Pedestrian Detection by Computer Vision

A thesis submitted in partial fulfilment of the requirements of Napier University for the degree of Doctor of Philosophy

Ivan Alaric Derrick Reading

June 1999

Abstract

This document describes work aimed at determining whether the detection, by computer vision, of pedestrians waiting at signal-controlled road crossings could be made sufficiently reliable and affordable, using currently available technology, so as to be suitable for widespread use in traffic control systems.

The work starts by examining the need for pedestrian detection in traffic control systems and then goes onto look at the specific problems of applying a vision system to the detection task. The most important distinctive features of the pedestrian detection task addressed in this work are:

- The operating conditions are an outdoor environment with no constraints on factors such as variation in illumination, presence of shadows and the effects of adverse weather.
- Pedestrians may be moving or static and are not limited to certain orientations or to movement in a single direction.
- The number of pedestrians to be monitored is not restricted such that the vision system must cope with the monitoring of multiple targets concurrently.
- The background scene is complex and so contains image features that tend to distract a vision system from the successful detection of pedestrians.
- Pedestrian attire is unconstrained so detection must occur even when details of pedestrian shape are hidden by items such as coats and hats.
- The camera's position is such that assumptions commonly used by vision systems to avoid the effects of occlusion, perspective and viewpoint variation are not valid.
- The implementation cost of the system, in moderate volumes, must be realistic for widespread installation.

A review of relevant prior art in computer vision with respect to the above demands is presented. Thereafter techniques developed by the author to overcome these difficulties are developed and evaluated over an extensive test set of image sequences representative of the range of conditions found in the real world.

The work has resulted in the development of a vision system which has been shown to attain a useful level of performance under a wide range of environmental and transportation conditions. This was achieved, in real-time, using low-cost processing and sensor components so demonstrating the viability of developing the results of this work into a practical detector.

Acknowledgements

My sincere thanks go my supervisors Dr. David Binnie and Professor Keith Dickinson for their guidance and encouragement.

David Binnie was particularly influential in the aspects of this work relating to the use of CMOS sensor technology and the design of image capture equipment described in Chapter 3. Professor Peter Denyer is also due my thanks for access to CMOS sensor technology.

Keith Dickinson supervised all the transportation aspects of this work and was also responsible for obtaining much of the funding which supported it. The work studying pedestrian crossing behaviour described in Chapter 2 was carried out as my part in his contract to examine the impact of Puffin crossings for the City of Edinburgh Council. He was also responsible for initiating the involvement in the Department of Transport's trial of volumetric crossings and most recently for setting up a Teaching Company Scheme to exploit the results of this research.

Special thanks are also due to Dr. Chuen Wan who was closely involved as a colleague in many aspects of this work. I'd particularly like to acknowledge his important roles in developing software and hardware for the Department of Transport trials, for contributing his knowledge of various 'C' compilers and for making his time-coding system available for the video capture process described in Chapter 7.

Professor Michael Hartley was of great assistance in providing motivation and encouragement and his role in the preparation of this thesis is also gratefully acknowledged.

I should also thank my colleagues Malcolm Thomson, Brian Stewart and Steven Webster for their contributions and companionship during various periods of this work. The technical assistance of Paul Chmielnik, Timor Dzinaj, Jamie Watson, Herbert Goh and Rupert Dickinson was also invaluable.

Finally I'd like to thank my wife Sarah for James, for Max and for pushing hard!

This work was largely funded by a grant from the EPSRC held by Professor Dickinson and Dr. Wan. It also benefited from valuable assistance with on-street testing contributed by the City of Edinburgh Council and the Department of Transport, all of which are gratefully acknowledged.

Contents

PAGE NUMBERING AS ORIGINAL

1 Introduction

1.1 Objectives

The last decade has seen a rapid expansion in the use of video monitoring systems enabled by technological and manufacturing advances which have resulted in ever improving performance at lower cost. Video monitoring in public areas is now commonplace, with prominent examples of its use including the provision of security in urban centres and the monitoring of transportation systems. The enormous quantity of data generated by these installations has led to a demand for automatic detection systems to analyse the video. These systems typically aim to identify important events for the attention of a human operator, or even to turn the received video into object-based information so that decisions can be made automatically. At present such video-based monitoring installations are still large scale and expensive propositions entailing the cost of not only the vision systems themselves but the often more significant costs of providing mounting masts (typically 30m high) and video cabling (often over several miles) to a central monitoring station.

The availability of ever increasing processing power and low-cost smart vision sensors has now reached the point that it is possible to contemplate the widespread use of autonomous, low-cost, embedded computer vision systems consisting of a camera and processing unit for standalone sensing tasks. This greatly increases the number of potential applications for vision-based detection systems. Performing image analysis locally has the benefit that the need for extensive video cabling would be removed as sensory output could be transmitted along low-bandwidth communication links or, if there is sufficient confidence in the results, passed directly to local (traffic) control units. Additionally, it means that detection tasks that were previously impossible or unreliable using other technologies may be soluble.

The work described in this thesis is an investigation into the feasibility of developing such an autonomous vision system for the detection of pedestrians in a transportation environment. In particular the task of detecting pedestrians waiting to cross the road at signal-controlled crossings is addressed with the principle objective of determining whether a vision-based detection system can be made sufficiently reliable and affordable, using currently available technology, to be suitable for widespread use in a traffic control system.

This is a challenging objective as although the use of computer vision systems in indoor production situations, where lighting and environmental conditions can be controlled, is now rapidly maturing the achievement of reliable outdoor operation is still an area of active research. Indeed, it is noticeable that the starting point for the design of production systems is usually the design of controlled lighting systems to eliminate many of the difficulties that were faced in this work.

In contrast to much previous work in computer vision this work is directed at a real, outdoor application where the detection system must operate under unconstrained environmental and transportation conditions. It will be seen that in earlier comparable computer vision work at least some of the consequent difficulties of these conditions have been removed by simplification of the problem domain. The most important distinctive features of the pedestrian detection task addressed in this work were that:

- the operating conditions were an outdoor environment with no constraints on factors such as variation in illumination, presence of shadows and the effects of adverse weather.
- the background scene was complex and variable and so contained image features that tended to distract a vision system from the successful identification of pedestrians.
- the camera's position was such that assumptions commonly used by vision systems to avoid the effects of occlusion, perspective and viewpoint variation were not valid.
- pedestrians could be moving or static and were not limited to certain orientations or to movement in a single direction.

- pedestrian pose and attire were unconstrained and detection was required even when details of pedestrian shape were hidden by items such as coats and hats.
- the number of pedestrians to be monitored was not restricted such that the vision system needed to cope with the monitoring of multiple targets concurrently.

This task was made more difficult by the need to attain reliable operation at realistic costs. The cost constraint meant that operation had to be achieved within the restricted computational and sensory resources of a low-cost embedded system. The reliability requirements were based on the performance specifications of the Department of Transport. Failure to meet these specifications could have serious implications for pedestrian safety. Accordingly a significant part of the work undertaken was directed at developing a means of evaluating vision algorithm performance on an extensive test set of image sequences – so as to be as representative as possible of the range of conditions found in the real world.

The investigations into detection at crossings described here are likely to be of wider interest as low-cost pedestrian detection systems would be of value in a variety of related situations in transportation as well as other domains. Potential uses in transportation include the estimation of demand at public transport waiting areas (e.g. bus stops, metro platforms and lifts), safety monitoring (to ensure no pedestrians are present in dangerous areas) and the automated gathering of usage statistics to assist in the assessment of the layout of public areas. Other application areas in which pedestrian monitoring is of interest include security, robotics, architectural planning, advertising and multimedia.

1.2 Summary and breakdown of Chapters

Chapter 2 lays out the transportation background to pedestrian detection and identifies the particular task of the detection of waiting pedestrians at road crossings as the main focus of this work. After discussing the underlying requirements for such detectors, various sensing technologies that have been previously applied to this detection task are reviewed and their relative merits established. Results are also given of a study conducted within the author's research group, that indicated performance problems with the currently favoured Active Infra Red detection technology. Thereafter the expected benefits of a computer vision based system in terms of improved performance and functionality are identified. Finally the task for a vision-based

detector is defined in more detail in terms of the operating environment, performance specifications and error criteria.

In Chapter 3 the selection and development of experimental hardware and software necessary to develop the vision system is described. In particular the constraints on the choice of image sensor and available computing power are clarified so that their impact on later work in algorithm development can be understood. Following this the development of an object-oriented software environment to allow fast, flexible prototyping of algorithms and the design of associated framegrabbing hardware are described.

Chapter 4 is an analysis of the detection task from the point of view of a vision system. On the basis of sample video captured from on-street test sites the main difficulties are identified in terms of constraints on camera position, weather and lighting effects and variation in object shape and behaviour. This provides the basis in Chapter 5 for a review of relevant work on the observation of the human body by computer vision systems with respect to the task specific difficulties of this work. The prior art is characterised both in terms of its application area and the assumptions upon which algorithms were developed. The validity of these assumptions, in the light of the analysis in Chapter 4, is then examined to determine the relevance of the methods developed to the current task

Chapter 6 is concerned with the development of algorithms to perform the detection task. This is guided by the difficulties identified in Chapter 4 and uses, where possible, techniques identified in the review of Chapter 5. A description of initial experimentation carried out to examine the efficacy of several vision techniques leads onto a description of the development of the final detection algorithm.

The final evaluation phase is described in Chapter 7, which addresses the problems of performing accurate evaluations, on a repeatable basis, using large data sets (sequences of video footage). Particular topics discussed are those concerned with generating manual analysis data as a reference performance, the automation of the process of comparing algorithm results against this reference and the associated problems of capture and storage of extended real-time image sequences. Results are given showing the level of performance achieved by the final detector system.

Chapter 8 concludes as to the success of the work completed and indicates the direction of current and future work.

2 The Pedestrian Detection Task

2.1 Introduction

This chapter starts by surveying the relevant background in transportation management necessary to establish the value of pedestrian detection and the reasons for the high level of interest in this area.

The incorporation of pedestrian detection at road crossings is then identified as a task of current importance and details of government initiatives in this area, based around a new type of pedestrian road crossing, are given. The operation of this crossing, known as the Puffin, and its use of pedestrian detection is then described and contrasted to that of the traditional Pelican crossing (section 2.3).

The availability of reliable pedestrian detectors is identified as being critical to the operation of the Puffin crossing. An examination of available information on detection technologies that have been applied to the Puffin's detection needs, and with what level of success, is then presented. Detection based on Active Infrared is found to be the currently favoured detection technology

The results of a comparative study, partially undertaken by the author, of an installation changing over from Pelican to Puffin operation using Active Infrared sensors are then presented (section 2.5). This study compared various measures related to the levels of safety and efficiency achieved for the two crossing types. Although the conclusions of the work were generally positive as regards the Puffin crossing there were clear indications that there was a need for better detection performance for pedestrians waiting at the kerbside. A more detailed analysis of the study results (section 2.6) directed at detector performance at the kerbside showed that, for the crossing studied, detection failure rates were unacceptable.

The limitations of the presently available detection technologies, along with the anticipated need for ever more sophisticated sensing modes in the future (section 2.7), provides grounds for investigating the potential of computer vision as a means to perform the kerbside detection role. Computer vision is then identified as having the potential to fulfil the range of requirements of such a detector if it can be made to work under the demanding operating conditions (section 2.8).

Finally an examination of the conclusions of the comparative study along with the specifications of the current Puffin, and proposed extensions thereto, is used as a basis for defining a specification of the performance requirements and operating constraints for a computer vision based detector of pedestrians waiting at the kerbside (section 2.9). This specification then defines the particular objectives of the research into the development and evaluation of a vision detector described in the remainder of this work.

2.2 Background

In transportation management and control the two principal, and often conflicting, goals are the achievement of high levels of both safety and efficiency. Safety, which can be quantified in terms of frequency of accidents, takes the highest priority whilst efficiency, in terms of traffic flow, is optimised so far as the constraint of maintaining safe operation permits. These objectives are particularly difficult to achieve in urban environments where it is common to find dense vehicle traffic flows coinciding with dense flows of pedestrians. At junctions and pedestrian crossings the intersection of these flows disrupts both flow systems, leading to an expectation of accidents and delays.

The expectation of safety problems is borne out by studies of the safety of vulnerable road users (i.e. pedestrians and cyclists) which have shown that the highest accident rates occur in urban areas and are concentrated at junctions (van Schagen 1991). In addition many pedestrian accidents were found to occur as a result of pedestrians making mid-block crossings at points both on and away from recognised crossing facilities such as Pelican crossings.

Other research has indicated that the consequences of undue pedestrian delay are more than just inconvenience in that safety and efficiency issues may be linked. Boyd (1995) observed that the tendency for pedestrians to non-comply at a crossing (i.e. to

cross outside of the pedestrian green stage when vehicles have priority) increases in proportion to the delay to which they have been subjected. Further evidence comes from the work of van Schagen (1991) who noted that large numbers of pedestrians (and cyclists) would violate red signals in order to reduce their delay. He also observed that some pedestrians would seek to reduce their delay by making less safe mid-block crossings, away from any recognised crossing point. Several studies have found such non-compliant behaviour at signals to be associated with higher risks indicating that subjecting pedestrians to delay has negative consequences for their safety (Tarko 1995, Garder 1989 and Cameron 1978).

The task therefore for traffic control at signal-controlled junctions and crossings can be summarised as being to interleave vehicle and pedestrian flows such that conflicts between them are minimised. To achieve this a sufficient level of service must be maintained to all traffic, such that delays do not reach the point where safety may be compromised. This is becoming ever more difficult as an increasing number of cars are placing more and more strain on existing road networks and traffic control systems. The scope to expand or re-shape these networks as a means of relieving this pressure is however severely constrained, particularly in urban areas, due to other demands on land use. It has therefore been necessary to find alternative means of overcoming this strain.

An important response to this situation has been to exploit technological advances in electronics and communications which have increased the speed, scope and complexity of available traffic signal controllers and hence allowed the use of more sophisticated, sensor-based, adaptive, traffic control strategies. The design of these control strategies has, for the majority of the past two decades, been principally targeted at optimising the movement of vehicles through road networks with respect to the minimisation of delay. During this period the requirements of pedestrians and other road users were treated as being of secondary importance their needs only being served, on the whole, to the point that the impact on vehicle delay was minimal (Hunt 1996).

This emphasis on optimising vehicle flow is illustrated by looking at the way that pedestrian crossing opportunities are included in the operating cycles of traffic control systems. In city centres signal control is often co-ordinated by a centralised Urban Traffic Control (UTC) system that seeks to co-ordinate the optimisation of flow over

networks of several junctions and pedestrian crossings. Examples include TRANSYT which operates based on fixed time and the increasingly popular SCOOT system which dynamically adjusts signal timings over the network to adapt to varying demands. Within such UTC systems the control strategy used is to only allow the vehicle flow to be interrupted at a fixed 'window of opportunity' within the total control cycle. Once a pedestrian demand has been received (via a push-button press at a crossing point) then the vehicle flow is only interrupted when the next 'window of opportunity' arrives at which point the pedestrian priority period is initiated. The time between a pedestrian pressing the button and receiving a response is therefore variable, being dependent on the time at which it was pressed with respect to this point in the control cycle. At some junctions faster response times are offered by offering two windows in each cycle (double cycling) even so it is clear that the requirements of pedestrians are being fitted into those of vehicles only at points where it is convenient for vehicle flow control.

The importance of walking as a mode of transport has however achieved increased attention in recent years with changes in Government policy responding to the impact of growing car populations (especially in urban areas) and public concerns about the environment and sustainability. The results of these changes are reflected in the increasing pedestrianisation of city centres and the encouragement of the public to make less use of their cars for short journeys and to move to public transport (Hunt 1996).

Since the late 1980's there has been a recognition that similar benefits to those obtained in vehicle control might be achieved in pedestrian crossing control by the use of more extensive sensory information on pedestrian activity (in the traffic control systems described above all information on pedestrian activity is provided by pedestrian actuation of push-buttons, which are not always provided at junctions). This has led to an increase in European and UK government interest in crossings incorporating automatic pedestrian detection and has resulted in several initiatives (e.g. under the EU Framework DRIVE I and DRIVE II programs) funded to explore the development of new detection-based crossing types. It should be noted that although it will be apparent that the extension of the same ideas to pedestrian crossing points attached to junctions would be beneficial, development to date has concentrated on proving the value of the new control methods, and the requisite

technology, at mid-block crossings. This work provided part of the basis for the definition of a potential replacement for the Pelican crossing known as the Puffin crossing which, as well as introducing the use of pedestrian sensors, also incorporates changes in the crossing layout and signal control strategies. The operational characteristics of these two crossing types along with available data on their performance is given in the next section.

2.3 Pedestrian Detection at Road Crossings

2.3.1 Introduction

In the UK the most common pedestrian crossing facility is the Pelican (PEdestrian LIght CONtrolled) crossing. It was introduced in 1969 as a replacement for the Zebra crossing system which had difficulty providing good service to pedestrians when vehicle flows were high and, conversely, to vehicles during high pedestrian flow (Hunt, 1990). A further shortcoming of the Zebra was that it was unable to operate as a co-ordinated part of optimising urban traffic control systems being operated purely according to pedestrian action.

In response to the need to improve the pedestrian facilities offered by the Pelican the Department of Transport has started trials of an alternative road crossing known as the Puffin (Pedestrian User-FFriendly INtelligent) crossing (D.O.T.,1993) which uses automatic detection of pedestrian presence to respond more flexibly to pedestrian demand to the benefit of both pedestrians and vehicles.

According to a regular survey of Local Authorities by the County Surveyors' Society, carried out in 1994, there were, in total, 11,183 pedestrian crossings operating at mid-block sites and also pedestrian crossing facilities were present at 47.2% of the 9,948 signalised junctions installed at that time (Traffic Control Users Group, 1994). Of the mid-block crossings the vast majority were Pelicans and 36 were Puffins. Trials of a variant on the Puffin designed to additionally control bicycle traffic (the Toucan crossing) were included within a category of 'other non-flashing' crossings which had 151 installations. The poor response by local authorities to the following survey two years later (Traffic Control Users Group, 1996) meant that the total numbers of crossings reported was lower. However, projecting from those authorities that did respond the number of mid-block crossings had risen by 5.4% and the percentage of signalised junctions with crossing facilities was up to 58.1%.

The operation of the Pelican and Puffin crossing types will now be described in more detail indicating the areas where the Puffin is expected to provide improved performance. The results of a study, partially conducted by the author, assessing the actual benefits achieved by the use of a Puffin relative to a Pelican will then be presented.

2.3.2 Pelican Crossing Operation

With the Pelican type of crossing the pedestrian presses a button to register a demand to cross, at which point a WAIT sign is illuminated. At an appropriate point following the pedestrian demand a pedestrian stage is initiated. The traffic receives a red signal and the signal opposite the pedestrian changes from a "red man" to a "green man" inviting the pedestrian to cross. After a short interval the "green man" flashes and the amber signal to traffic flashes. This indicates that pedestrians currently on the crossing should continue to cross and that if the crossing is empty, vehicles may start to move. This period is followed by a "red man" and green to traffic thus completing the pedestrian stage. The operating cycle and typical time settings of this crossing type are given in Appendix A.

From the detection point of view the button press can be interpreted as a detection of the arrival of at least one pedestrian. Interestingly, although it would be expected that this would amount to a completely reliable detection of presence there is evidence from the study described in section 2.4 that pedestrian failure to make use of the button means that this is not the case.

The Pelican crossing has disadvantages from the perspectives of both pedestrians and of vehicles. Firstly pedestrians may feel under pressure to cross quickly as the "green man" starts to flash and as vehicles edge forward during the flashing amber phase. In addition drivers may be stopped during a pedestrian stage, even though no pedestrians are present, since pedestrians have already crossed during a natural gap in the traffic before the pedestrian stage has commenced - this situation is commonly referred to as a "*false call*".

2.3.3 Puffin Crossing

In addressing the Pelican's problems the Puffin differs in three main areas. These are the use of automatic pedestrian detection, changes in the position of the pedestrian signals and changes in overall signal control to ensure clear priority.

The Puffin makes use of automatic pedestrian presence detection in two ways. Firstly, detection is carried out on the crossing zone to determine if moving pedestrians are still present. The controller can then, when necessary, extend the allowed crossing time (period of pedestrian priority) to assist slow moving pedestrians (up to a pre-set maximum). This detector should only be sensitive to movement and in trials to-date has been of a passive infrared type similar to those used in building intruder alarms. At the kerbside area the Puffin uses a second detector to determine if any waiting pedestrians (who may or may not be moving) are present. This detector has two roles. In **demand confirmation** pedestrian actuation of a push-button is, as with the Pelican, used to register a request to cross but a pedestrian must additionally be automatically detected at the kerbside to validate the demand. For **demand cancellation**, once a demand has been successfully registered, the waiting area is monitored and if no pedestrian is detected (for more than a pre-set minimum period) then the demand will be automatically cancelled thus avoiding a false call. This kerbside detector may be either a surface-mounted or an above-ground type corresponding, currently, to pressure mat and active infra-red systems respectively.

Figure 1 and Figure 2 below give an overview of a Puffin crossing. They serve to illustrate the different features of the Puffin crossing's layout with respect to the more familiar Pelican (the figures were taken from the trial site studied for the work described in Reading et al, 1995).

Figure 1 shows the crossing zone from the viewpoint of a pedestrian about to set off to cross the road. Note the absence of the Pelican's green/red man indication at the far side of the road. For the Puffin the green/red man indication is instead given on the near side (as shown in Figure 2) with the intention that it will better serve as an invitation to **start** crossing only when it is illuminated. This change however has caused some unease in pedestrians familiar with the Pelican as, not being able to see the green man in front of them, they feel unsure of the state of crossing priority whilst they are crossing.

Figure 1: *View over a Puffin crossing. It can be seen that the Pelican's green man signal at the far side of the road is no longer present. This indication is instead given on the nearside (the uppermost of the two boxes mounted on the nearest signal pole).*

Figure 2 gives an overview of the crossing apparatus on each side of the road. The button and green man display units are shown occupying separate boxes with the latter on top. The figure also illustrates a typical position available for mounting an above-ground kerbside pedestrian detector (shaded grey box) shown pointing down at the waiting area.

