion*, vol. 11, no. 4, pp. 311–324, 2010.

[9] A. Torabi, G. Massé, and G.-A. Bilodeau, "An iterative integrated framework for thermal–visible image registration, sensor fusion, and people tracking for video surveillance applications," *Computer Vision and Image Understanding*, vol. 116, no. 2, pp. 210–221, 2012.

[10] C. C. Sun, G. S. Arr, R. P. Ramachandran, and S. G. Ritchie, "Vehicle reidentification using multidetector fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 5, no. 3, pp. 155–164, 2004.

[11] A. Sumalee, J. Wang, K. Jedwanna, and S. Suwansawat, "Probabilistic fusion of vehicle features for reidentification and travel time estimation using video image data," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2308, no. 1, pp. 73–82, 2012.

[12] S. Du, M. Ibrahim, M. Shehata, and W. Badawy, "Automatic license plate recognition (alpr): A state of the art review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 311–325, 2013.

[13] R. W. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal*, vol. 29, no. 2, pp. 147–160, 1950.

[14] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, vol. 10, 1966, p. 707.

[15] S. Wang, S. Pedagadi, J. Orwell, and G. Hunter, "Vehicle logo recognition using local fisher discriminant analysis," in *5th International Conference on Imaging for Crime Detection and Prevention (ICDP-13)*. IEEE, 2013.

[16] A. P. Psyllos, C.-N. Anagnostopoulos, and E. Kayafas, "Vehicle logo recognition using a sift-based enhanced matching scheme," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 322–328, 2010.

[17] A. Psyllos, C. Anagnostopoulos, and E. Kayafas, "Vehicle model recognition from frontal view image measurements," *Computer Standards & Interfaces*, vol. 33, no. 2, pp. 142–151, 2011.

[18] Y. Kanwal, I. Arta, and J. Ali, "Comparative analysis of automatic vehicle classification techniques: A survey," *International Journal of Image, Graphics and Signal Processing*, vol. 4, no. 9, pp. 52–59, 2013.

[19] T. Bayes, D. W. Bunn, H. Raiffa, R. Schlaifer, and D. Von Winterfeldt, "An essay toward solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society of London*, vol. 53, 1984.

[20] F. V. Jensen, *An introduction to Bayesian networks*. UCL press London, 1996, vol. 210.

[21] J. L. Kelly, "A new interpretation of information rate," *Information Theory, IRE Transactions on*, vol. 2, no. 3, pp. 185–189, 1956.

[22] A. P. Dempster, "A generalization of Bayesian inference," DTIC Document, Tech. Rep., 1967.

[23] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976, vol. 1.

[24] R. R. Yager, L. Liu *et al.*, *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008, vol. 219.