The signal aspects and sequence differ from those at a Pelican crossing in that the flashing amber and flashing green man sequences are no longer used. This is important as with the Puffin it is, at all times, unambiguously clear whether it is vehicle or pedestrian traffic that has priority. This removes the intimidation that many pedestrians felt during the flashing amber period during which vehicles tend to start revving their engines and beginning to move onto the crossing area. A full description of the operational characteristics of the Puffin crossing can be found in Department of Transport 1993.

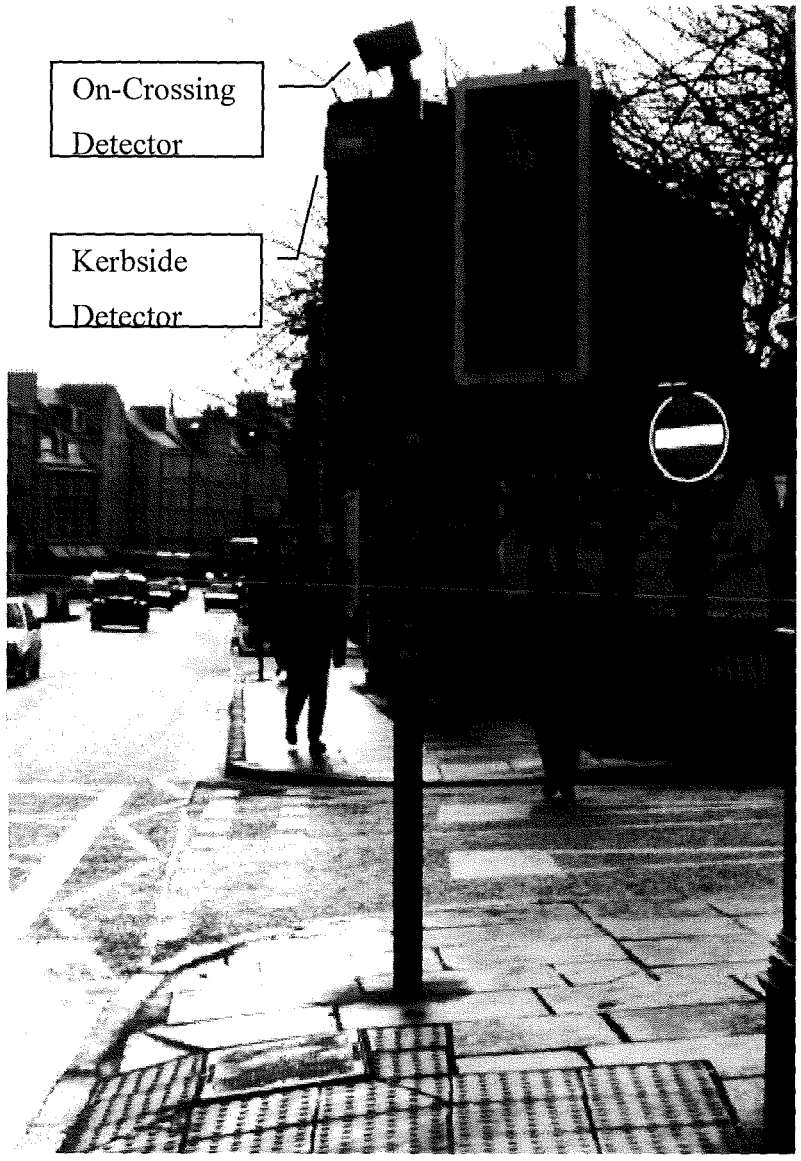


Figure 2: *Signal pole equipment of a Puffin crossing. The centre of the figure shows the invitation to cross signal (red/green man) mounted so as to clarify its role as being an invitation to start to cross. The two detection units can be seen mounted at the top on the signal pole, one pointing down over the kerbside and the other left onto the crossing area.*

With the above factors in mind the Puffin crossing has been designed to provide a safer crossing environment for pedestrians and reduced delay for vehicular traffic. The crossing time is extended to allow the crossing area to clear before vehicles are signalled to move making it less likely that pedestrians will be intimidated by waiting traffic. The automatic cancellation mechanism means that vehicles will not be stopped for a crossing stage where no pedestrians are present.

2.3.4 Conclusion

It is evident that pedestrian detection performance will be critical to the effectiveness of these new Puffin crossings and their acceptance by the public. With this in mind the available data on the operational performance of trial crossings utilising currently available detector types is discussed below concentrating on the role of the kerbside detector.

2.4 Pedestrian Kerbside Detection Technologies

Most current transportation sensing methods were developed for vehicle detection and are as such unsuitable for immediate application to pedestrian sensing. Certainly for many detector types, which are based on properties of vehicles that are not held in common with pedestrians, there is no obvious way of adapting them to pedestrian detection. Inductive loops, for example, rely on a vehicle's metallic construction and sonic detectors on the existence of engine noise. Another problem with the adaptation of vehicle sensing methods is that pedestrians are smaller and are not constrained to follow well-defined spatial paths so detection must take place over a wide field-of-view. This means that methods such as the use of ultrasonic detectors that rely on objects moving under a detection gate are unsuitable.

Some methods such as microwave and active infrared detection systems have however been found to be amenable to adaptation to pedestrian detection and are discussed below. A further approach that has been applied is the use of pressure sensitive mats. This is a migration of technology from safety systems in industrial environments where such mats are used for the automatic cut-off of machinery when human operators come too near to dangerous areas.

It should be noted that the discussion below is based on the information available from published studies which have been completed on the operation of trial installations of pedestrian-responsive road crossings using these various types of pedestrian presence detector. Although quantitative trials must have taken place before these pieces of equipment were licensed for use on the public transportation system this information is not in the public domain. Also the detectors discussed are industrial products, so information on their performance and technical details on their operating principles are also not always available for commercial reasons.

2.4.1 *Microwave Detection of Pedestrians*

Microwave detectors were used of part of the European Commission's DRIVE transport telematics programme under a project entitled "An Intelligent Traffic System for Vulnerable Road Users" and again in a following DRIVE II project "Vulnerable Road-User Traffic Observation and Optimisation (VRU-TOO)". This work, which preceded the specification of the Puffin crossing in the UK, examined the used of the detectors for the pre-arrival detection of approaching pedestrians to trigger automatic demand registration (Rothengatter, 1994). The detectors were commercial vehicle detection units operating at around 10.5 GHz which had been adjusted to be sensitive to a minimum approach velocity typical of pedestrian rather than vehicle movement at 1.5 km/hour (Ekman, 1992)(Rothengatter, 1994). One of the criteria for choosing microwave detectors for these projects was that being of the above-ground type they required no alteration to the pedestrian footway and could be easily mounted on the existing signal pole of the crossing.

Although these detectors were found to be satisfactory for the purposes of these studies they do however suffer several drawbacks when considered for use in kerbside detection at Puffin crossings. Detection performance was reported at 80% of approaching pedestrians (van Schagen, 1991) which although sufficient as an additional feature to push button triggering is too low for the Puffin's requirements where there are safety consequences attached to detection failures, see section 2.9.2.1. There was also some evidence that in certain circumstances that vehicle movements, or reflections of movement, could cause false triggering of the detectors (Rothengatter, 1994). Furthermore a fundamental constraint of this Doppler-based technology is that objects under observation must be moving whereas the Puffin specification insists that all waiting pedestrians, moving or not, should continue to be detected.

A further limitation of the microwave devices is that they were designed to detect objects within a relatively narrow field-of-view over an operating distance ranging from a minimum of 5 metres to a maximum of 30 metres (Rothengatter, 1994). The limitation in field-of-view requires a number of detectors to cover all possible pedestrian approach routes. The minimum operating distance means they are not suitable for mounting on the crossing's signal pole if the waiting area, which only extends for about 4 metres, is to be covered.

An operational problem was that the detectors were observed to be prone to false triggering by pedestrians passing-by who did not intend to use the crossing. In practice it was therefore found necessary to collect data beforehand on the existing routes of pedestrians to enable the detectors to be pointed at only those approach directions found to correspond to pedestrians that are likely to use the crossing (Rothengatter, 1994).

In conclusion, although the work described above has shown that microwave detectors can be usefully used for pedestrian detection, the attributes of these detectors mean that they are not suitable for the particular requirements of the Puffin specification in their present form.

2.4.2 Pressure Sensitive Mat Detection

These devices operate by monitoring the transmission of light through a fibre optic strand which has been embedded into a mat. When a pedestrian stands on the mat their weight deforms the mat material causing the embedded fibre to be bent. The bending of the fibre causes an increase in light loss that can be detected. Being surface mounted, installation and maintenance involve considerable disruption to the footway around the crossing site leading to greater expense.

A European Commission DRIVE project on Pedestrian Urban Safety System and Comfort at Traffic Signals (PUSSYCATS) sought to incorporate technical improvements that were better adapted to the behaviour and needs of pedestrians into road crossings. As part of this work the Transport Research Laboratory (TRL) published a report (Davies, 1992) describing preliminary tests at the TRL and subsequent trials at two signal controlled junctions sites in West Sussex and London using pressure mat detectors. Overall, the results showed that Puffin operation could successfully reduce delays to both pedestrians and vehicles as pedestrians benefited from the longer pedestrian phases and vehicles from the elimination of false calls. Nevertheless, so far as the detectors were concerned the author concludes that due to problems with installing and maintaining the pressure mat kerbside sensors an alternative kerbside detection system may be worth considering. The use of active infrared detectors is suggested by the authors as a possible alternative for consideration. A similar conclusion, as regards pressure mat detectors, was reached by

the Dutch participants in the project (Levelt, 1994) who also advised looking at alternative detection techniques.

A UK project (reference N107) run by the Department of Transport's Urban and Local Transport Directorate Research Committee (ULTDRC) was terminated early as the pedestrian detection equipment at most of the trial sites was not working properly resulting in sluggish operation of the crossings after conversion from Pelican to Puffin operation. The project was completed in May 1993 with no final report being published due to its premature termination.

A second variation on the pressure mat sensing method is based on piezoelectric sensing. No published data has been found on its performance, however even if detection performance improves upon fibre-optic based mats, it is likely to suffer from the disadvantages of requiring relatively high production and installation costs (due to invasive installation) and a lack of flexibility in varying the detection area when compared to above-ground sensors.

2.4.3 Active Infrared Detection

These devices operate by actively illuminating the detection area with a near infrared source and monitoring the reflected signal with an array of infrared sensitive photo-diodes. The use of near infrared rather than visible wavelengths avoids any awareness by pedestrians of the illumination. It should be noted that operating in the near infrared, where there is significant power in the solar spectrum, the problematic effects of shadows and reflections could affect the performance of these detectors in a similar way to vision systems operating in the visible portion of the spectrum (these matters are discussed in more detail in Chapter 4).

Although no technical description of the operation of these devices was found in the literature it is reasonable to assume that modulated illumination was used to help exclude the effects of shadows due to relatively slow changes in solar illumination. High-pass filtering could then be used to observe just the reflected signal resulting from the active illumination (shadows due to the infrared source itself will be invisible to the sensing array as the source is mounted next to the sensors such that shadows are hidden behind the objects causing them). These detectors are aligned mechanically by sighting along the top of the casing and that they have no facilities for flexibly defining the active detection area

Active IR devices are the currently favoured detection method for Puffin installations although no previous independent studies on their performance were found to be available (in the public domain). The next two sections describe work performed by the author as part of studies on a Puffin crossing operating with these detectors. They indicate the detection performance achieved in practice.

2.5 A Comparative Study of Pelican and Puffin Performance

2.5.1 Introduction

The planned conversion by Lothian Regional Council of a pedestrian crossing in Edinburgh from Pelican to Puffin operation presented the opportunity for the research group to perform a before-and-after study of the effects of the changeover. The kerbside pedestrian detectors used on the Puffin installation for this work were of the active infrared type and the on-crossing detection was performed using passive infrared devices.

The study performed was broad reaching in the aspects of crossing performance examined. Comparative measurements of various transportation parameters (e.g. delay to vehicles, delay to pedestrians, pedestrian compliance, pedestrian crossing time and crossing occupation time) for the two crossing types were made along with Puffin specific measures of extensions in crossing time and cancellations of crossing requests. Various observations of pedestrian behaviour and of detector performance were also made.

2.5.2 Results and Discussion

The main conclusions of the comparison of Puffin and Pelican operation under similar conditions are summarised below.

- There was an increase in delay to vehicle traffic at the Puffin crossing by between 0.4 and 1.0 vehicle hours per hour, or by approximately 5 seconds for each stopped vehicle. However the stressful conflict between the pedestrians and vehicles during the flashing amber period was eliminated and replaced by stages of clearly defined priority.

- The level of non-compliance at a crossing was influenced by a variety of factors of which frequency of pedestrian stages may be the most significant. There was some evidence that non-compliance decreased on the Puffin crossing although the proportions of pedestrians anticipating the green man (by observing vehicle signals) at the Puffin and Pelican crossings were not significantly different.

- For all classes of pedestrians the mean times to cross the road at the Puffin site were higher than those during Pelican operation. When considering all pedestrians, the difference observed was only 0.3 seconds although it was somewhat greater (0.46 seconds) in the case of the elderly. A possible explanation for the increase is that, firstly the removal of the flashing amber to vehicles reduced the encroachment of vehicles onto the crossing reducing the associated harassment and secondly the removal of the flashing green man has taken away the "hurry" feeling felt by some pedestrians. One of the aims of installing Puffin crossings is to reduce pedestrian stress. Observed changes in the pedestrian crossing time seemed to suggest that this was achieved.

- The mean occupancy time of the Puffin was about 3 seconds greater than that for the Pelican. In addition the Puffin had a much greater variation in occupation time. Analysis of the observations also showed that during 1.5% of Pelican crossing cycles, pedestrians were still crossing when green to traffic had started, although there were no observations of pedestrians on the Puffin crossing when traffic was being shown the green aspect.

- Delays experienced by pedestrians were related in the main to the frequency of pedestrian stages rather than the detailed operation of the crossing in terms of signal aspects.

Of particular relevance to pedestrian detection was the observation that delay at the Puffin crossing was found to be increased by two additional factors: the observed pedestrian behaviour in registering demands to cross via the push button and faulty pedestrian detection. These points are discussed in more detail below.

2.5.2.1 Demand Registration

So far as pedestrian push-button usage was concerned two different issues were examined. Firstly the likelihood that any pedestrian will request a pedestrian stage via the push-button if the "wait" light is not illuminated and secondly the proportion of cycles where the first pedestrian to arrive presses the demand button.

The percentage of people who pressed the push-button on arrival if the WAIT lamp was not illuminated is shown in Table 1 below, for three time periods over four days of observation. The percentage figure represents the number of pedestrians who pressed the button on arrival (if the wait lamp was not illuminated) divided by the total number of pedestrians who arrived while the wait lamp was not illuminated and so had the opportunity to press the button.

DAY (TYPE)	9.00-10.00	12.00-13.00	16.30-17.30
1 (PELICAN)	95.3%	77.3%	80%
2 (PELICAN)	89.9%	78.3%	85.8%
3 (PELICAN)	91.3%	73.1%	92.7%
4 (PUFFIN)	70.4%	72.1%	67.5%

Table 1 *Percentage of pedestrians that use the push-button on arrival*

Aggregating over the different crossing types the percentage of pedestrians that used the push-button on first arrival was 82% for all the Pelican days and 69% for the Puffin day. The reason that fewer pedestrians push the button when the wait lamp was not illuminated for the Puffin crossing compared to the Pelican may have been due to the lack of familiarity with the new crossing compared to the old style Pelican crossing. The effect of this over a given period would be to provide fewer pedestrian stages for the given pedestrian flow and hence increase overall pedestrian delay.

The number of cycles in which the first arriving pedestrian pushes the button is shown as a percentage of the total number of stages in Table 2 below:

DAY (TYPE)	9.00-10.00	12.00-13.00	16.30-17.30
1 (PELICAN)	95.1%	80.4%	84.7%
2 (PELICAN)	90%	84.4%	90.8%
3 (PELICAN)	88.9	82.3%	96.1%
4(PUFFIN)	79.7%	72%	73.4%

Table 2 *Percentage of stages in which first arriving pedestrian presses the button*

These results indicate that, even with the familiar Pelican, in nearly 12 percent of cycles the demand registration is not made by the first arriving pedestrian. The incorporation of an Automatic Demand Registration function (based on pedestrian detectors) into the crossing's specification could therefore provide considerably better service to pedestrians.

2.5.2.2 Pedestrian Detection

Considering the pedestrian detection aspects of the study, a strength of the Puffin crossing, in principle, is the cancellation of pedestrian demand when pedestrians are no longer waiting to cross. During the survey of the Puffin crossing, seven valid cancellations were noted thereby improving the efficiency of the Puffin operation. Unfortunately, over the same period eleven false cancellations were identified which resulted in increased delay to pedestrians, who had to wait for a second demand to be registered at the push button. In addition to the above, detection deficiencies resulted in 7.6% of valid requests being undetected leading to further increases in pedestrian delay at the Puffin. The reasons for these failures were that either pedestrians were standing outside the designated zone, or that the detectors simply failed to register pedestrian presence. In either case, there appeared to be scope for further improvements of the pedestrian detection system.

The overall conclusion of the study was that the Puffin crossing operated in a more flexible manner allowing, where necessary, more time for pedestrians to cross. The pedestrian detection enabled unwanted pedestrian stages to be cancelled to the advantage of both pedestrian and vehicle traffic, although failures in the method of pedestrian detection caused unnecessary delays and highlighted the need for improvements.

2.6 A Study of Pedestrian Kerbside Detection Performance

This study having flagged problems with the detection system, a follow-up study was performed including more detailed analysis on the reasons for detector related problems at the kerbside (Reading et al, October 1995). Detection failures were split into the two categories of detector failures and detection system failures to differentiate between failures of the detector unit to respond to pedestrian presence and errors related to pedestrian waiting position, respectively. The latter category was devised to take account of pedestrian behavioural tendencies to wait outside of the specified zone of coverage of the detector, a common example involved pedestrians pressing the button by reaching around from behind the signal pole. The results, broken down into these two categories, are given in the tables below with respect to the demand confirmation and demand cancellation roles of the crossing.

	Unconfirmed Demands due to Pedestrian Waiting Position	**Total Unconfirmed Demands**	**Total Demands**
Sample Day 1	6	19	230
Sample Day 2	9	18	169
Total	15	37	399

Table 3: *Puffin performance at Automatic Confirmation of Pedestrian Demand*

	Erroneous Cancellations due to Pedestrian Waiting Position	**Total Erroneous Cancellations**	**Total Valid Cancellations**
Sample Day 1	3	11	7
Sample Day 2	3	9	12
Total	6	20	19

Table 4: *Puffin performance at Automatic Cancellation of Pedestrian Demand*

A total of six hours of operation were analysed over two sample days. Over this admittedly short sample period, necessitated by the intensive manual data analysis involved, it can be seen that even when the detection failures due to pedestrian waiting position are excluded the number of detection errors was still very high. The data for automatic cancellations is of particular concern as the number of erroneous decisions by the detector even exceeds the number of valid ones on the first sample day. These failings are very important given there are safety implications in pedestrians being ignored by the crossing, making it more likely that they will become frustrated and cross during a period of traffic priority. The successful cancellations did represent a saving in terms of delays to vehicle traffic but perhaps this was achieved at too high a cost in terms of increased pedestrian risk.

2.7 Extended Detection Requirements

Before going on to look at the use of vision systems it is useful to try and anticipate future requirements of these detectors. Consideration of the historical development of pedestrian crossings indicates a trend towards utilising increasing amounts of information on pedestrian activity to improve levels of service. The Zebra used no (mechanised) detection, the Pelican used pedestrian operation of the push-button as a detector of pedestrian demand and most recently the Puffin has been based around two binary detectors of pedestrian presence.

This trend looks set to continue with new crossing types being introduced which incorporate corresponding increases in the detection requirements. An example is the Toucan crossing, which is already in operation, and extends the Puffin to accommodate cycle lanes and so requires the additional detection of bicycle presence.

A further extension of the Puffin using a volumetric detector at the kerbside is also under consideration by the Department of Transport. These crossings (referred to as Volumetric Puffins) extend upon the basic Puffin's use of presence detection by using detection of the number of waiting pedestrians. This volumetric measure of demand is then used to dynamically increase the frequency of pedestrian stages in proportion to the number of pedestrians that are waiting. Aside from decreasing delay to pedestrians during periods of heavy flow, this feature is expected to reduce the frequency of non-compliant crossings (i.e. occasions when pedestrians starting to cross the road whilst the green man is not showing) and hence improve safety. The effects of volumetric

response and the performance of sensors capable of measuring the volume of pedestrian demand, have been examined in on-street trials by the Department of Transport. Vision-based detection equipment based on the work described in this thesis was one of the systems monitored in these trials (see Chapter 7 on evaluation).

Projecting forwards, the trend towards increasingly complex sensing makes it likely that in the future detectors that can make behavioural distinctions will be in demand. The Puffin crossing current requires detection of all pedestrians in the waiting zone regardless of their reason for being there. In practice it is however rarely valid to assume that all pedestrians present are waiting to cross as many will merely be passing-by or arriving from other side of crossing. There would therefore clearly be an advantage in discriminating between these categories. Given that 'passing-by' and 'waiting' pedestrians overlap spatially, this discrimination is likely to be based on the extraction of temporal behavioural attributes, in addition to position information, by the detector. The value of this classification in avoiding erroneous detections was identified by Ekman (1992) during work on microwave detection methods and by Hunt (1992).

A further improvement in the level of performance would be to track the trajectories of each individual pedestrian as they move through the scene - a difficult task considering the large numbers of pedestrians that may be involved and the high degree of occlusion between them. If achieved this could provide important feedback to transport system designers. For example, it would enable examination of the effects of crossing layout and timings on the actual waiting times of individuals and their movement patterns during waiting as well as identifying unsafe behaviours such as crossing violations.

The results of the study described in section 2.5 on pedestrian use of the push button indicated that it may also be advantageous if the registration of a demand to cross could be made automatically by a pedestrian sensor. Such an automatic demand registration (ADR) system would be more convenient for pedestrians (particularly the mobility impaired) but would demand very high confidence in the reliability of the presence detectors. The indications from the studies on detector performance reported above are that this level of confidence has yet to be achieved. Verification of the likely benefits of ADR at road crossings has come from simulation work. Having identified the importance of improving pedestrian facilities Hunt (1996) compared a

range of alternative operating strategies for pedestrian crossings. An important feature of all these strategies is that they were based on an assumption that reliable detection of pedestrian presence was available. Modelling the use of pedestrian detectors for ADR on pedestrian arrival at mid-block crossings their simulation studies show improvements could be expected in terms of:

- a reduction of the percentage of pedestrians that cross during the *red man* period with the accompanying expectation of safer operation.
- a better balance in the mean delay imposed by the crossing on vehicles and pedestrians.

Having examined the problems of current detectors and the anticipated future detection demands the next section can look at the role that a computer vision system might play in meeting the needs of pedestrian detection at Puffin crossings.

2.8 Pedestrian Kerbside Detection by Computer Vision

2.8.1 The Potential of Vision-Based Detection

The study described above (section 2.5) supports the case for the use of Puffin crossings in place of Pelicans. However it also indicates that reliable detection performance will be important in obtaining the advantages of the Puffin as well as being critical for ensuring pedestrian safety. The discussion in section 2.4 presented data from studies using microwave, pressure sensitive mat and active infrared devices for the detection of pedestrian presence. The results suggest that it is questionable whether any of these methods are yet able to provide this function to a level sufficient to meet the Puffin's presence detection needs at the kerbside. Furthermore, given the need for extended detection modes, which are even more demanding in terms of reliability and the extraction of information on pedestrians, it seems reasonable to conclude that significant technical challenges remain in the design of pedestrian detectors. This section will examine the potential of a computer vision based system to meet these challenges.

The most important aspect in which vision-based detectors differ fundamentally from the methods described above is that the underlying sensing is performed at a much higher spatial resolution distributed over the field-of-view (e.g. typically > 100,000 sample points from an image sensor). So far as presence detection is concerned this

extra information is available to aid discrimination between genuine pedestrians and other objects that might cause false triggering. The high spatial resolution also makes extended functionality, such as volumetric detection, bicycle detection and the tracking of individual pedestrian movement for the extraction of behavioural information feasible.

A further important advantage of the spatial resolution of a vision detector comes from the sensory input being in a form that is meaningful to human observers. For the user one of the most important aspects of a detector is the means of specifying which areas of the footway are to be part of the active detection zone. Active Infrared devices rely on mechanical sighting methods for alignment. This not only restricts the accuracy of alignment and but also limits the shape of the detection zone to being a simple rectangle. In contrast, vision systems offer a great deal of flexibility. During installation they can be aligned electronically by supplying an image to the installer on which they can simply draw, with a paint program, to specify detection zones (see Appendix B). This means that zones of complex geometry can be catered for which is an important consideration in practice. Although surface mounted mats allow some flexibility in specifying zone shape, they are much more difficult and expensive to realign if changes in the detection area are to be made. A related potential advantage of vision is that a detector could automatically monitor its alignment and signal if it has been accidentally knocked out of position. This feature would be of significant value to operators as they report this is a frequent problem in practice for transportation sensors.

Being an above-ground (non-contact) detection method vision has, like the infrared devices, all the practical advantages of ease of access for installation and maintenance over surface mounted mats.

A further practical concern expressed in various studies of the use of pedestrian detection is that of vandalism of the sensor (Ekman, 1992). This was anticipated to be particularly problematic if the sensor was perceived to be a video camera - even when it was in fact a microwave detector. Although none of the above referenced studies had any problems with vandalism, one trial reported evidence of tampering with the sensor - and these studies were of a small scale. Vision based systems have a clear advantage in this respect compared to alternative technologies as only the camera need be mounted on the signal head with the (relatively expensive and vulnerable)

processing units positioned remotely in a control box at the other end of a coaxial cable. Solid-state video cameras of very small physical size are now commonplace with lens apertures, which are all that need be exposed to the outside world, which can be as small as a pinhole. For example the camera used in this work had an external lens aperture of only 8mm.

From an economic point of view it is also timely to look at the use of vision systems as the cost of both sensing and processing hardware continues to fall and the practical use of vision systems for real applications continues to expand.

On the negative side, the advantages gained from greater spatial resolution mean that a relatively large quantity of data needs to be processed for each assessment of the scene. This incurs the well-known heavy processing and storage demands associated with image processing systems. Nowadays however these systems are rapidly becoming more economic due to hardware advances in both sensor and processor systems. This has meant that high unit cost vision systems for monitoring road traffic are now becoming more commonplace although whether practical systems can yet be created at low-cost for widespread use, is one of the questions the work will seek to answer. The other main drawback of the higher spatial resolution is the difficulty in deriving the algorithms to reliably process the image data to extract the correct measurements of pedestrian activity. This task will be the main objective of this work.

The case for the suitability of vision for extended detection modes is supported by events in a trial of the volumetric Puffin funded by the Department of Transport. Prototype volumetric detectors (based on any technology type) were invited to take part. At the start of the project around 14 alternative detection systems were offered for evaluation. After two preliminary evaluation stages the two detectors selected for long-term on-street trials (one of which was the detector designed by the author and described in this thesis) were both based on computer vision technology. Chapter 7 includes more detail on the trial process. This provides some indirect evidence that the manufacturers of the non-vision detectors described above were unable to successfully convert their devices to volume sensitive operation.

2.8.2 Conclusion

When compared to alternative sensor technologies, computer vision sensors offer the potential for better performance, an enhanced specification and greater usability in

terms of easier visual alignment, installation and maintenance. The next section goes on to layout the vision detection requirements addressed by this research in more detail and to define appropriate terminology.

2.9 Requirements of a Vision Detector of Pedestrians at the Kerbside

2.9.1 Introduction

It is clear from the above that there is a need for reliable pedestrian detection that is not being met by currently available devices. The following section seeks to define a set of detection objectives for this research into computer vision based detectors. These objectives incorporate the current detection requirements of the basic and volumetric Puffin crossing types introduced above as well as extended functionality that might be provided by a computer vision system. Much of what follows is therefore based on the specification (Department of Transport 1996 and 1997) for the Puffin crossing and extensions thereof appropriate for the volumetric Puffin. Other aspects of this specification are based on experience of the practical problems associated with using vision systems in the field particularly when they are to be installed and maintained by personnel with no knowledge of vision. Finally, key constraints on the detector design in terms of impositions of the transportation regulations and the range of operating conditions are incorporated.

2.9.2 Detection Modes

Through examination the categories of pedestrian responsive crossing described above, the author decided to address the demand for the following categories of pedestrian detection. The nature of errors related to these detection categories are defined as a basis for reference during the evaluation work described in Chapter 7.

2.9.2.1 Binary Detection

An output signal is required for the basic Puffin crossing specification indicating either the presence or absence of pedestrians in the waiting area. In evaluating binary detection performance, the safety of users of the crossing must be of primary importance. A pedestrian ignored by the crossing detector is liable to become frustrated with increasing delay and is more likely to cross when it is unsafe. Pedestrian safety is therefore most likely to relate to the number of **false negative binary detections**, defined as those occasions where a waiting pedestrian is missed by

the binary detector. For efficiency of vehicle flow, it is also important to minimise the number of **false positive binary detections**. These are defined as occasions where a presence signal is given in the absence of waiting pedestrians thus leading to unnecessary stoppage of traffic.

For the purposes of this work evaluation will be performed with respect to the above categories. It should be noted however that the Department of Transport specifications incorporate certain precautionary measures in the controller specification of the Puffin crossing as a guard against binary detection errors. A timing device stretches the signal from the kerbside detector for a **hold period** as a cover against unreliable detector operation if, for example, the subject temporarily moves out of the detection zone in the act of pressing the push button. A similar timing device is used to hold the output of the push button. The hold period is presetable in the range of 1 to 2 seconds in steps of 0.2 seconds. In addition, a **time-out period** occurs after a pedestrian demand has been accepted and the detection of pedestrian presence has ceased. At the end of this time-out period, the demand is cancelled. This feature effectively allows a measure of security in case a false negative detection should occur, of up to the time-out period. This period is presetable in the range of 2 to 5 seconds in steps of 0.2 seconds.

2.9.2.2 Volumetric Detection

An output signal is required giving a measure of the number of waiting pedestrians at the kerbside as required by the Volumetric Puffin crossing type. For evaluation, a measure is required of the accuracy of the counts. In practice the required accuracy becomes less stringent for higher volumes as shown, for example, in the banding shown in Table 5 below which has been taken from the specification of a Department of Transport trial of volumetric pedestrian detectors.

In the author's opinion an improvement would be to allow for some overlap between these categories in the assessment process otherwise particularly high accuracy is required at the category boundaries. For the purposes of evaluation, see Chapter 7, the vision system will be assessed with respect to the actual number of pedestrians present as defined by a manual reference count.

Volumetric Category	Pedestrians Actually Present
0	0
1	1
2	2-4
3	5-7
4	8-12
5	13-17
6	>18

Table 5: *Categorisation of pedestrian group sizes in to seven levels reflecting the decreasing need for accuracy with increasing volume*

2.9.3 Operating and Design Constraints

2.9.3.1 Pedestrian Size

The Puffin specifications (Department of Transport 1997) include definitions of both maximum and minimum sized pedestrians which are used to define the worst case conditions for avoiding false positive and false negative detections respectively as is described in section 2.9.3.3 below on the detection zone.

A minimum sized pedestrian is defined as having a height of 1.0m, a width of 0.5m, a depth of 0.2m and a mass of 20Kg with the form and dynamic properties of an average child aged five years old.

A maximum sized pedestrian is defined as having a height of 2.0m, a width of 0.75m, a depth of 0.35m and a mass of 80Kg with the form and dynamic properties of an adult.

Pedestrians should also be detected if they are seated in a wheelchair or pushchair.

2.9.3.2 Pedestrian Attire

Detection performance is to be maintained even when pedestrians are dressed in worst case conditions. This includes the wearing of waterproof or cold weather clothing.

2.9.3.3 Detection Zone Dimensions

The detection zone is the area of the pedestrian footway over which the detector is required to produce a response. It is defined by a two dimensional region on the

surface of the footway within which a pedestrian should be standing to be detected.

Figure 3 shows a plan view of a typical detection area as specified for the Puffin crossing. Note that the distance between the specified detection zones and the road edge is not part of the specification. A representation of the road area has been shown here merely to make the orientation of the detector environment clearer.

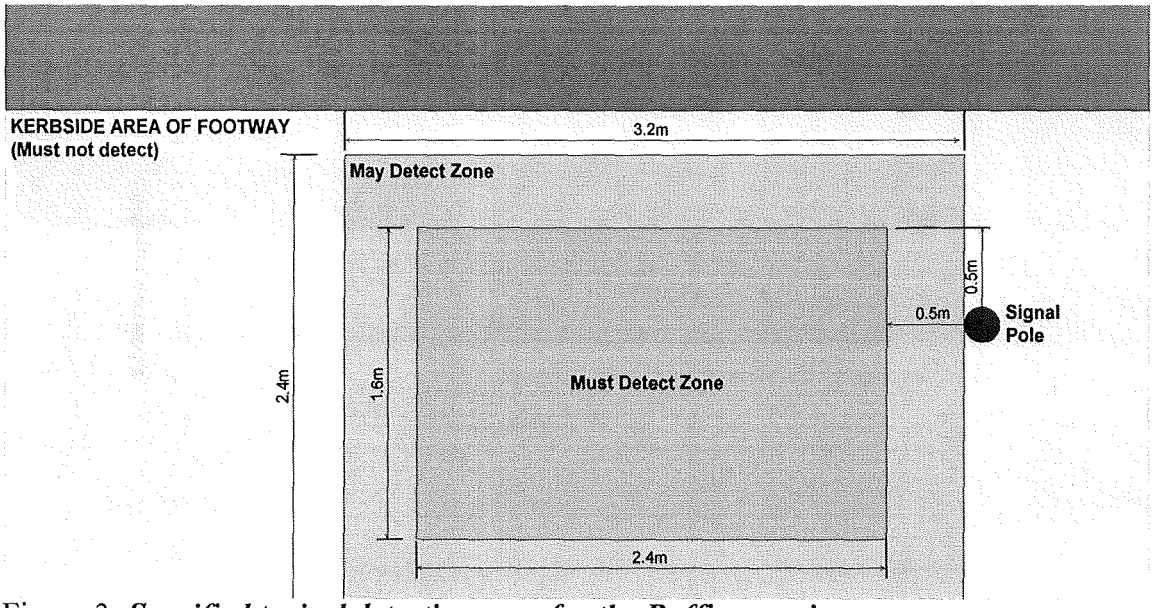


Figure 3: ***Specified typical detection zone for the Puffin crossing***

In the above figure, three categories of detection zone have been specified the central 'Must Detect' zone being surrounded by a 'May Detect' zone which in turn is surrounded by a 'Must Not Detect' zone. The performance requirement for the 'Must Detect' zone is that a minimum sized pedestrian must always be detected in this area. In the 'Must Not Detect' zone a maximum sized pedestrian must not be detected. Between these zones lies a 'May Detect' zone reflecting the fact that detection errors with respect to both the above criteria are tolerable. Obviously it is desirable not to have to make this distinction and preferable to have a clearly defined sharp boundary between the 'Must Detect' and 'Must Not Detect' regions. However the existence of the 'May Detect' zone is a recognition of the fact that, being of finite extent, the precise position of a pedestrian is difficult to define due to the lack of a clear ground-level reference point on the human body.

In this work the objective will simply be to detect all pedestrians within a single user specified zone. Tolerance will be incorporated by allowing any pedestrian who is only partially in the detection zone to be optionally detected but all pedestrians fully in the zone must be detected.

One important advantage of vision systems over other detector technologies which lack any spatial resolution in the sensor is that detection zones of arbitrarily complex shape should be able to be specified. In the figure, this area is shown as a simple rectangle whereas in practice it is advantageous to allow a more flexible specification of the area for the following reasons:

- Many kerbside waiting areas are of more complex geometry such that it is necessary to be able to specify a non-rectangular area over which detection should occur.
- A transportation engineer may wish to bias detection towards certain regions of the waiting area. Typically this would involve broadening the detection zone near the signal pole if most pedestrians tend to wait in this area
- A transportation engineer may wish to disable detection in certain regions of the waiting area. This may, for example, be required to avoid false positive detections due to overlap of the detection zone with regions where pedestrian presence is not associated with the crossing e.g. where pedestrians are just passing-by.

As a confirmation of this value of flexible zone specification it should be noted that the specifications for the Volumetric Puffin's detector (Department of Transport, 1996) specify the detection zone in terms of a grid of non-overlapping blocks of 0.25m side and further state that diagonal zone edges (presumably of higher resolution) would be desirable.

The use of a detection area of any shape within the sensor's field of view will therefore be an objective in this work. The 1.6m by 2.4m dimensions in Figure 3 above are only specified as typical for the purposes of the specification and in practice the must detect area will often need to be much larger. A practical implication of the use of flexible detection zone specification is that an installation engineer must be able to adjust detection zones whilst on-site, thus requiring access to video images.

A somewhat surprising feature of the specified 'Must Detect' zone shown in the figure is that it only extends to within 0.5m of the base of the signal pole (and hence the push button) position. As pedestrians are very likely to wait in this area after a button press it would be better to extend the zone right up to the base of the pole. One reason for its inclusion may be a recognition of the practical problem that for an above-ground detector mounted on the signal pole the view of the area near the pole

base is commonly at least partially obstructed by the presence of other traffic equipment also mounted on the pole. In the trial specifications for the volumetric Puffin this distance is lower at 0.3m. In this work detection of pedestrians standing up to as near as possible to the signal pole will be sought.

2.9.3.4 Crossing infrastructure

The vision system should require no alteration to the crossing area's structure or appearance. The reasons for this, aside from the costs implicated in such changes, are that many aspects of the crossing environment such as pavement shape, colouration and texturing that have been designed with other purposes in mind. Texturing, for example, is used to help the partially sighted to know when they are standing on a crossing's waiting area.

Above-ground detector units should be mounted at or near the top of the existing signal pole that carries the pedestrian request button. This limits effective camera mounting height to under 3.5 metres (less in many practical circumstances) and has a significant effect on the vision task as it removes the possibility of simplifying algorithm design by mounting the camera such that it has a plan view of the scene.

Another related constraint is that the mounting position of the detector should be inconspicuous to avoid attracting the attention of vandals. Not only does this mean keeping the size of the box small and grey (or black) coloured it also means the protective window behind which the camera is mounted should be opaque from the pedestrian viewpoint.

2.9.3.5 Response Time

Both the basic and the volumetric Puffin specifications require that the output of their detectors should respond to the arrival or departure of pedestrians within 500ms.

2.9.3.6 Environmental Conditions

Detector performance must be independent of all weather and environmental conditions appropriate to the normal range of UK weather including fog, heavy rain and snow at all times of the day or night. Significant parameters affected by these conditions include temperature, humidity, vibration and illumination. It should also ignore the effects of movement in the specified 'Must Not Detect' detection zone such as the shadows of passing pedestrians and vehicles. More detail on the consequences of environmental variation for vision system design are given in Chapter 4.

2.9.3.7 *Economic Constraints*

To be considered viable for widespread use, a detector should be suitable for eventual commercial implementation at a production cost of around £300. This figure is based on the total cost of instrumentation of a crossing using currently available alternative detectors. It may be that a somewhat higher figure would be acceptable given the extended functionality of a vision-based detector. Even so, the budget is very restrictive for a computer vision system.

Although this work was not tied to a particular budget, emphasis was placed on achieving the required functionality at minimal cost. The effect was to impose some restrictions on the types of equipment that could be used as is discussed further in Chapter 3.

2.9.3.8 *Failure Conditions*

For this application system reliability is crucial. A detector should therefore incorporate self-monitoring, reporting and recovery from failure conditions. Attention must therefore be paid to ensuring fail-safe behaviour. This amounts firstly to ensuring that any failure is identified by the system and secondly that a safe response is made including recovery where possible.

The important failure modes identified were:

- Camera lens obscuration (due to accumulated dirt)
- Loss of alignment (due to movement of the camera mount)
- Loss of video signal (due to cable break or power loss to camera)
- Loss of power (to detector processing system)

In all these cases, upon identification of a failure the detector should respond by signalling a continuous pedestrian presence (otherwise a permanent false negative detection state might occur with the accompanying safety implications of increasing pedestrian delay).

Failures must also be reported to the controller so that appropriate maintenance operations can be initiated. So far as loss of power is concerned both the basic and the volumetric Puffin specifications require that from a power-up or reset of the detector a maximum period of 10 seconds may pass before the output of the detector is valid.

As important as the mechanical and electrical reliability requirements is the need for logical stability within the software. It must be able to operate correctly in all environmental conditions and continuously for months on end.

Due to the need to concentrate on the fundamental detection tasks it was considered not to be possible to address these important aspects of a detector in this work.

2.9.3.9 Installation and maintenance

The relatively wide range of functionality offered by a vision based detector, necessarily requires a method of allowing the user to interact in setting parameters, performing alignment and specifying detection zones. This meant that the specification of measurements of the installation to assist in calibration had to be limited to simple linear measurements with a high degree of error tolerance. The installation method developed is described in Appendix B and referred to with respect to calibration in Chapter 6.

2.9.3.10 Mechanical and Electrical Requirements

The mechanical requirements relate chiefly to the casing of the camera head and were dealt with in collaboration with industrial partners who were familiar with the design of transportation equipment. The mechanical mountings used were based on simple bracket arrangements as used by current transportation detectors. The vision detector had therefore to tolerate the usual amounts of movement and vibration inherent in such mountings. This matter is discussed further in Chapter 4.

2.10 Conclusions

The demand for pedestrian detection from the Department of Transport and hence the transportation industry for the control of pedestrian facilities is apparent. However the introduction of pedestrian sensing has been hampered by technological difficulties as indicated by the results of studies of current detector performance cited above.

It is the author's belief that the use of vision has significant advantages to offer over other sensing technologies in terms of providing a wider variety of more detailed measurements of pedestrian activity. These benefits can however, only be appreciated if operation within the demanding range of constraints given above can be achieved.

This work therefore studied the use of computer vision for pedestrian detection concentrating on the particular case of the detection of pedestrians waiting at road

crossings. The central objective was the achievement of reliable performance with respect to binary and volumetric detection in an on-site situation.

Now that the task of detecting pedestrians waiting at the kerbside of a pedestrian crossing has been defined in some detail the next chapter describes the selection and where necessary development of equipment to support the capture and analysis of video of pedestrian activity.

3 Equipment Development and Selection

3.1 Introduction

This chapter describes the selection and development of tools required to support the computer vision work described in the remainder of this thesis. As was indicated in the last chapter, this work was primarily concerned with examining the feasibility of using computer vision to detect pedestrians. A major part of this was the development and evaluation of algorithms to perform the computer vision functions of the task. However, a further aspect was to show that the detection could be achieved within equipment constraints that would realistically reflect what would be economic for the system to be suitable for eventual, widespread commercial use.

Clearly, it would only be appropriate to consider the details of exploitation once a set of proven algorithms had been established using a research system. However, knowledge of the eventual constraints necessary to ensure commercial exploitation would be viable did have an important influence on the equipment used and thus indirectly on the algorithm development process.

For this work, the main practical consequences of the economic constraints on equipment were on available processing power and on the choice of image sensor. The latter had important consequences for the quality of image data passed on to the rest of the detector system for analysis. The former put the onus on the development of algorithms that were computationally efficient. These points are discussed in greater depth within sections 3.3 and 3.4 after the framework of the hardware components of a vision system has been explained in section 3.2.

During this research work, the main requirement in terms of development equipment was for a laboratory system on which the vast majority of the algorithm development and evaluation work could be carried out. The prime considerations here were to have

equipment of sufficient performance (speed of operation) to allow evaluation to proceed whilst at the same time maintaining maximum flexibility and observability. In addition, there was also a need to allow the laboratory-developed algorithms to be able to be evaluated in on-street trials requiring a real-time response. This requirement was an influence on the choice of processing environment used for development described in section 3.4. The migration of developed algorithms to an on-street trial system is then described in section 3.6 followed in section 3.7 by an examination of some issues concerning the viability of eventual commercialisation. The mechanical aspects of mounting test camera's for data gathering and trials are covered by section 3.8.

Finally, section 3.9 gives an overview of the MISA (Multi-scale Image Sequence Analysis) object-oriented software environment written by the author to support both the laboratory and on-street trial versions of the hardware. This software proved to be a reliable and flexible basis for algorithm development and very much formed the backbone of this work. Its development and maintenance was probably the single most time consuming aspect of this project.

3.2 Vision System Hardware

Figure 4, shows the major components of a computer vision system and the form of information flow between them. It operates as follows: Light reflected or emitted by world objects is focussed onto an image sensor consisting of a two-dimensional array of light-sensitive pixels. The pixel values are scanned row-by-row and the measured intensities are output as a sequential stream of video data mixed with synchronisation signals. An image capture device (or frame-grabber) is then responsible for converting the analogue stream into digital form and storing each frame such that it is accessible to the processing system. The processing system then executes algorithms to reach decisions as to what is happening in the world system. It then controls actuators that make a world response.

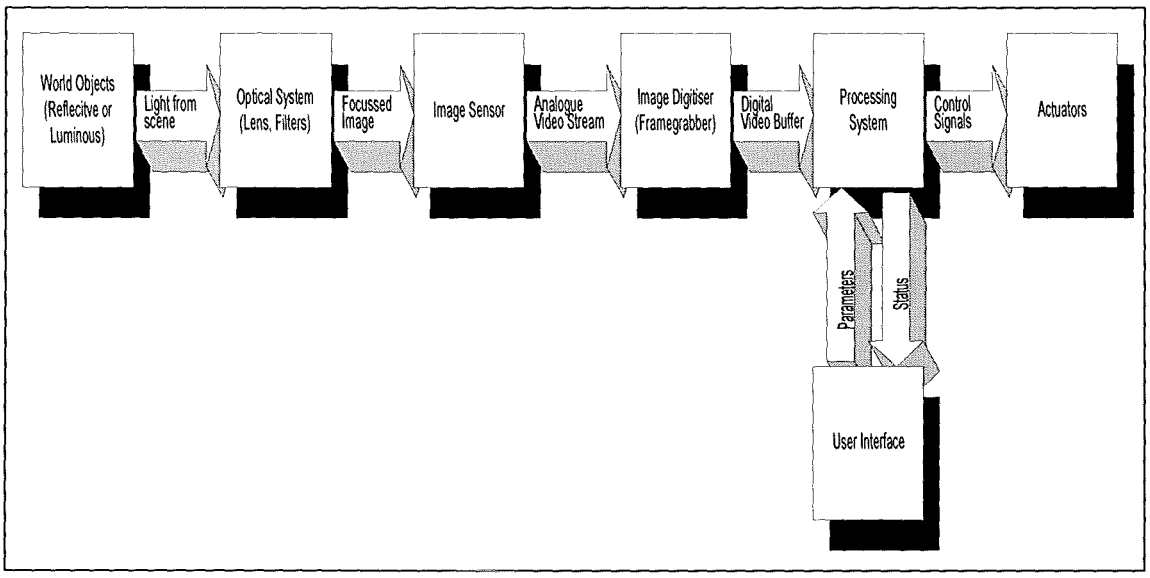


Figure 4: *Breakdown of the hardware components of a vision system.*

The attributes of world objects and their effect on image formation are the subject of the next chapter. Interaction with the user is required to allow specification of the active detection zone and the entry of calibration and operational information, shown as the User Interface block. These processes are described in Appendix B which reproduces a user manual written to allow the detection system to be independently evaluated (see Chapter 7). So far as actuation is concerned, once the processing system has reached a decision on pedestrian presence/volume it is a simple matter to turn this into a relay closure to indicate the situation to a crossing controller. The remaining, intermediate, components of the system however merit further discussion.

The combination of the optics, sensor and the electronics to produce the analogue video signal will be referred to as a camera, and is often supplied as single unit. The selection of cameras is considered in the next section which is then followed by a discussion of the image capture and processing system requirements.

3.3 Camera

The performance of a vision system as a whole is limited by the resolution and quality of the incoming image from the camera. These performance requirements have however to be balanced against system constraints such as the computational load imposed on the rest of the system and economic factors. Industrial machine vision cameras, even monochrome ones, cost upwards of several hundred pounds. Economic

constraints therefore determined that this system would have to operate using a basic mass-produced, monochrome, charge-coupled device (CCD) camera currently available at around £100.00 in one-off quantities, when packaged with a good quality lens. As will be seen below, an examination of the shortcomings of these cameras shows many of their problems also apply to the more specialised cameras, so there would, in any case, have been limited advantage in using the more expensive devices.

The decision was taken to operate from a monochrome camera initially, rather than a colour one, as they were less sensitive and a factor of three times more expensive in addition to requiring more processing power to handle the increased volume of data coming from the sensor. There was also a concern that methods developed for daytime operation based on colour would not function well at night under street-lighting. The value of colour as a means of distinguishing objects from the background and ignoring the effects of shadows is reviewed in Chapter 5 with respect to prior art work in this area. Support for the suitability of the monochrome CCD devices comes from the fact that, despite poor image formation under some conditions, a human was found to have no difficulty performing the detection task viewing monochrome video footage of the kerbside waiting area. This test can be seen as a minimum constraint on image quality in that, if the Human Visual System (H.V.S.) cannot operate from the image it is likely that for a machine the task will be impossible.

Parameter	**Value / Range**	**Units**
Field of View – Vertical	55	Degrees
Field of View – Horizontal	74	Degrees
Exposure Control	1/60 to 1/10,000	Seconds
Resolution – Vertical	582	Lines
Resolution - Horizontal	500	Lines
Minimum Illumination	1	Lux
F- Number	2.0	
Focal Length	4.48	mm

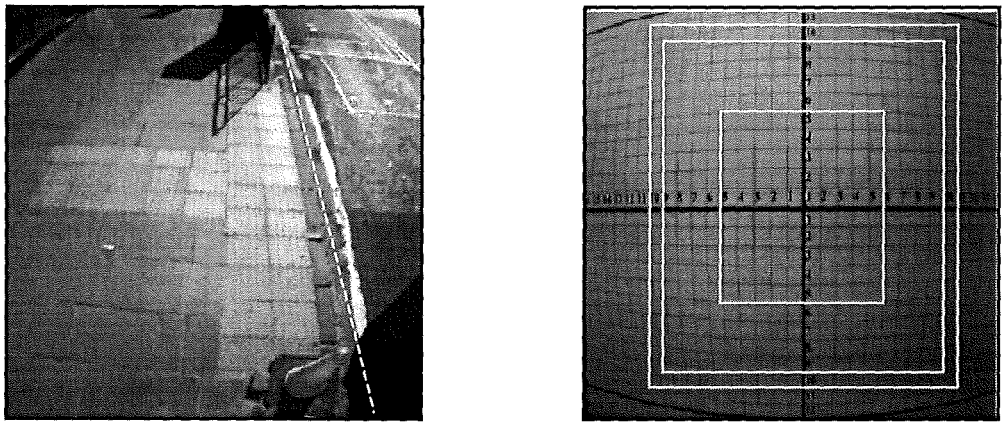
Table 6*: Parameters of CCD Camera used for Development.*

A device of typical performance for such devices was chosen, its main specifications are presented in Table 6. Of particular importance is that it has a wide field of view, which is necessary to allow operation close to the subjects under observation over a large pedestrian waiting area. This situation also constrains the resolution of camera, which must provide sufficient detail on pedestrians at the far side of the waiting area.

The following sections look at three areas of the camera's performance that were relevant to its use for this task namely optical distortion, automatic exposure control and response at low light-levels.

3.3.1 Optical Distortion

A negative consequence of the low cost lens' wide field of view, was the resultant barrel distortion of the image. The standard equations for perspective projection (see Chapter 6) are based on the assumption of a simple pinhole lens model. Under a pure perspective transformation, all straight lines in the scene should map to straight lines in the image. It can be seen however from the sample image shown in Figure 5*(a)* that straight lines such as the white line along the edge of the road are curved. A straight, dashed, line has been overlaid on the image to illustrate the degree of curvature.



(a) *(b)*

Figure 5*: Non-linear distortion due to wide-angle lens is demonstrated by the curvature of the white line at the roadside, (a). In (b) an image of a grid, which should consist of only straight lines under pure perspective distortion, is shown. The overlaid rectangles show how the lens design has kept the centre of the field flat with curvature increasing rapidly towards the edges of the image.*

Figure 5*(b)* shows the degree of distortion on an image of a rectangular grid. The fit of the central 5 by 5 unit square shows that the centre region of the optical field is approximately flat. The outer pair of overlaid rectangles show at 10 units, about 1 unit of error between a line at the x-axis and the same line 10 units above it. As most of the distortion is towards the edges, it was decided to wait and see if it had a bearing on detection performance. If necessary photogrammetric methods are available (Mohr, 1996) that can be used to cancel this kind of non-linear distortion at some computational expense.

3.3.2 Automatic exposure control (AEC)

The vision system will only perform well if it is provided with a clear image under all conditions by the camera. This requires a wide range of exposure control to form images in conditions ranging from bright sunlight to night-time.

Typical scene illuminations vary over a dynamic range of six orders of magnitude. The bottom range of a charge-coupled device's (CCD) sensitivity is defined by the 'dark' current at which point the effect of noise starts to dominate over that of incoming light. The designed illumination range that a CCD can accommodate then typically extends over a range of about 1000:1 which is then digitised to 256 levels. To make best use of the dynamic range the CCD includes an automatic exposure control (AEC) mechanism that varies the integration time, and hence sensitivity, of the sensor pixels to allow it to operate over a wider range of illumination conditions. Automated systems within the camera dynamically set the AEC according to the instantaneous illumination falling on the pixel array and attempt to find a good compromise setting for the range of intensities arriving from the scene. A typical implementation of this is for the camera to adapt to equalise the number of very bright and the number of very dark pixels or to target a pre-set average level. Incorporation of these AEC mechanisms is a consequence of the cameras having been designed for use in consumer video equipment where the goal is to allow it to autonomously form the best image possible without requiring manual adjustment.

The camera used had a wide range of electronic exposures (from 1/60 to 1/10,000 second) and on the whole its AEC mechanism did an excellent job of adjusting the exposure to maintain good image quality in a wide range of illumination conditions.

From a vision system's point of view there were however three problems to be addressed:

- The vision algorithm's lack of knowledge of the state of the camera's AEC process.
- The temporal range of illumination to be accommodated.
- The spatial range of illumination to be accommodated within a single image.

As the AEC process is contained completely within the camera, it is beyond the control of the vision software. Consequently, the adjustments made by the AEC mean that, so far as the image analysis software is concerned, there will be sudden changes in image intensity that are of unknown amplitude and which occur essentially randomly in time. These changes are a function of incident illumination levels interacting with the internal AEC control algorithm used by the camera. Image analysis algorithms will therefore have to be tolerant of this effect.

The second bullet point above refers to the fact that the absolute range of illumination to be accommodated varies greatly between the relatively low intensity of artificial street lighting at night-time through to bright sunlight. Practical experience showed that this range of illumination exceeded the capabilities of the camera's electronic exposure control system. A tool commonly used in vision systems to expand the range of illumination that can be accommodated by an image sensor is a mechanical aperture that can limit the amount of light reaching the camera lens. The use of such mechanical devices however also means significantly greater costs and introduces reliability concerns. The method chosen in this work was therefore to optimise the choice of window material through which the lens received light from the scene. An empirical approach was adopted whereby a fixed level of attenuation was provided by an optical filter over the camera lens that was just sufficient to cope with the brightest illumination conditions. The camera then needed to be sufficiently sensitive to also produce a good image at the lowest (night-time) light levels.

It was found that the insertion of filters necessary to cope with peak solar illumination led to a relatively low-contrast range in night-time images, as can be seen in Figure 6a. However as pedestrian activity was still discernible in these images, it was decided to work with the low-contrast night-time images and to attempt to compensate algorithmically and thus aim to avoid the need to provide active

supplementary illumination at night. The actual amount of filtering used was chosen at a compromise level so as to improve the images obtained at night-time which on occasion resulted in some loss of detail on bright objects, see Figure 7.

(a) (b) (c)

Figure 6 *Contrast Range: Between the capture of the two images above, two and a half seconds have elapsed and lighting levels have remained constant. It illustrates how the contrast between a pedestrian and background scene can vary significantly as a function of the reflectivity of their clothing. The contrast between the central pedestrian and the background in (a) is about 5 grey-levels which is 12 times less than it is for (b) at 60 grey-levels. The third frame is the difference image ((a-b)+128), the common area to the frames is of value 128+-3 grey levels indicating that no exposure change has taken place between the capture of the two frames. Source WEPS6_1_2622-2627.*

For this application the use of a dark filter in front of the lens had a secondary benefit as it makes the camera less visible from the pavement and therefore less likely to attract unwelcome attention.

Referring to bullet point three, limitations on the spatial range of illumination that could be imaged meant there were also some conditions where the automatic exposure control mechanism failed to produce a detailed image. The problem lay in the fact that the exposure control parameter available to the camera's AEC mechanism (and indeed to all current, mainstream image sensors) was global to the entire sensor array. This meant that setting the exposure to correctly image the bright areas caused

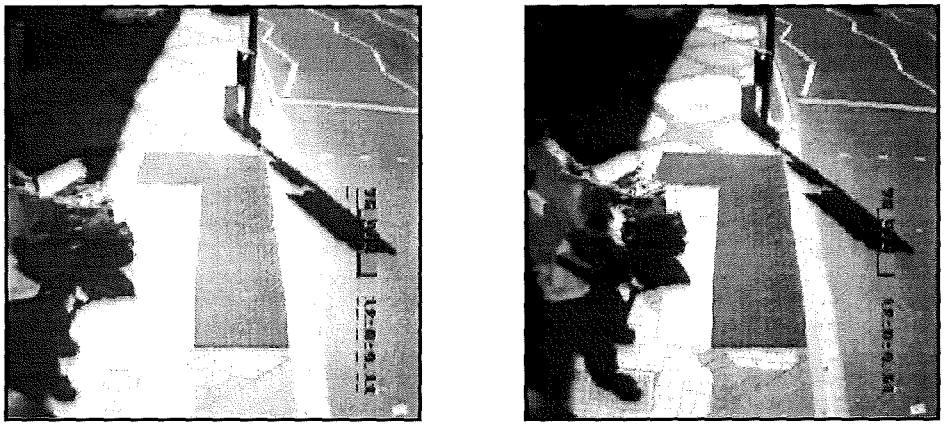
insufficient sensitivity in the dark areas producing uniform black areas with no internal detail. Setting the exposure to see detail in the dark areas however caused saturation in the bright areas of the image also meaning that detail was lost.

(a) (b)

Figure 7 *Over Exposure: This figure shows an effect of the limitations on the exposure range of the camera. When the pedestrians are within the shadow of the signal pole (a), details of their clothing can be clearly discerned. Out of the shadow, the level of reflected light is too great for the upper limit of the camera's exposure control (b) leading to saturation (white areas in the image). Source BC5_1_6456-7.*

This problem typically occurred during bright illumination conditions, see Figure 8a, where part of the image is brightly lit whilst other parts are lie in shadow. Under these conditions, the AEC mechanism tends to choose the middle ground such that often there is little contrast in either the dark or the bright areas. Problems due to this mechanism are exacerbated when the image content is such that the AEC control system sits near a decision level causing image brightness to fluctuate rapidly, Figure 8. Such changes can be triggered by variations in noise levels or, as in the figure, by the arrival of objects.

Further examples of the above effects can be found in the next chapter where they are discussed in relation to the conditions that cause them to become evident.



(a) (b)

Figure 8: *Two consecutive images showing global image intensity change due to the effect of the camera's automatic exposure control mechanism interacting with a building shadow. Under these conditions, the camera can change rapidly between very different exposure settings due to the passing of a pedestrian through the scene. Test images BC4_1_3530, BC4_1_3531.*

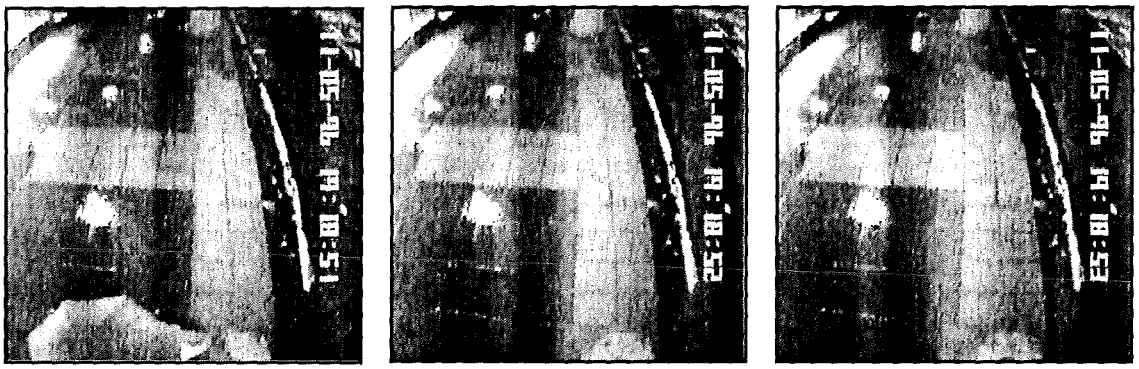
As discussed above the limits on the camera's exposure range necessitated the use of optical filters to reduce the incoming light intensity. Initially a neutral density filter was used. However, learning a lesson from photography, it was realised that the use of a Polaroid filter could provide the required attenuation along with the additional benefit of reducing unwanted reflections in the image.

This was of particular benefit when the pavement surface was wet, as light was partially polarised in the horizontal plane after reflection off the water surface. Therefore as the detector was always oriented horizontally and parallel to the pavement making the polarising axis of the Polaroid vertical meant that the polarised portion of the reflected light was removed. Even when the surface was textured due to fine pavement detail, it was found that a useful reduction in the intensity of reflection could be achieved resulting in an improved image of the underlying scene.

3.3.3 Banding at Low-light Levels

At night, dark vertical bands occur moving horizontally across the image as shown in Figure 9 where a band is initially at the centre of the image *(a)* and moves to the left *(b), (c)*. The images shown have been enhanced by histogram equalisation to make the effect more evident. The actual level of darkening within the bands is about 10%.

These bands move at a constant rate of one passage of the image every 4 seconds. Their contrast was low as were the spatial gradients at their edges so it was hoped that algorithmic methods would be able to overcome the effects of this artefact. In practice, it was found that the algorithms developed were unaffected by them.



(a) *(b)* *(c)*

Figure 9: *Camera banding artefact which occurs at low light levels is a dark vertical band traversing the image (Centre of (a) moving progressively left in (b) and (c)). The frames were captured at approximately half-second intervals Source WEPS6_1_9650,52,54*

More detail on the operational characteristics of the camera in real-world environmental conditions is given in the next chapter which, among other things, looks at the effects of changing weather on lighting conditions and image formation.

3.4 Processing System

Research into computer vision and image processing is computationally demanding due to the high data rates and volumes involved and as such has proved a fertile area for the development of high-speed processing architectures, particularly parallel ones. In the past, this constraint has meant that research into vision tasks has been closely linked to customised hardware design which was often a prerequisite to achieving sufficient speed of operation for algorithmic ideas to be evaluated. Even with the increasing availability and power of programmable processing systems vision research nowadays often has to address the question of balancing flexibility against achieving the required speed of operation. The former represents the strength of

software the latter that of hardware. The chosen balance may differ significantly according to the phase of system development in particular whether the effort is directed at development and proving of the method or at final, high volume, implementation. For development, we are concerned with flexibility at sufficient speed of operation.

For the purposes of development and test during this work, the standard PC platform was chosen. In performance terms, this offers very low costs for the power available due to the enormous volumes they are being built in amidst strong competitive pressure amongst manufacturers. The PC platform is also a very convenient selection from the point of view of software development due to the wide range of peripherals and programming tools available. This choice was further supported by the availability of embedded PC systems, which offered a convenient route from a desktop development environment into a compact unit of industrial construction standards (see section 3.6) for on-street trials.

Given the above, the processing capability of a state-of-the-art embedded PC was taken as a measure of the amount of processing capability under which real-time operation (i.e. the desired 500ms response time) must be achieved. It was assumed that, if operation could be achieved in this environment, subsequent transfer to an alternative uni-processor system would then be viable. As image sequence analysis is well known for its heavy demands in terms of computational load, the limitation to a PC environment was a significant constraint on the detection algorithm's design. The author's response was to seek computationally efficient computer vision methods to compensate for this shortfall in processing power. These are discussed further within Chapter 6 on algorithm development.

A method was developed to remove the real-time constraint from the development process by the use of extensive pre-digitised sequences. This also brought further advantages for development of image analysis methods by allowing repeatable, off-line evaluation. The advantages of such an approach are confirmed by the increasing calls in vision research circles to standardise algorithmic test and comparison. The method developed for capturing these sequences and their role in algorithm development is covered at greater length in Chapter 7 on evaluation.

3.5 Image Digitisation

Throughout the duration of this work commercial frame-grabbers became ever more widely available at decreasing cost. This trend was driven by the arrival of mass-produced chip-sets aimed at video acquisition and processing for multimedia consumer products. The commercial frame-grabbing boards used in this work were all based on the Chips and Technologies PCVideo chip set. A benefit of boards based on this chipset was that they incorporated hardware scaling functions intended to simplify the sizing of on-screen video display in multimedia products. These scaling functions were used for performing the resolution control functions (described in Chapter 6) without incurring any software processing overhead.

3.6 Migration from Development to On-Street Trial System

An important reason for choosing the IBM PC (as opposed to workstations) as the basis for the laboratory development system was the existence of a large number of commercially available boards offering a convenient route to embedded implementation. The system pictured in Figure 10 was put together in this way for use in on-street trials of the detection algorithms. The standard PC104 board is approximately 3 inches square and contains all the features of a standard PC motherboard including disk controllers, serial ports, parallel ports, video drivers and keyboard connection. These connections have been brought out to the front panel in the system shown. Expansion is achieved by vertical stacking on the PC104 connector (which is a reformatted version of the PC's standard ISA bus). In the figure, expansion boards have been added to provide a frame-grabber, a bootable solid state (flash) drive and a relay driver. The casing shown is a standard 3U 19 inch rack case designed for mounting into a traffic control box and contains two PC104 systems (one for each side of the test crossing) as well a switch mode power supply. Figure 11 shows the interior of the roadside cabinet used for trials at a test site in Bracknell.

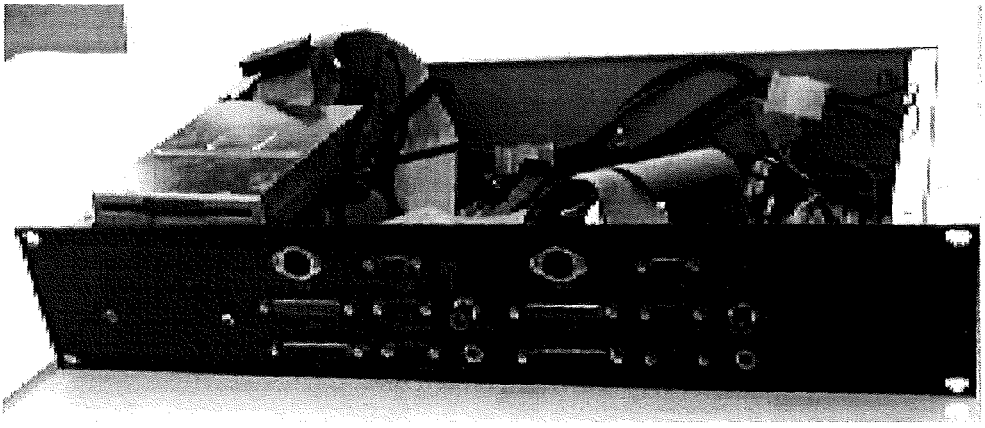


Figure 10: *PC104 embedded PC system used for in-street trials of algorithms.*

Software developed on the desktop PC system was transferred to the embedded system by attachment of a disk drive or the temporary addition of a network card onto the PC104 stack. The only alteration required to the software was a recompilation with the appropriate drivers for the PC104 frame-grabber and the sections of the program that drove the relay card enabled.

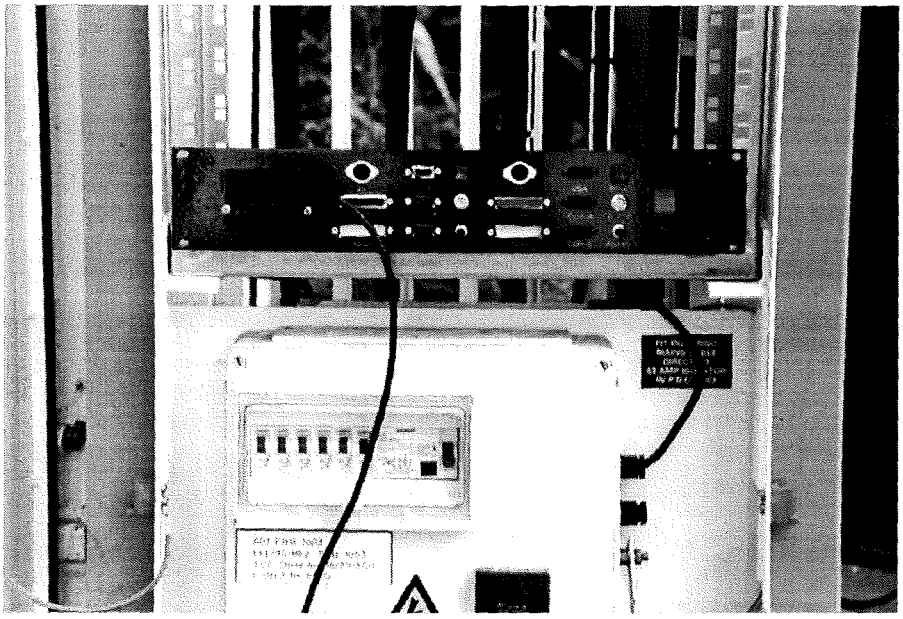


Figure 11: *19inch roadside equipment cabinet used for on-street trials of algorithms*

The processor on the motherboard was a 486 DX4 running at 25(75) MHz. As noted above the processing capability of this device imposed an effective upper limit for processing power available to the algorithm for on-street trials.

It should be noted that this route to an embedded system is expensive. The system shown cost around £1500 to assemble (for each PC104 stack) however this was

justified in the context of development particularly given the high level of compatibility with the laboratory development system. Routes to higher volume production are discussed below.

3.7 Migration from Development to Commercial system

In the early stages of this work collaboration with a local company led to experimentation with a new form of CMOS image sensor developed by VLSI Vision Limited (now Vision Limited) from research at Edinburgh University. This sensor differs from solid state imaging devices based on charge coupled device (CCD) technology in that it can be produced on a standard CMOS production facility. Aside from the benefit of decreased production cost the main advantage of these devices is that, any silicon area unused by the imaging array can be used to include on-chip processing of the image data. The manufacturers have demonstrated this by integrating fingerprint recognition onto a single silicon die (Anderson 1991) and producing a range of single-chip application specific image sensor (ASIS) cameras. Such devices combining sensing and processing functions are commonly referred to as Smart Sensors.

The advantages of the use of this new technology were examined by looking at image capture using the ASIS devices. Normally a frame-grabber must extract synchronisation clocks from the analogue video stream to digitise it. This adds considerable complication to the capture hardware. The ASIS device however provided these signals as additional outputs to save the capture system from having to extract them. Most importantly, it supplied a non-standard pixel-clock synchronised to the ideal pixel sampling instants in the video stream.

Three generations of frame-grabber were designed and built, by the author, to capture images from the ASIS family of cameras. Details of the final design of this board can be found in Appendix C. This board demonstrated how an image capture system, complete with interface to a PC, could be designed for a component cost of only £50. It is pictured in Figure 12 and can be seen to consist of only three integrated circuits and a few passive components. This degree of compaction was achieved by the use of a Field Programmable Gate Array (FPGA) to contain all the interface and control logic. This was at a time when commercial grabber systems were only available for a cost of at least ten times this figure. The final version of the frame-grabber, along with

a version of the software described below, is currently used for the teaching of image processing within the university and has formed the basis for various hardware projects.

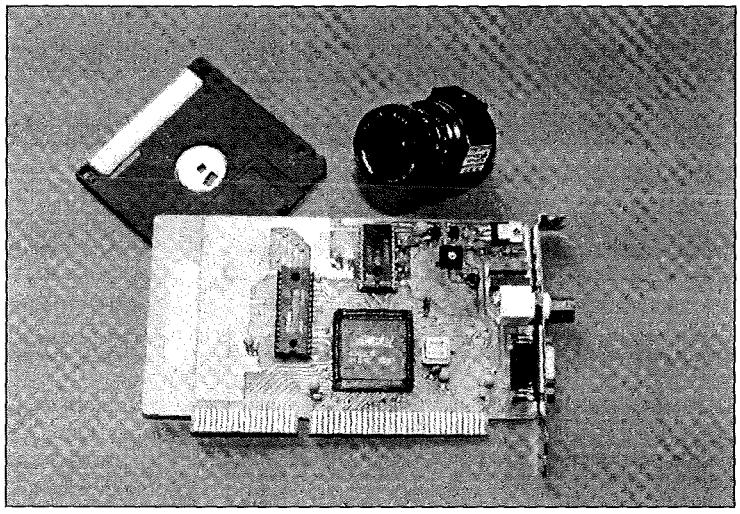


Figure 12: *Frame-grabber designed for ASIS image sensors*

The conclusion of looking at this technology was that the benefits of CMOS technology for image sensors with respect to CCDs only come into play when large production volumes are being considered. It is only at this point that the small difference in price becomes important and the effort involved in custom exploitation of integrated processing circuitry becomes justified. Therefore, although the CCD camera was considered the more appropriate choice for development (it offered significantly better image quality), the use of Smart Sensor technology may be appropriate as a route to large volume production once algorithms are mature.

So far as processing is concerned, in the authors opinion, a custom hardware implementation is only justified when algorithm design state has been developed to the point that its performance has been proven and can therefore be frozen or when it is the only way to allow algorithm development to proceed (e.g. when it is being hampered by a lack of speed). In the past, situations where custom hardware design has been necessitated were more frequent as limitations in available general purpose processing power meant it was otherwise not possible to properly develop and

evaluate algorithms for real-time video applications. The benefits of hardware implementation only become clear when a customised architecture is to be generated to accelerate particular specialist functions that are not handled well by general-purpose processors.

Nowadays customised hardware design is becoming much easier and is now a routine engineering task for volume implementation. A significant factor in this is the advent of hardware description languages (e.g. VHDL) and accompanying design routes to FPGAs (as used in the frame-grabber design described above) and ASIC (application specific integrated circuit) implementations. These routes are well established with the choice between them being a question of production volume. Despite the above, custom design is only economic for manufacture in volume once the necessary design and test effort have been considered.

As the development system is constrained to operation on a standard desktop PC and as there are many DSP devices of comparable and superior processing power at low cost, there is every reason for confidence that no specialist image processing hardware will be necessary for commercial implementation of the algorithms. This belief is borne out by recent exploitation work carried out with industrial partners (see Chapter 8).

3.8 Casing and Mechanics

Mechanical casings to enable on-street evaluation and capture of test video were produced by industrial collaborators as were all the necessary fittings to attach it to signal poles. Figure 13 shows the casing used for the majority of the work which was supplied by Microsense Systems Limited. The left of the dual front windows is used by the camera the other is unused (the casing was originally designed for a different purpose). Insertion of the electronics is via a removable base plate through which connections to the outside world pass. The front window has six screw points to allow attachment of optical windows in front of the camera. All removable sections are waterproofed by compressed rubber seals.

Figure 13: *The camera casing used for on-street trials and video data-collection shown installed at the Princes Street test site.*

3.9 Vision System Software

The other essential tool required for this task was a software environment to tie all the hardware together. Accordingly a package, which will be referred to throughout this document as the Multi-resolution Image Sequence Analysis (MISA) environment was written by the author. As with the hardware development it was important, that this was flexible enough to support development but also portable for use in an embedded trial system. The term multi-resolution refers to a means of reducing computational load by reducing image resolution to the minimum level sufficient for a task (see Chapter 6).

Initial work was under the 16bit DOS operating system that rapidly became limiting due to the small amount of PC memory that could be accessed and the lack of a debugging facility, which made software development difficult. This code was therefore converted to operate under 32 bit extended DOS which eliminated memory limitations, however it was still not possible to debug programs as compiler and third-party debuggers all failed to work on programs producing graphical output. In the end it was found necessary to move to a Windows environment to permit debugging, however the multi-tasking features of Windows meant that these programs were not suitable for time-critical operations associated with functions such as image capture.

The eventual solution was to write versions of MISA to run under two different operating systems, DOS32 and Windows, for use according to how their characteristics matched particular tasks. The roles of these versions are summarised in Table 7.

Required Function	Windows	DOS
Program Debugging	Yes	No
Operation in Embedded System	No	Yes
Real-time Control for Sequence Capture	No	Yes
Device Independence for Inter-machine Portability	Yes	No
Executable Program Size	~2Mb	~500Kb

Table 7: *Comparison of merits of DOS and Windows operating systems for various tasks*

The Windows version was used for development because of its debugging support and the additional benefit of device independent graphics that made moving between host PCs much easier (the DOS32 version required different drivers according to the graphics card installed). The lack of reliable real-time response was not a problem in this role as it operated from pre-captured sequences and so was buffered from all real-time constraints.

The DOS32 version was used where accurate real-time control was required during sequence capture and for porting algorithms from the development system to the embedded system for on-street trials. Aside from the real-time control problem, it would in any case have been impractical to run Windows on an embedded system.

Although no comparative measures have been made it is also probable that the DOS version executed faster as there was less overhead in maintaining the graphical display and dealing with background event messages. The software was designed with a set of 'switches' to allow various display-oriented functions to be selectively disabled. Thus during development, maximum feedback on algorithm operation could be displayed to the user, whereas once ported to the embedded system they could be disabled to maximise the speed of operation. Other measures used to maximise execution speed were the pre-calculation of loop constants and the almost complete

avoidance, throughout algorithm design, of floating point arithmetic. Other functions of the software for performing roles such as automated evaluation, relay driving and the generation of log files were similarly controlled by software switches.

The programming language used was initially C and then progressed to C++. C++ is an object-oriented language such that data is encapsulated with a set of functions to manipulate it as an entity known as a class. Once the class' structure had been established advantages of development speed, reliability and portability were gained through modularisation of the image class and its separation from display and capture functions - which were operating system and hardware dependent respectively. The use of C++ is sufficiently widespread that porting the software to other platforms would not be difficult. For a commercial implementation it may prove necessary to convert back to standard C as most digital signal processor (DSP) systems only support assembly language and C compilers at present.

The Symantec C++ compiler was chosen in preference to the more widely used Borland and Microsoft compilers as it was the only one that allowed access to memory by physical address. This was important as it provided a means of accessing image data from the commercial frame-grabbers, which was memory-mapped into part of the PC's address space.

Figure 11 gives an overview of the class hierarchy that constitutes the basic object set of the MISA software suite. It can be seen that the classes derive from three stems. The class Point and its base class Location encapsulate the display-related properties of each derived object. The List class and its base Instance encapsulate attributes that allow derived classes to be incorporated in linked lists. The Mask class was used to handle user specified detection zones. A brief description of the roles of the most important classes for image handling is given below:

- Image: This object encapsulates the handling and processing of image data. Its data members consist of a two-dimensional array of image data, an array of pointers to index into the array and parameters specifying the image's dimensions. A set of function members (around 180 of them at present) has been defined to perform image processing operations on the data.
- ImageStream: The flow of image data through the software is handled by this class which makes the application programs' communication with

sources/destinations of images transparent. Algorithms can therefore be written independently of the image source/destination type. This makes it relatively easy to add further image capture options without restructuring application software.

- ImageInStream is derived from the ImageStream base class and supports input from a variety of hardware and stored image sources. Hardware sources are either frame-grabbers based on the PC video chipset or one of the in-house developed video capture systems that were designed as part of this work (section 3.5). The normal mode of operation during development was however to operate from sequences of images stored on disk or CD-ROM in either Windows device independent bitmap (*.BMP) format or a more compact proprietary (*.BIG) format. Such sequences are captured in real-time to CD-ROM, using the method described in Chapter 7 and then stored in an MSDOS directory with the filenames consisting of a stem name prefixing a sequentially numbered identifier.

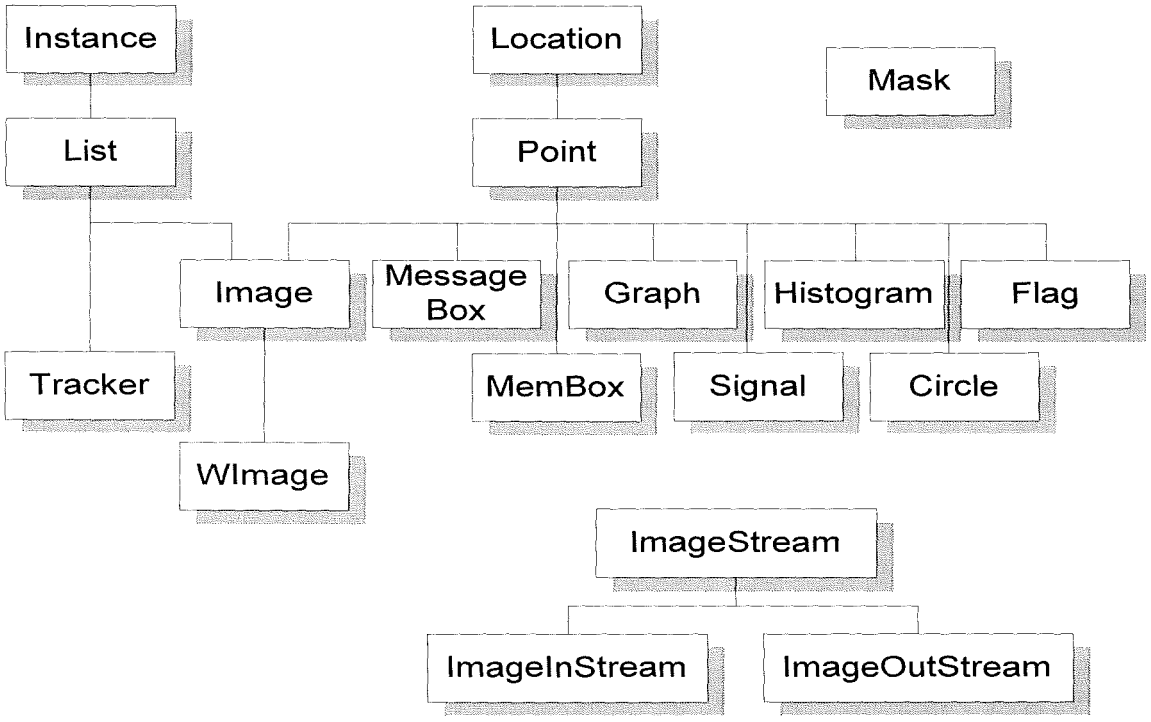


Figure 14: *Class hierarchy of the Multi-resolution Image Sequence Analysis (MISA) software. Base classes are shown at the top with those inheriting their characteristics linked from below.*

- ImageOutStream: The organisation of this class is very similar to ImageInStream and shares common characteristics for indexing through a sequence of images

derived from the ImageStream base class. Output image streams are directed to file storage using either of the image formats defined above or the additional option of output to an Excel compatible format for examination in a spreadsheet. The main uses for this class were in sequence capture and the selective capture of partially processed images from any stage of the algorithms, for later review.

Although they are not illustrated above, the MISA package developed also included many additional objects to encapsulate particular algorithmic functions. Examples are objects to perform background modelling, detection zone mask specification and 3-D pedestrian modelling.

3.10 Conclusions

This chapter described the selection and design of a set of equipment and software tools to act as a platform for the development and implementation of computer vision algorithms for pedestrian detection. It also described the consequences of these choices for the algorithm design phase which is described in a later chapter.

Although the main concern was to define the development platform, it was also necessary to provide a route to an on-street trial system and to use components that would be commercially realistic for a practical detector. A PC based development platform was selected and the route described for porting to an embedded system to allow on-street trial of algorithms. The constraint of commercialisation realism had the effect of limiting the camera used to being a mass-produced, monochrome CCD and of restricting the available processing power.

The difficulties due to the choice of camera were described with respect to the environmental conditions that caused them to become prominent. In particular the effects of the global control and limited dynamic range of the camera's exposure control system were highlighted. This showed that pedestrian detection algorithms would have to be tolerant of sudden, unpredictable changes in global image intensity and also of a wide range of pedestrian contrast. Rather than using additional components such as mechanical apertures or supplementary lighting to address these difficulties it was decided to seek algorithmic methods to compensate for the image sensor's deficiencies.

Software for image sequence manipulation (MISA) was written to provide an environment for the development of algorithms. In particular the requirement to

analyse sequences and to process image representations of varying resolution and field-of-view necessitated its writing as no commercial and available package offered these features. This had the advantage that there was complete freedom of access to the code without commercial limitations being imposed. Once developed and tested, it has proven to provide a stable basis for experimentation and extension of its facilities. It also offers ease of portability between environments due to its containment of the image processing functionality in an underlying set of C++ image classes that are platform and interface independent. MISA has been used by several other Napier researchers in their computer vision work and a cut down version is in active use as a teaching tool, all of which are testaments to its usefulness.

The next chapter goes on to examine prior knowledge of the task and identify difficulties the vision system had to overcome.

4 Analysis of the Vision Detection Task

4.1 Introduction

In recent years, an important realisation in the field of computer vision and the related fields of artificial intelligence and robotics has been the value of designing systems that are oriented at specific real-world tasks. This contrasts with earlier work, which attempted to create general-purpose systems capable of 'vision' or 'consciousness'. The value of this change in emphasis from general-purpose to task-oriented design is supported by observations of biological visual systems which have evolved in many very different forms each specialised according to the environment and behaviours of the particular organism of which they are part.

In the case of vision the objective of the general-purpose approach was to invert the image formation process into a complete description of the 3D world under observation (Marr 1982). The intention then was to extract the required information for any particular task from the resulting complete object-based representation. Work following this approach found that, for all but the simplest and most artificial of environments, the complexity of the inversion process and extremely high computational demands made gaining the world description a practical impossibility. Even extensive work in the field of parallel processing, often prompted by the demands of vision research, failed to overcome this barrier to progress.

In contrast, a task-oriented approach utilises the particular constraints of the operating environment and an understanding of the level of scene description required to achieve a particular goal as a means of limiting the system's complexity. The vision system's work is then confined to the extraction of a set of task specific measurements from the image that are sufficient to perform the task in hand. In robotics this approach is exemplified by the work on 'subsumption' architectures (Brooks, 1991)

and in computer vision much of this activity is described under the banner of 'active' vision (Blake, 1992). In practical terms, the active vision approach manifests itself in the use of methods which direct a vision system's attention spatially, temporally and in resolution, according to task specific bounds. Image features are then analysed at a level of detail just sufficient to achieve the measurements required of the vision system, resulting in a reduction in the complexity and quantity of computation required (e.g. Burt, 1991).

The pedestrian detection task, as specified earlier, is extremely demanding in computer vision terms, as it has to deal with multiple, deformable, moving objects in an outdoor environment. Despite this, it must be implemented at very low cost and hence with limited computational and sensory resources. This state of affairs led the author to adopt a strongly task-oriented approach. Therefore, wherever possible, knowledge of the problem domain was incorporated to aid in generating a solution - rather than trying attempting to solve the vision problem in a general way.

As part of this approach, the principle goal of this chapter is to collect and examine prior knowledge of the detection task with the aim of identifying information that could be used to constrain the task of the vision algorithms. This body of information includes knowledge from the literature, from the author's observation and that which is encapsulated in the objectives given at the end of Chapter 2. This analysis provides the basis for the following:

- The assessment, in the next chapter, of related prior art in the field of computer vision where comparison of the constraints of the current task and those in previous work is used to assess their relevance to this work.
- The identification of the main opportunities (distinguishing characteristics between scene components) for exploitation by the vision system in separating pedestrian from non-pedestrian objects, in Chapter 6 on algorithm development.
- The determination and characterisation of environmental factors that affect image formation and the consequent identification of the major problems that the vision system would have to overcome was used in system and algorithm design (Chapter 6).

- The definition of a test data set, based on an examination of the range of operating conditions to be tolerated, that was sufficiently representative as to allow an expectation of reliable operation in long-term use, see Chapter 7.

The next section starts by describing the structure of the localised scene in which the detector must operate. In this section (4.2) the parameters and terminology used to describe the crossing environment are defined. In addition a set of assumptions that were made regarding the scene structure and its relationship to camera and pedestrian position are justified and the relationship between the required camera field–of-view and pedestrian size discussed. The consequences of the resulting projection of the scene into the camera image are then illustrated in section 4.3.

The characteristics of pedestrians and other objects in the crossing environment are then examined (section 4.4). As the vision system was essentially required to perform a classification task, the characteristics that were of most interest were those that most strongly distinguished pedestrians from other object types that might appear, either directly or indirectly, in the image. The important other object types (e.g. crossing infrastructure, pushchairs, wheelchairs, litter and leaves) are identified and their characteristics studied.

Section 4.5 then goes onto discuss the varying illumination and other environmental factors under which the detector must operate and thereafter the nature of the resulting visual artefacts (such as shadows and reflections) produced in the video image (section 4.6).

Data on the attributes of the transportation system in terms of pedestrian and vehicle behaviour are also presented with a view to identifying factors that might assist in the extraction of pedestrians (section 4.7).

It should be noted that much of the data presented in this chapter is derived from digitised test sequences. In the figure captions, reference information is given which indicates which test sequence the information is taken from and the position of any particular video frames within the sequence. In Chapter 7, a description can be found of the characteristics of these sequences.

4.2 Scene Structure

Schematic representations of a detector observing the kerbside waiting area of a crossing are given in Figure 15 and Figure 16. They define the terminology that will be used to describe the crossing infrastructure in this document. The details of the dimensions and detector mounting heights of the test sites used in this work can be found in Appendix D.

The following working assumptions were made with regard to the scene and its relationship to pedestrian position:

1. The signal pole was at right angles to the surface of the kerbside.
2. The kerbside area was flat over the detection zone to be monitored.
3. The near edge of the detection zone would start from a point vertically below the camera lens.
4. No structural alterations could be made to the scene to assist the operation of the vision system.
5. The reference point for a pedestrian's position would be defined as a point on the ground plane directly below their head.
6. The full height of pedestrians standing in the detection zone should fall within the camera's field-of-view.

Although the first of these assumptions will not be true of all sites (i.e. those on hills of appreciable gradient) the work described in this document could be adapted to deal with such cases without too much difficulty. The second assumption was made to simplify the task of relating objects' world positions to those in the camera's image plane. It is valid in all cases of the author's experience. The third reflects the fact that the near edge of the detection zone should start at the pedestrian request push-button which is mounted on the same pole as the detector.

Concerning the fourth assumption, most waiting areas have various structural features such as textured (tactile) paving and lowered kerbs that are designed to aid (particularly disabled) users of the crossing. The introduction of additional features to aid detection (as has been used by previous workers, see Chapter 5) is not acceptable to the Department of Transport.

It was further decided that in this work, the position of a pedestrian would be defined with respect to a point on the pavement directly below their head. This reference was chosen in preference to head position (which is more likely to be visible) as the user-specified detection zone is also referenced to the ground plane.

Finally, it was observed that if for example only the lower part of a pedestrian was visible, then the judgement of pedestrian presence would be become less confident, being based on a reduced amount of information.

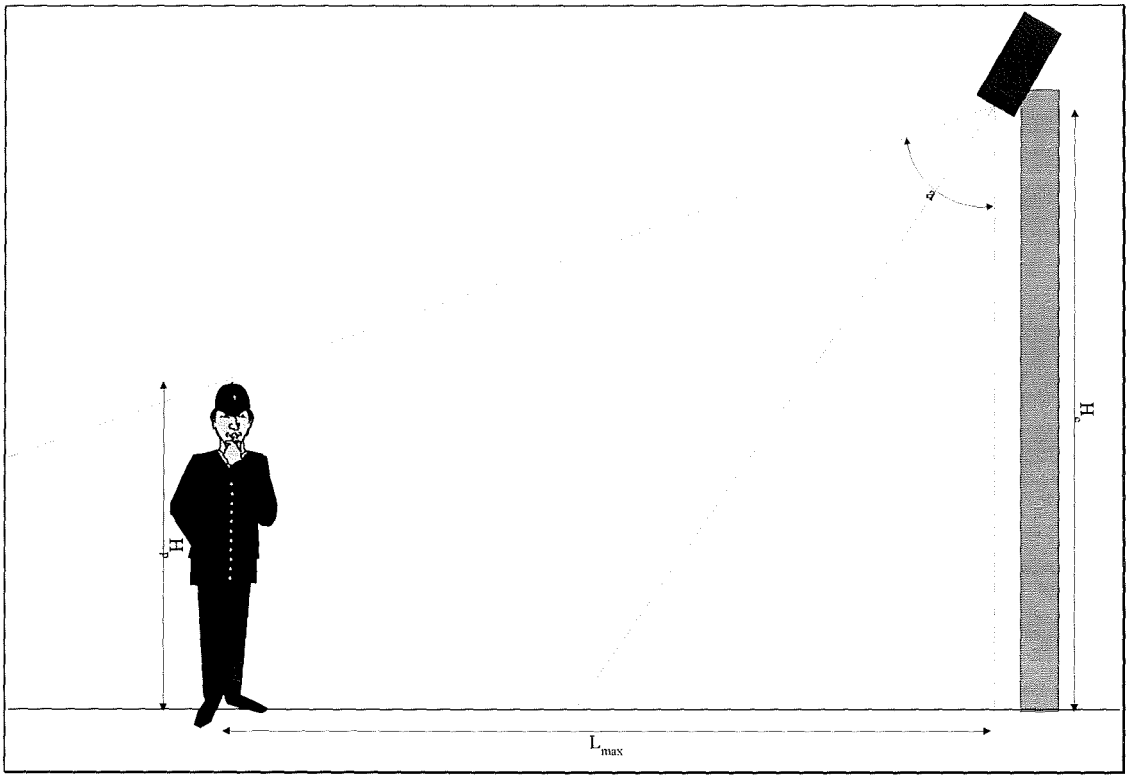


Figure 15: *Side view indicating how constraints such as pedestrian height and the camera's field of view and mounting height are linked to the length of detection zone that can be used.*

It was therefore assumed that the entire projection of a pedestrian standing at the far side of the detection zone must be visible in the image. The camera's field of view had therefore to extend over a much greater range than the detection zone - by an amount dependent on the maximum pedestrian height, see for example **Figure 15**. This introduced a further difficulty in that unwanted background image activity, due to pedestrians and vehicles falling within the field of view, but being outside of the detection zone, needed to be ignored by the vision system.

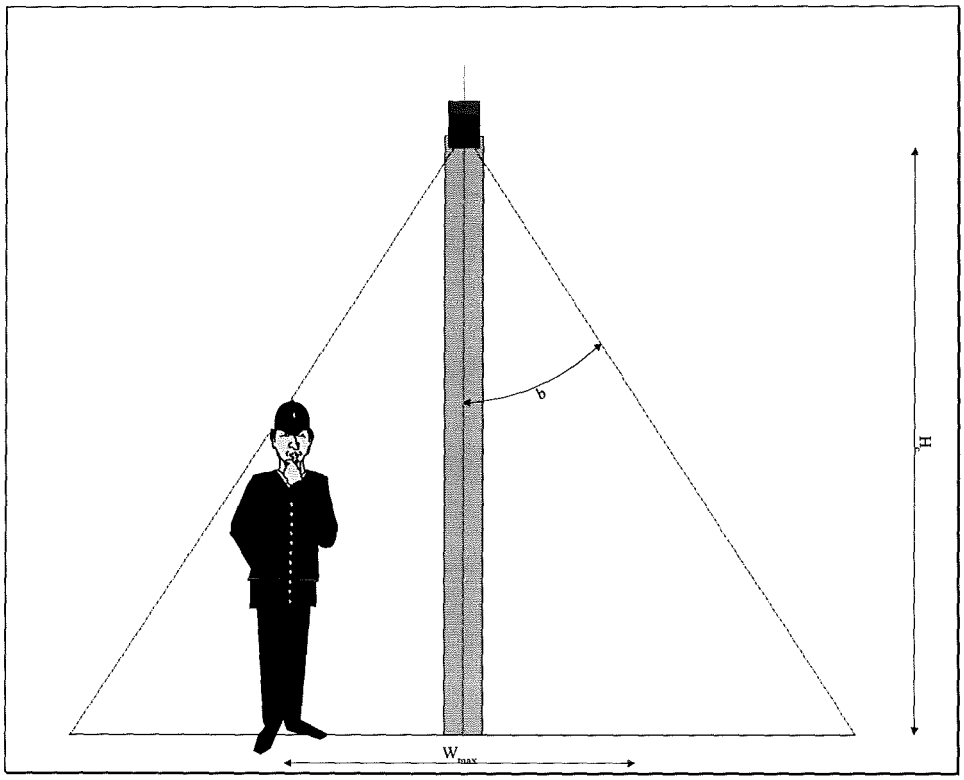


Figure 16: *View parallel to the roadside showing how constraints such as pedestrian height and the camera's field of view and mounting height are linked to the width of detection zone that can be used.*

The camera selected for this work (see Chapter 3) had been designed for compatibility with television standards and therefore its vertical and horizontal field-of-view corresponded to the standard aspect ratio of 4:3 used by television screens (being 74° horizontally and 55° vertically). However, for this task the vertical field-of-view needed to be the larger and so the best match of the camera's aspect ratio to that of the scene was achieved by using it on its side.

A common simplifying assumption used in prior work was that the camera had a plan view of the scene. Under these circumstances, particularly from a high viewpoint, objects under observation all appear at the same scale and do not occlude each other. For this task however the projection of the scene into the camera image is more complex. Consequently occlusion is commonplace and the effects of changing viewpoint and perspective throughout the detection zone become significant - and must be taken account of in the detection algorithms. These matters are discussed in the next section.

4.3 Image Projection

The image formation process can be viewed mathematically as a projection of three-dimensional space onto the plane occupied by the image sensor. This process is known as a perspective transformation (Mundy, 1992). It produces the familiar distortions of the world seen in photographs and paintings with a reduction in scale with distance from the observer and a convergence of parallel lines to a horizon line.

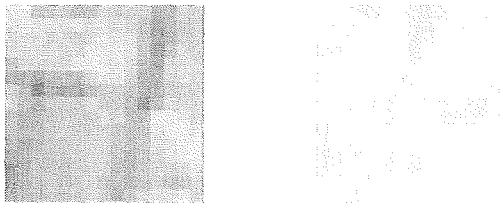
In many applications of computer vision, various assumptions are used to simplify effects of the perspective projection. Orthographic projection, for example, removes the effect of scaling with distance and affine projection preserves the parallelism of lines that are parallel in the real world. A special case of affine projection is the, frequently used, weak perspective assumption which assumes that the distance to an object is large in comparison to its dimensions. In this case, the approximation used is an isotropic scaling of the object according to the distance to a reference point within it, again preserving parallelism of lines.

However, as described above in the section on scene structure for this task the camera was positioned close to the objects under observation and furthermore their size was significant with respect to their distance from the camera. In consequence, there were strong perspective effects leading to significant variation in scale and the loss of parallelism.

Figure 17: *The effects of perspective causing a change in scale and viewpoint as a function of pedestrian position in the scene are illustrated by the above sequence. Underneath the pedestrian is shown clipped from each frame to make the effect on scale, viewpoint and image resolution more apparent. Source WEPS4_1_3068-88.*

In addition to scaling effects, the camera's proximity to the scene meant there was a large degree of variation in the camera's viewpoint of objects. The view of a pedestrian ranged from a plan view when they approached close to the camera mounting-pole to almost a side-on view at the far edge of the detection zone. This meant computer vision methods based on feature matching were unlikely to be useful over extended trajectories, as details of pedestrian images could not be relied upon to be consistently visible as they moved throughout the scene. An example of the scaling and viewpoint changes of a pedestrian image between near and far points from the camera is given in Figure 17.

The proximity of the scene and objects to the camera also meant there was significant spatial variation in the effective world resolution and so the image features discernible at the front and rear of the scene were very different. For example, the textures of the underlying pavement (Figure 18) and the detail visible on pedestrians' faces and bodies varied significantly. This indicated that feature analysis based image processing methods would be limited to only those features which could be expected to exist as a detectable pattern over the full range of viewing scales.



(a) (b) (c)

Figure 18 *Scaling of Texture: Sub-images of block paving illustrate variation in texture with increasing distance from the camera (a) foreground, (b) midground, (c) background. Extracted from test image BC4_1_9998.*

Finally, a further important consequence of the camera's position with respect to the scene was that the viewpoint was sufficiently oblique that pedestrian occlusion of each other was the rule rather than the exception. It was therefore not valid to assume the entire pedestrian outline would be consistently visible to the camera. This ruled out the use of vision methods that match detailed models of pedestrian outline to image data. The high degree of self-occlusion that occurred with variation in pose would also be problematic for some of these methods.

4.4 World Objects Categorisation

From video analysed in the course of this work, various significant types of object that might occur in the scene (world system) were identified. As was identified in section 4.1 the vision system will be required to distinguish between pedestrian and non-pedestrian activity in the scene. Accordingly the following discusses the most important characteristics of pedestrian structure and behaviour (with respect to this task) and thereafter the attributes of non-pedestrian objects.

4.4.1 Pedestrian Attributes - Structure

The 3D structure of a pedestrian has a large number of underlying degrees of freedom which result in a highly deformable shape. Comprehension of this shape is further complicated by its being projected into the image plane. The collective state of the components of a human body is referred to in the literature as its pose.

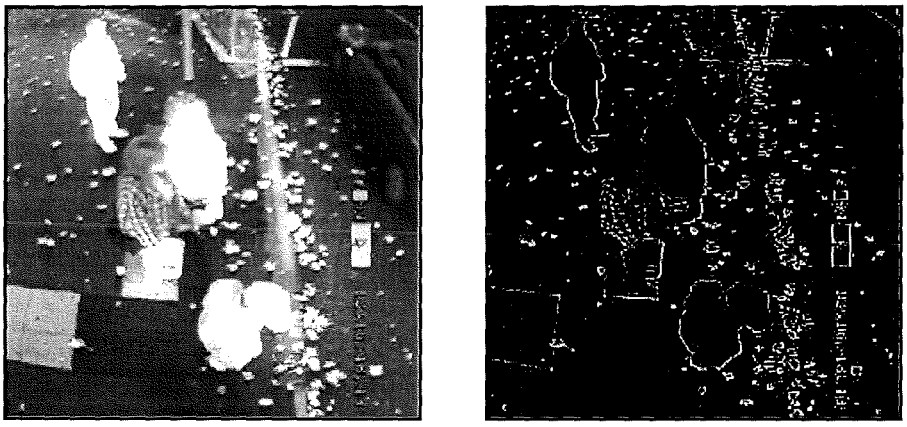
A popular approach used in the past has been to model the body by breaking down the pose to a set of rigid components (upper arm, forearm etc.) connected at joints of varying degrees of freedom. Even the simplest of such jointed cylinder model as employed by previous researchers allows an enormous number of possible body poses (Attwood, 1989). Some work has reduced the number of poses by only considering a restricted range of movements (e.g. constant velocity walking or a certain set of dance movements).

Such simplifications are unlikely to be valid for this work as account will have to be taken of the full range of pedestrian behaviour, as well as the effects of the deformable nature of clothing, see below.

A factor in shape variability not mentioned in any of the prior work, is that of individual variation in body shape. There is however, clearly a large difference in 3D form between the tall and thin and the short and overweight. In the prior art it is normal to concentrate on people of "normal" build.

The effects of variation in attire (e.g. coats, hats, skirts) is another area which has been largely ignored in previous work. The effects of attire manifest themselves in two ways: shape and patterning. Although much normal clothing fits closely to the body and only adds slight distortion to the basic body shape, looser clothing (such as coats and jackets) generally follows body shape but adds considerable distortion. Free

hanging areas of coats will, aside from perturbations due to the effects of wind and inertia, hang vertically under the influence of gravity.



(a) (b)

Figure 19 *Patterned Attire: This figure shows some of the practical difficulties in the use of shape analysis algorithms based on fitting image edges to detailed body component models. The over-exposed pedestrian in (a) presents only an outline in the gradient image (b) that would be insufficient for many of these model-matching methods. The checked jacket shows how the majority and strongest of edge images can be due to surface colouration rather than being related to the physical boundaries between objects as is usually implicitly assumed in these methods. Similar problems for these methods may occur due to surface features on pedestrians (such as creases in clothing) and surface features on the footway due to leaf patterns and textured paving (see bottom left of image). Source WEPS6_1_6801.*

The other important effect of attire is due to surface markings on clothing which are often responsible for stronger edge discontinuities in the image than the boundaries of pedestrians, see Figure 19. Similar problems can occur as a consequence of creases in clothing which may produce edge features, particularly when they are enhanced by localised shadows in bright sunlight. Items such as hats obscure head shape so that this is not suitable as a reliable basis for detection.

Carried items (such as bags, and trolleys) are a further source of distortion of the shape of pedestrian outline. As these items are associated with the accompanying presence of a pedestrian, it is important that they should not lead to a failure to detect the pedestrian.

4.4.2 Pedestrian Attributes - Behaviour and Motion

Pedestrians are on the whole able to move freely within the detection area, although barriers are sometimes used to separate different streams of pedestrian movement. They are liable to move around in a fairly random manner, particularly whilst waiting to cross, providing no basis for assuming flow will be restricted to particular directions of movement. Furthermore, pedestrians may be moving or static and will not necessarily be performing regular walking rhythms when they are in motion.

An observation that may be of value, however, was that the maximum time a pedestrian (as observed in the image plane) remains completely motionless is limited to a few seconds. This information is used to discriminate between transient objects such as pedestrians and more permanent objects in the scene using the 'persistence' filter described in Chapter 6.

The graph in Figure 20 reproduced from Smith, 1993 presents data gathered during railway research by Ando, 1988. It relates age to walking speed for free flow conditions i.e. where the movement of each individual is unimpeded by obstacles or by the crowding effects of other pedestrians.

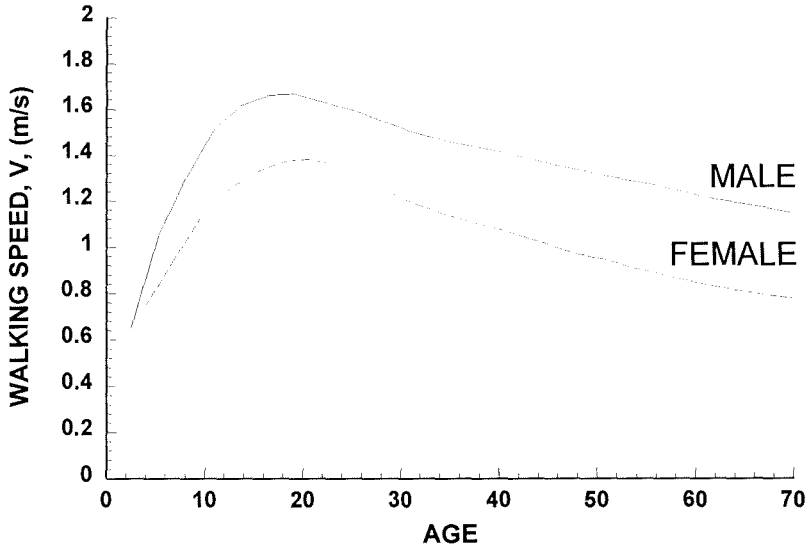


Figure 20: *Walking speed variation as a function of age, from Ando et al*

The graph shows a maximum expected pedestrian velocity of about 1.7 m/s corresponding to a 20-year-old male. From this data it would therefore seem safe to assume that all pedestrians (certainly all those that could be considered as waiting) will be moving at or below a velocity of 2 m/s (or 4.5 mph).

4.4.3 Non-pedestrian Objects

From the test video gathered, it was seen that bicycles frequently passed through the detection zone – often whilst being ridden. For the Puffin crossing there is no obligation to detect them however the system must detect any accompanying pedestrians, who should be walking.

Wheelchair users are frequent users of crossings and needed particular attention as they are more vulnerable than most pedestrians. Unlike pushchairs, a wheelchair need not be accompanied however detection is aided by the fact that their size is such that they occupy a spatial area greater than a pedestrian of about half-adult height. The fact that their shape is markedly different from that of a normal pedestrian indicated that any spatial model used in the detection algorithms would have to be sufficiently generic to include them (alternatively multiple models could be used).

The presence of litter and leaves is particularly difficult to cope with using vision Figure 21 shows the effect of leaves in the scene. It can be seen that they are as brightly reflecting (even in the overcast conditions shown) as the pedestrian (who is wearing light coloured clothing). In fact, the contrast between the leaves and the background is often greater than that for many pedestrians who wear darker coloured clothes. This would be problematic for image processing methods based on fixed pre-processing thresholds that would tend to pick out leaves in preference to the pedestrians. The density of edges (regions of high intensity gradient) present would also be likely to confuse many methods of the vision methods that attempt to find the best match between a projected shape models and edges in the image data.

A further problem with leaves is that when dry they often move intermittently with the wind. In this progression of frames, a number of leaves are being blown through from left to right by a small whirlwind creating a background of moving, high-contrast edge information. This would cause difficulties for the many vision methods based on extracting objects of interest from the background by their motion,

particularly as their temporal behaviour can appear pedestrian-like often consisting of short stop/start movements in a similar velocity range. The presence of moving leaves is likely to cause similar problems for vision methods that assume all moving edges represent pedestrian activity.

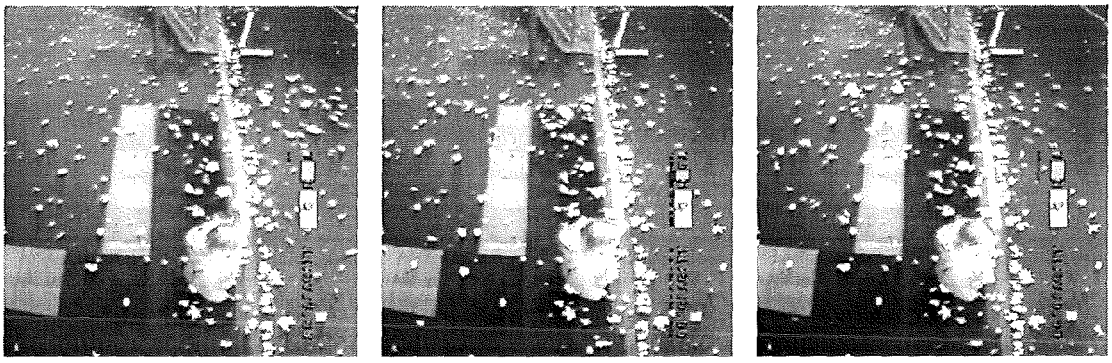


Figure 21: *The effect of leaves in the scene. Source BC5_1_4317-21.*

Although no examples are shown, the effect of litter is similar to that of leaves as it usually consists of small pieces of highly reflective paper, which are also liable to move with the wind.

4.5 Environmental Factors

4.5.1 Vibration

Typically, mast-mounted cameras for traffic monitoring sit at heights of nine or more metres above the ground. At such heights, some movement of the camera caused by wind buffeting is almost inevitable due to a lack of rigidity of the mast. Vision systems designed to analyse the video from these cameras have to take account of this motion by tracking the movement between image frames. Even with the low mounting height of the pedestrian monitoring camera it was anticipated that some camera motion due to wind buffeting might also occur, although it was expected to be of much smaller amplitude. It was also anticipated that the close proximity to the road would make it likely that vibration of the camera mount would occur due to the nearby passage of large vehicles.

In practice however, the effects of vibration were found imperceptible on trial data and the camera's mechanical structures and fixings were sufficiently rigid that the effects of vibration could be neglected. The importance of this observation is that it

could then be assumed that the camera was stationary with respect to the scene so permitting the use of background modelling methods described in Chapter 6.

4.5.2 Weather

The observed effects of weather conditions on the image formation process were mainly indirect in nature typically affecting lighting conditions or influencing pedestrians to hurry or vary their choice of attire in bad weather. Observations of the most important effects of the weather as far as a vision system is concerned are presented below.

Wind High winds in combination with cloud cover on sunny days can lead to rapid global illumination changes. It also increases the amounts of movement of litter. When winds are particularly strong buffeting might be expected to lead to vibration of the camera mount however this was not found to be a problem in practice.

Cloud When cloud conditions are overcast, solar illumination is diffused leading to an absence of shadows. Decreased cloud density can lead to the appearance of soft shadows when sunlight is only partially diffused. Frequently clouds are moving and thus modulate the light from the sun causing unpredictable sudden changes between direct and diffuse illumination.

Rain Rainfall itself was not directly visible in the image whilst actually falling. This was due to its small drop size and high velocity of movement compared to the camera's exposure time. The most important effect of rain on image formation was the resulting change in reflectivity of wet surfaces. In combination with point light sources this led to specular reflections - as is illustrated in the image of Figure 26 in the section on reflection below. Further indirect effects of rainfall on pedestrian behaviour were the increased wearing of coats and hats and the consequent changes in their shape and appearance.

Snow This condition was not observed in the trial data. However it is anticipated that the brightness from reflected sunlight may cause the sensor's exposure control apparatus problems. As snow falls more slowly than rain, it also may be expected to appear in the image.

Fog This condition was not observed in the trial data. It is considered unlikely that there would be a significant attenuation in viewing clarity over the operating distances used although there could be some loss of contrast.

4.6 Visual Artefacts

An object may appear in many different ways when imaged according to the prevailing environmental factors. Marr (1982), identified the four major factors that contribute to the appearance of an object in an image as: geometry, reflectance, illumination and viewpoint. For this task, all four of these factors are liable to vary. The effects of geometry and viewpoint variation were covered in the preceding sections. In the following sections, the influence of illumination (in terms of shadows and highlights) and reflectance will be covered.

4.6.1 Shadows

One of the major difficulties in the analysis of outdoor scenes by vision systems is discrimination between shadows and real objects. Shadows are formed by objects obstructing the illumination of all or part of a scene from a light source. They therefore have shape and temporal characteristics related to those of both the light source(s) and the object(s) casting them.

To better understand the problems presented by shadows test video was observed to determine the major classes of shadows that occurred. The results are shown in Table 8 in terms of the object/illumination combinations that caused them.

Category Label	Object	Light Source
Pedestrian	Pedestrian or Signal Pole	Sun
Vehicle	Vehicles	Sun
Global	Clouds	Sun
Texture	Pavement Surface	Sun
Building	Fixed Structures & Buildings	Sun

Table 8: *Categories of shadow*

Several sources of illumination were able to influence the scene and throw shadows (e.g. vehicle headlights, street lighting, shop lighting and moonlight). However, the table indicates that the only shadow types considered likely to lead to difficulties for the vision system were solar in origin. Headlights did cast visible pedestrian shadows onto the waiting area but being close to the scene and having relatively distributed light sources, their boundaries were softer. This is less of a problem for vision

analysis methods, which are typically edge based, than the high contrast solar shadows discussed below.

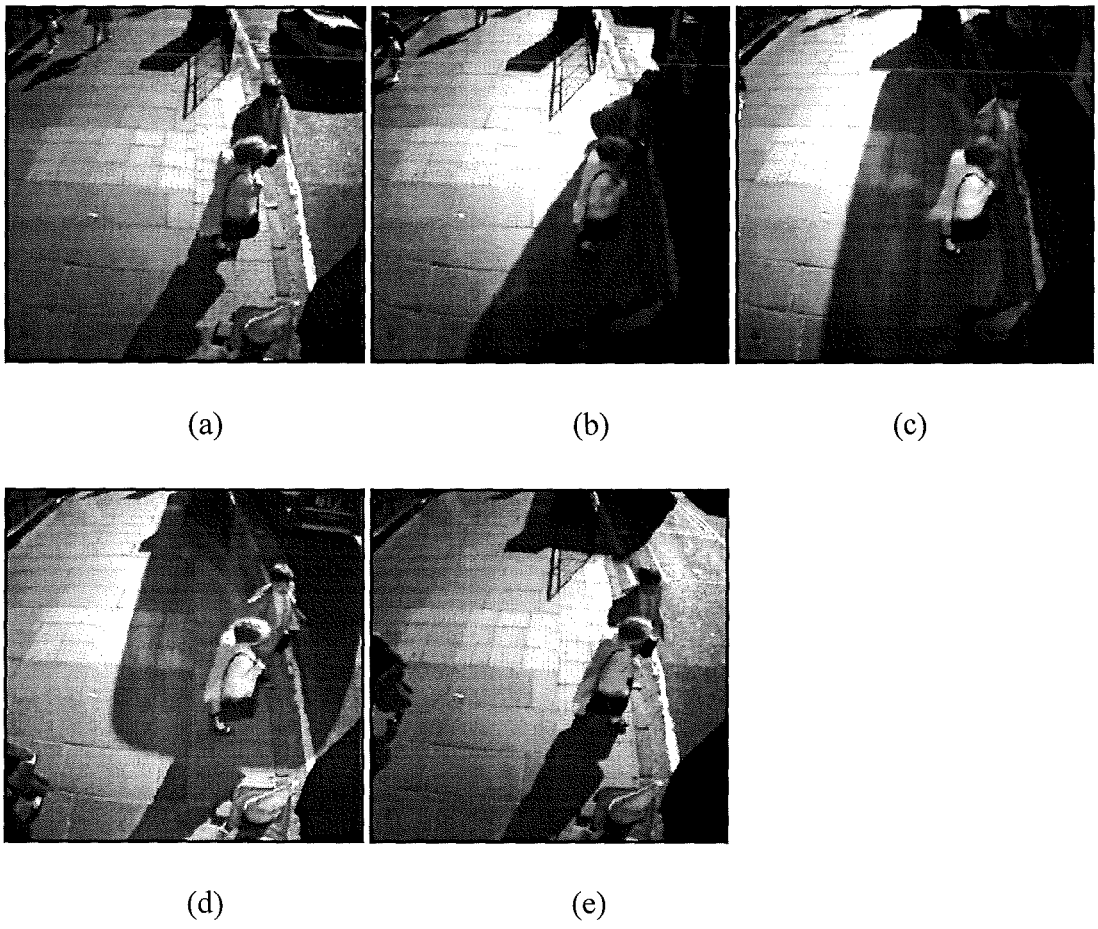
Pedestrian shadows were characterised by a high edge density with a long perimeter for their area. Their temporal characteristics are closely related to those of pedestrians themselves and so would be likely to cause problems for motion analysis methods. An example is shown in Figure 22, which also demonstrates how signal pole shadows can have a comparable orientation and shape to a pedestrian.

Figure 22: *Pedestrian and Signal Shadow. This figure shows how a pedestrian's shadow can be of approximately the same shape, size and contrast as an actual pedestrian. The shadow of the signal pole also exhibits features of similar structure to the pedestrian shadow. Source BC5_1_8273.*

As an example of vehicle shadows, Figure 23 shows the effect of a bus' shadow at the Princes Street test site. Points of interest are the induced change in exposure level shown by the brightening of the pavement in frames (b) and (c) and the large difference in contrast of pavement features according to whether they are in shadow. The strong edge features around the shadow's perimeter can be seen to move at high speed in comparison to pedestrians. Nevertheless, the aperture effect means the shadow perimeter's principal velocity component, which is parallel to the kerbside, is not necessarily apparent in the image along edges that are also parallel to the kerbside. The most important distinctions between vehicles shadows and pedestrian ones are their higher velocity and lower edge density (gradient signal per unit area).

The effect of global shadows caused by passing cloud cover is shown in Figure 24. The shadow strength due to pedestrians and signal pole can be seen to increase significantly within the elapsed time of only 4 seconds between the images. This is

troublesome for a vision system as much pedestrian activity occurs over similar time scales creating problems for discrimination based on the time of persistence of a pattern in the image (e.g. background modelling methods). A further related shadow problem is that texture profiles in the surface of the pavement are often picked out by bright sunlight, producing a strong contrast between highlights and shadows. The amplitude of this additional 'noise' can be very high and is likely to be difficult for texture analysis methods to tolerate. The high local intensity gradients produced would be likely to cause difficulties for vision methods using the matching of edges to wire-frame models.



(a) (b) (c)

(d) (e)

Figure 23: *Vehicle Shadows (Source WEPS4_1_6599-6607)*

Shadows due to buildings vary sufficiently slowly so as not to trouble motion analysis or background modelling methods. Their main negative effect is linked to the global nature of the exposure control mechanism in the image sensor - as was described in Chapter 3. Figure 25 shows the effect on image formation of the sweep of a building's shadow through a detection area. Of particular interest is the effect of the exposure on the detail visible in the sunlit and shadowed portions in the scene. In frame (c) the

centre portions of the pavement texture are saturated to white (over exposure) whilst in (d) the shadowed left-hand side is underexposed - also removing detail. This contrast variation would adversely affect much previous work that has been based on background modelling and comparison techniques. Note that the time elapsed between each of the frames in the figure is 12 minutes and so has a bearing on the choice of adaptive time constants.

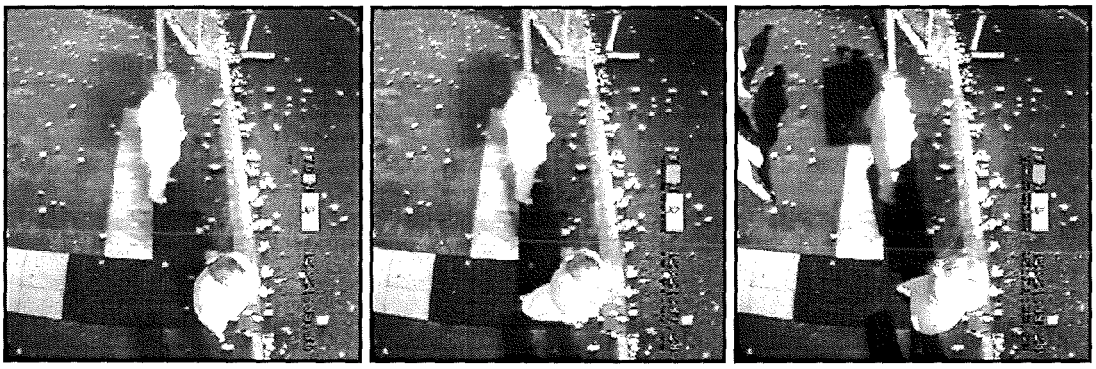


Figure 24*: Global shadows due to passing cloud cover. Source BC5_1_7375-83.*

In general shadows can be considered as consisting of two components, which are distinct in their effect on the vision system, namely their body and boundary.

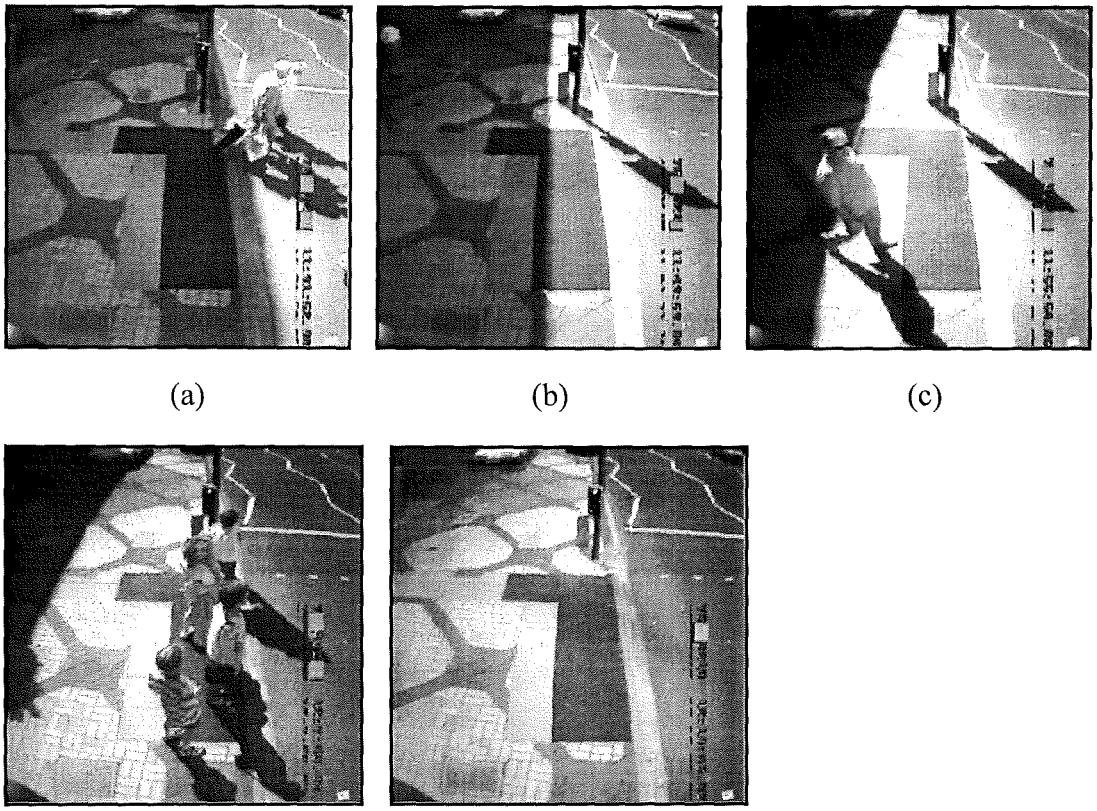


Figure 25 *Building Shadows. Source BC4_1_0001,1501,3001,4501,6001.*

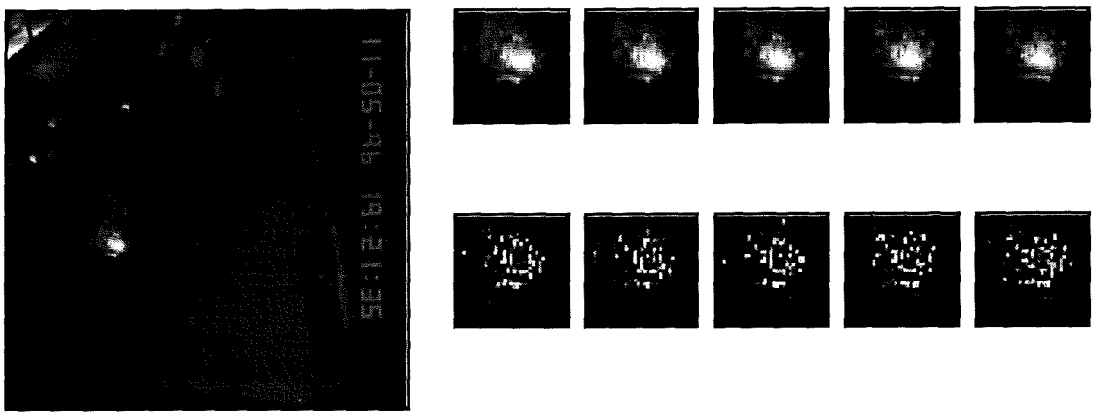
Within the body of the shadow, there is uniform darkening of the underlying scene however the pattern of the scene elements should still be discernible under most conditions. Algorithms seeking to match these patterns under varying shadow conditions must therefore incorporate some invariance to this variation in contrast.

The boundary of the shadow produces a strong gradient feature that obscures any underlying scene information and so cannot be handled in the same way. Most vision systems at some point rely on an edge-based description of the scene often matching edges to object models to identify them. Finding a way to deal with these boundaries was an important part of this research.

4.6.2 Reflections

In many computer vision systems surfaces in a scene can be considered to be Lambertian (Schalkoff, 1989) i.e. to diffuse light uniformly in all directions. However under some circumstances objects can be highly reflective and cause distracting artefacts in images. Typically, these include metal poles (which are part of the crossing infrastructure), objects carried by pedestrians and wet surfaces after rain. The highly reflective surfaces can lead to these specular reflections when the incident rays from a light source are coplanar with the reflecting surface's normal and with the reflected rays. They are particularly troublesome for image sensors and hence for the vision system as they lead to spatially localised bright spots that may cause the sensor to have difficulty in choosing a good value for its exposure time, Figure 26.

The direct effects of heavy rain on image formation are shown in Figure 26. In (a) six areas of specular reflection can be seen where light sources (from shop window displays on the left of the main image) are reflected in rain water on the wet pavement surface. The sequence (b) shows how the falling rain affects the image texture of the largest of these areas whilst (c) shows the spatial gradients in (b). There is no discernible effect on images due to rain in the air, as its time of passage through the projected area of an image pixel is very small compared to the camera's exposure time.



(a) (b, top) (c, bottom)

Figure 26 *Rain: Source WEPS6_1_999,3,5,7,91.*

In practice conditions between the above extremes can occur leading to a 'sheen' from the pavement surface which is of lower intensity than a specular reflection but which covers a larger area.

4.6.3 Highlights

Two sources of highlighting were observed. At night, car headlights caused bands of light and shadow to rapidly traverse the waiting area of the crossing. As was mentioned above these effects are not as troublesome to the vision system as daytime shadows. This is because headlights, being close to the scene, are only approximately a point source so the boundaries between shadows and highlights are somewhat softer in intensity gradient around their boundaries than those due to sunlight.

The highlighting effect of headlights was most problematic when it reflected off the clothes of pedestrians at night artificially producing a sudden large increase in their contrast in the image. This effect was found to cause difficulty in setting detection sensitivity thresholds for the vision algorithms. A pedestrian who was not illuminated resulted in a comparatively weak signal such that they may have been missed if adjustment of levels had been based on these more strongly contrasted pedestrians (see Chapter 6).

4.7 Transportation Factors

4.7.1 Pedestrians

This section looks at the characteristics of pedestrian group behaviour. As would be expected, occlusion by pedestrians of each other increases with flow. However, even at low flows occlusion was seen to be significant, as pedestrians tended to bunch together by waiting in a string along the kerbside or in the groups in which they arrived. The occlusion problem was further exacerbated by the effects of the intersecting flows of crossing, arriving and passing-by pedestrians.

It was also of interest to consider the maximum level of pedestrian volume to be expected. A Department of Transport trial of sensors for the Volumetric Puffin chose an upper level of occupation at 18 pedestrians for an average sized crossing. In practice the maximum occupation levels for the five test sequences used in this work (described in Chapter 7) were 11, 12, 18, 12 and 2 pedestrians. An example of the distribution of pedestrian volumes is given in the figure below.

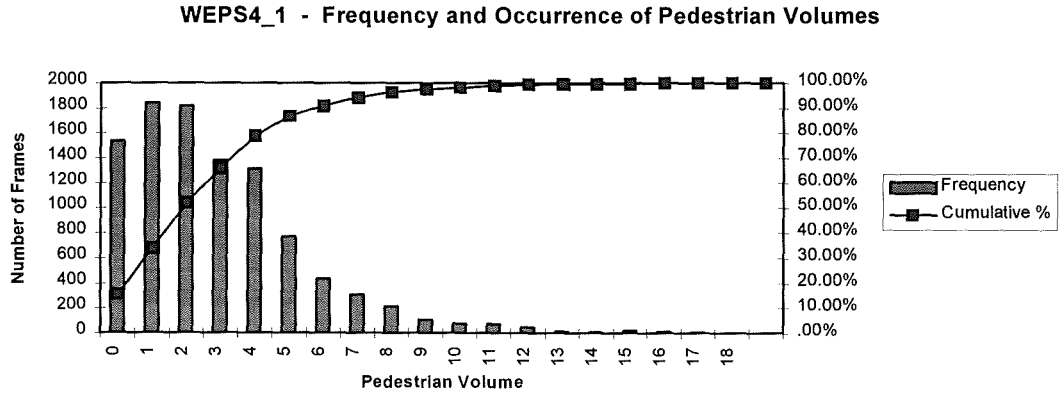


Figure 27: *Distributions of number (volume) of pedestrians occupying the waiting area over a 10,000 frame evaluation video sequence.*

The importance of knowing the maximum expected volume was that, in the detection algorithm (described in Chapter 6), it affected the response time in volumetric mode due to the way pedestrians were extracted sequentially. Knowledge of this figure meant the maximum response time could be limited by capping detection once the maximum occupation level had been attained. In operation, an engineer installing the system could select a maximum value appropriate for the area to be monitored and the level of use.

4.7.2 Vehicles

As was discussed earlier the effect of passing vehicles on image formation occurs through the projection of their shadows onto the pavement area. Although traffic flows were not monitored for this work, a previous study at a crossing in Edinburgh found flows of up to 1300 vehicles per hour on a single carriageway, city centre road. Clearly this figure would be higher for multilane roads, but given the interaction will mainly be with vehicle shadows from the near lane, the above figure is a useful guideline.

During the period when pedestrians are waiting, vehicles will have a green light and are so are likely to be moving past at a comparatively high speed. The expected range of velocities for vehicle movement will start from zero, for queuing traffic, up to 40 mph, although the observed velocity of their shadows projected onto the pavement the may often be higher than this. Given the above it seems reasonable to assume that, much of the time, vehicle shadows will move quickly through the image in comparison to pedestrian movement.

4.8 Conclusion

This chapter has identified the most important effects of the environmental and transportation constraints under which the vision task had to be performed. The important conditions and objects likely to be encountered were identified and parameters of importance to a vision system estimated. This information forms the basis from which an assessment of the prior art can be made in the next chapter and from which algorithm design (described in Chapter 6) can proceed.

Examination of scene structure indicated that there was significant variation in pedestrian images due to viewpoint and perspective. This suggested that at least a basic calibration phase would be necessary to establish the relationship between objects detected in the 2D-image plane and their likely position in the 3D world. Additionally this was likely to preclude the use of many feature-matching techniques as not only would features vary significantly in scale and orientation over the scene but they would also have a high probability of being occluded.

Perhaps the main problems identified for a vision system were those due to the possible presence of small high contrast objects such as litter and leaves and those due to the strong shadows produced by direct sunlight. An important conclusion arising

from examination of pedestrian variation and behaviour was that any structural model used to represent a pedestrian would probably have to be quite broadly defined. If this were not the case then the system would have difficulty coping with the variation in basic pedestrian shape let alone incorporating objects such as wheelchairs and dealing with variation in pedestrian attire. This contrasts with models used in previous work on human body detection which often sought to model at the level of individual limbs. These matters are considered further in the review of prior work in the next chapter.

A positive point was that as the camera could be considered stationary with respect to the scene. Background modelling techniques could therefore be examined as a means of separating pedestrian activity from other, more permanent, image content such as building shadows. It was also observed however that rapid changes due to surface texture illumination and camera exposure control responses meant that some components of the background structure would be spatially and temporally complex. Consequently many prior art background techniques, that had often been based on assumptions of relatively simple illumination conditions, were unlikely to be applicable without modification.

This chapter has identified the most important factors in the application of vision to detecting pedestrians waiting at the kerbside of a pedestrian crossing. The next chapter categorises and reviews relevant work from the academic literature in computer vision to identify techniques that are applicable to this task.

5 Review of Computer Vision Applied to Pedestrian Sensing

5.1 Introduction

This chapter examines systems from the literature in computer vision that have been applied to pedestrian and human body detection. To set the scene a broad overview of the range of motivations that have led to work in this area is given and from this the application areas most relevant to this work are identified.

This leads on to the main discussion that categorises the prior work according to the task-specific assumptions upon which they were based. The validity of these assumptions is examined with respect to the objectives and particular difficulties relevant to this research, as identified in earlier chapters. From the examination of previous work on detection algorithms it is found that a widespread approach has been the use of background modelling, to diminish distracting, permanent features of the scene, followed by the application of spatial models to extract pedestrian positions. Due to the likely importance of these methods previous activity is reviewed in more detail in sections 5.3.1 and 5.3.2.

The application-orientated nature of this work meant it was also considered useful to examine the wider literature in computer vision for techniques that were expected to be important in the completion of this work. Accordingly there are sections reviewing alternative sensor technologies, vision techniques for minimising computational load, and performance evaluation methods.

5.2 Prior Art Categorised by Application

In the computer vision field there has been a considerable amount of work directed in various ways at sensing the human body. A first distinction that can be made between

the various bodies of work in this area is the scale at which people have been observed. At the smallest (external) scale is the area of biometrics. Here the objective is to match measurements of various body parts to stored representations as a means of recognising or verifying a person's identity. Biometric patterns that have been researched include the shape of faces, palms and fingerprints and vein patterns in the wrist and iris. These tasks require the very specific analysis of particular body parts at a level of detail beyond that which is appropriate to the objectives of this work. Although some of the pattern matching techniques used in these applications may be applicable, their operation is generally assisted by controlled illumination to ensure a clear high-contrast image is obtained for analysis. At the top end of the scale, crowd monitoring techniques have tried to assess the characteristics of dense groupings of people in terms of the resulting image texture or the periodicities in image change resulting from head movement.

Between these extremes there are a number of application areas involving observation at the scale of individual people. Here the motivations have included:

- Advertising: Interest has been shown in counting the number of pedestrians within user-defined areas of shops and exhibitions as a means of gauging the level of interest in a display (Tsuchikawa 1995).
- Architecture: Pedestrian monitoring for both building management and planning has been undertaken (Waterfall 1981, Sexton 1993, Sexton 1994, Zhang 1995).
- Human posture and motion study: Vision based extraction of body pose has been used in the analysis of movement for athletics (Lerasle 1996, Akita 1984), dance (Kakadiaris 1995, Campbell 1995), as well as for rehabilitation and disability studies (Chelette 1996, Guo 1994, Long 1991, Leung 1995, Meyer 1997).
- Multimedia systems: Work has been published on a vision-based human computer interface (Wren 1995, Wren 1997) and on its use for allowing user interaction with virtual objects (Azerbayejani 1996):
- Robotics: Pedestrian detection has been used in this area as a means of robotic control (Chen 1992), for the automatic recognition of command gestures (Okawa 1991) and for navigation (Mori 1994, Yasutomi 1996).

- Security: Examples of vision applied to security monitoring include methods for tracking pedestrian activity (Baumberg 1994, Tesei 1995) and means of identifying certain classes of movement (Polana 1994).
- Transportation: There have been numerous attempts to apply vision in transportation systems including pedestrian counting methods (Bartolini 1994, Hwang 1983, Khoudour 1996, Rossi 1994, Sato 1993, Tang 1991, Xidong 1993), pedestrian volume measurements (Lu 1990, Regazzoni 1996), tracking across roads (Rohr 1994, Kinzel 1994, Richards 1995), estimating demand at lifts (Schofield 1994, Schofield 1996), the detection of overcrowding (Regazzoni 1996) and driver assistance (Heisele 1997).

Of the above application areas, the transportation domain must rate as one of the most demanding as it requires operation under unconstrained outdoor conditions combined with strict requirements on performance, response time and reliability. In compensation however the level of detail of pedestrian activity required, for this research at least, is relatively low.

Looking at the objectives of previous work in terms of the level of detail of information extraction required, they can be categorised in order of increasing difficulty, through:

- The detection of pedestrian presence (e.g. Surveillance and Robotics safety)
- The detection of pedestrian volume (e.g. Advertising)
- The tracking of individual pedestrian motion (e.g. Architecture, Surveillance)
- The determination of body pose (e.g. Posture study, Robotics Training and Gesture control)
- The identification of a class of activity (e.g. Robotics Training and Gesture Recognition)

As might be expected, it is apparent from examining this work that the higher the level of information being sought, the greater were the simplifying assumptions made by the authors. Indeed much work on posture analysis and activity identification started from the premise that information on the position of body joints had already been extracted manually or by unspecified 'low-level' processes. Even within previous comparable work addressing the first two categories (with which this work is

concerned) it was still found possible to identify important assumptions that would affect their applicability to this task.

The next section examines the nature of the simplifying assumptions used in previous work which, either explicitly of implicitly, formed the basis of the algorithms developed by their authors. A central issue in assessing the relevance of the techniques used was therefore an examination of the validity of these assumptions against the particular difficulties of this work - as detailed in the previous chapters.

5.3 Prior Art Categorised by Assumptions

To make their tasks tractable previous authors have, quite reasonably, utilised task-specific assumptions to simplify the demands on the vision algorithms. In the discussion below, prior work is assessed with respect to the assumptions used and their validity as regards the current task. In consequence, methods of value to this work are identified. It should be noted that it is common for several of the assumptions discussed below to be made use of within a single piece of work.

In the case of a free moving camera operating in an environment occupied by both important and distracting objects, objects of interest can in principle be extracted by the direct matching of spatial models to the image data. However achieving this is difficult in practice, as algorithms must be tolerant of distractions to the matching process due to the unwanted components of the scene. A slight simplification is to assume that the majority of the scene occupies a single frame of reference (relative to the camera) so that attempts can be made to differentiate between it and objects of interest by extracting the strongest global motion field. An example would be the analysis of images from a camera mounted in a moving vehicle. Again this is difficult due to the computational demands of determining the motion field and then identifying background objects which, despite their common world velocity, have image velocities that vary according to their position in the image.

However for the vast majority of vision work, and for practically all the work reviewed, the further assumption was made that the camera was stationary with respect to the majority of the scene whilst objects of interest were in motion, relative to this static 'background' frame of reference, for at least some of the time.

Many previous workers have based the identification of pedestrian activity on the detection of this relative movement. Within this category of application some work

has been aimed at the direct recognition of pedestrians from image related motion signatures associated with regular rhythm of walking (Mori 1994, Yasutomi 1996). This is obviously only applicable to tasks where the assumption of regular and observable pedestrian motion can be made. Others have made the assumption of continuous pedestrian motion, without requiring the rhythmic pattern (Rourke 1992, Polana 1994, Leung 1987, Cai 1995, Gu 1996, Sato 1993, Hwang 1983, Lu 1992, Rossi 1994, Rohr 1994, Sexton 1993, Sexton 1994, Sullivan 1995, Tsuchikawa 1995, Tsukuyama 1984, Vannoorenberghe 1997, Yasutomi 1996). This allows techniques such as inter-frame differencing to be used to identify pedestrian activity and in some cases is supported by the additional assumption that the only moving objects were pedestrians (Rourke 1992, Sexton 1993, Sexton 1994, Lu 1990, Hwang 1983) to simplify the task still further. This latter assumption is clearly not valid when time-varying shadows are present. An additional spatial constraint on motion that has also been appropriate to several prior pieces of work is that pedestrian movement was limited to a single direction of movement and/or constrained to cross a detection line (Tsuchikawa 1995, Bartolini 1994, Rossi 1994, Sato 1993, Sexton 1993, Sexton 1994, Tsukuyama 1984). This was used to localise analysis and so reduce computational load.

The use of motion analysis under such constraints is efficient at separating-out pedestrian activity at low computational load. It can also be effective in the presence of shadows so long as their motion/variation is sufficiently slow as to be negligible within the time period over which the motion is estimated. This amounts to a need for image changes due to environmental changes to happen more slowly than those due to pedestrians do. Unfortunately however pedestrians at crossings could not be guaranteed to be in continuous movement and shadow patterns would frequently vary at least as quickly as pedestrians would.

In cases where continuous pedestrian motion could not be assumed previous researchers have looked at the optical flow between image pairs (Richards 1995, Velastin 1994) as a means of identifying pedestrian activity or directly attempted to track grey-level features (e.g. Allsop 1997). There must also be some doubt as to whether these methods could successfully distinguish between the motion behaviour of a pedestrian and that of their (more strongly contrasted) shadow. The presence of litter and leaves in the scene is also likely to be problematic for these methods as they

would provide multiple high-contrast, mobile points to distract the algorithm and interfere with the search for feature correspondences.

Motion analysis is just one example of the approaches that have been applied to the separation of pedestrian activity from background scene. Perhaps the simplest cases are those where assumptions could be made about the complexity of the background scene or even where assistive alterations could be made to it. Under such controlled conditions it was possible to arrange for the background to be of uniform intensity to simplify object segmentation (Okawa 1991, Akita 1984,Tsukuyama 1984) or even to add alterations to the scene such as white lines (Bartolini 1994, Lu 1990), checkerboard patterns (Glachet 1995) and retro-reflective stripes (Khoudour 1996). Unfortunately such simplification of alteration of the background is not an option for this work.

Assumptions related to the illumination of the scene have also been popular as a means of easing the separation of pedestrians from the background. Most common was the assumption that there were no shadows present or that they were sufficiently diffuse so as not to cause a problem (Tsukuyama 1984, Cai 1995, Akita 1984, Papanikolopoulos 1995). Some groups overcame the problem of shadows by arranging the camera to have a horizontal viewpoint. (Rohr 1994, Cheng 1997, Meyer 1997, Akita 1984, Gavrila 1996, Mori 1994, Richards 1995) to remove shadows on the ground plane from consideration - as they were no longer visible in the image.

The examples given in the last chapter illustrated that assumptions simplifying the lighting and background scene as invoked in the above work could not be made at pedestrian crossing sites. In addition the camera was constrained to be mounted on the signal pole so that the advantages of side-on or plan viewpoints could not be exploited.

For a task of the complexity of that addressed by this work a more realistic approach was to accumulate a model of the background based on observation of its long-term characteristics. A significant volume of work has addressed such background modelling methods and so they are reviewed in a separate section below which also looks at how they have been made use of to distinguish pedestrian from background signal. Thereafter a review of model based approaches to turning this pre-processed data into information on individual pedestrian activity in section 5.3.2.

5.3.1 Prior Art in Object/ Background Separation

A recurring theme in computer vision is that of finding pre-processing methods for the separation of data of interest from a more permanent background scene. In using a background modelling approach there are two aspects to be considered. Firstly there is the means of forming the background model and secondly there is the means of matching it against the current image from the frame-grabber so as to separate out the objects of interest for higher level analysis. Approaches to these tasks that have been identified from the prior art are discussed below.

Background modelling methods that were reviewed are summarised in the table below. The simplest method of obtaining a background model that has been used is to require a manual operator to judge when the scene is free of activity and to trigger the capture of an image that can be used from then on as a background reference. Velastin (1993) has for example made use of this method. This approach, in addition to its undesirable reliance on manual operator contributions, is limited to very simple environmental conditions where unpredictable changes in long-term scene content do not occur. Clearly any changes between the capture of the background and the capture of a live image will be reflected in comparisons.

Others have made use of non-linear methods to filter out short-term events. For example Baumberg (1994) used a temporal median filter. This would however have required too much storage for the time scales over which this pedestrian detector had to operate. It implicitly assumed that scene was usually empty and that objects of interest were moving quickly with respect to the temporal span of the filter. A scheme proposed by Leung (1987) required the occurrence of 'n' identical samples to cause a pixel update where the value of 'n' required was proportional to the distance (difference) of the new candidate value from the original. The reliance of this method on there being extended periods during which pixel intensity was essentially constant (or within a small range) is a weakness as this degree of constancy is unlikely to occur over extended periods under realistic outdoor lighting conditions and levels of pedestrian activity. Indeed Hwang (1983) gives examples of how the time required to acquire backgrounds, even when even quite short periods of constancy are required for update, can take a very long time to occur. A further alternative is based on the mode of a pixel's intensity over a time period (Okawa 1991) which again makes implicit assumptions about the scene that are not valid for outdoor vision applications.

Fathy et al (1995) introduced a combined method of background updating using edges for illumination immunity. New background values were based on a threshold chosen from the histogram of inter-frame differences. They also averaged any new value with the current background value to smooth-out transitions. It should be pointed out however that it is only edge direction and not magnitude that exhibits the sought illumination immunity.

The above examples illustrate how the response of many groups to performance difficulties with these methods has been to make the background generation process progressively more complicated by the addition of more and more rules to the modelling process in response to empirically identified problems. Underlying models of the basis for the expected separation are usually absent and furthermore the behaviour of the rule based systems becomes harder to understand intuitively and their responses correspondingly more difficult to predict. Particularly hard to predict are methods of the selective updating type (e.g. Sato 1993, Vannoorenberghe 1997, Lu 1992). These make decisions at the current frame which are dependent on a decision from the last iteration, such that the effects of an error could propagate without limit.

Other workers have sought to track slow variations in the background scene by the pixel-wise application of temporal low-pass filters such that the effect of temporary objects is averaged out. The use of a Finite Impulse Response filter applied over the last 'n' samples of the image sequence causes practical problems of storage and calculation requirements, which are prohibitive for practical time constants (see Seed 1988). In response workers such as Seed (1988) and Papanikolopoulos (1995) formed a background from a pixel-by-pixel Infinite Impulse Response (IIR) temporal filter. This allowed a long-term average to be generated quickly using less calculation and storage but weighted the most recent images with greater significance. Seed et al also identified difficulties due to intensity quantisation in applying this filter using the 8 bit integer data representations to which they were restricted. Richards (1995) used a variation of the IIR filter method where the time constant parameter of an IIR filter was dynamically varied according to an 'importance factor'. Unfortunately generation of the 'importance factor' was not described.

Method	References	Description
Manual Capture	Velastin 1993 Yin 1996 Xidong 1993	A human operator triggers the capture of a background image when the scene is empty. The reference may be periodically updated by the operator to track changing conditions.
Majority Vote	Okawa 1991	The most common grey-value from a set of sample images is chosen. The implicit assumption is that the background is visible in >50% of sample images. The effects of noise and large storage requirements are likely to be problematic.
Most stable intensity	Long 1991	The background becomes the most stable grey-level over a given sequence period.
Block stable	Sexton 1993 Sexton 1994 Sexton 1995	The image divided into non-overlapping blocks. If a block is steady (undefined) for n cycles then it becomes background. This may take a long time to occur for outdoor scenes.
Random Updating	Seed 1988	Background pixels are randomly replaced with current ones. This leads to impulse type errors but they are spatially unconnected so easily removed by binary/median processing. The replacement rate determines the update time constant
Update on no motion	Lu 1992	Background pixels are replaced when no motion is detected as judged by calculating the difference from the last background.
Temporal Median	Baumberg 1994	The background is the median value of the time series for each image pixel. This method would need significant storage to cover a prolonged time period.
IIR Average	Papanikolopoulos 1995 Michalopoulos 1991 Richards 1995 Seed 1988	The background is a recursive average of the incoming signal with long time-constant.

Slope limited updating	Seed 1988	A limit is imposed on the rate of change of background update.
Non-linear recursive edge map	Vannoorenberghe 1997	The background becomes a combination of the edge strength of the last background value and that of current image. The proportions of the mix are controlled by a factor that depends upon squared difference between the reference and current edges of the last iteration.
Statistical Measures	Sato 1993 Tsuchikawa 1995	The background is characterised by its mean and variance over a fixed length time block. Pedestrian motion and good pedestrian contrast are required. The decision to include a new frame as part of the statistical data for background model is based on a decision made using the existing background model.
Selective updating	Seed 1988	Only those areas identified as background in the last iteration are used as the basis for updating the background in the current cycle.
Neural Network	Schofield 1994 Schofield 1996	Based on the recording of multiple (binary) patterns. It assumes a training set of background images is available. A sobel edge image is binarised against the average background edge image. Each ANN processor is a content addressable RAM, which stores a '1' for all patterns from the training set. The proportion of '1's produced with actual data indicates similarity to the training set. It is effectively a means of storing and comparing to a set of alternative local background patterns.

Table 9: *Background Modelling Methods*

The approach taken in this work, was to be realistic about what could expected of a background removal process and accept that the overlap in the motion and contrast characteristics of pedestrians and shadows meant no perfect separation would be achievable. Accordingly, as will be discussed further in Chapter 6, a simple temporal averaging based method was preferred, as its behaviour and limitations were easily understood and yet it could be expected to be useful in reducing the impact of the background scene on later processing. In using this method, it was accepted that the

tracking and automatic compensation for rapid illumination effects, such as cloud shadow, by the background model would not be possible.

Once it was accepted that the background model would never be identical to the true background, the importance of finding a matching operator that would usefully compare background and live images, and yet be as invariant as possible to illumination differences, became apparent. A matching operator was therefore sought which would provide a means of identifying changes of pattern in the image data whilst removing the effects of differences due to illumination and shadow bodies (as opposed to boundaries) from the result of the matching process.

Matching methods that have been used in the prior art are summarised in Table 10. By far the most common method of comparison used was pixel-by-pixel subtraction, resulting in a difference image. The absolute values of the difference image were then usually binarised by comparison against a fixed, empirically chosen, threshold value. Although this approach can be effective under controlled and constant (e.g. indoor) lighting conditions it is not useful when even simple global changes in illumination occur.

Sexton (1994) extended this by comparing the first order differential (sobel magnitude operator) of the current and background images rather than the raw intensity levels. This method was investigated by the author but determined only to be of value when used on uncluttered background scenes with little edge information. If this was not the case background regions of high gradient coinciding with those in the current image caused unwanted removal of image edges. This loss of signal was found to be particularly critical when the entire pedestrian projection only occupied relatively few pixels at the far edge of the detection zone - a problem exacerbated by the use of reduced resolutions to speed analysis.

Bartolini et al (1994) developed a system to analyse passenger movement through bus doorways in outdoor conditions. They identified the importance of avoiding the use of binarising thresholds to identify edges as these were found to be too difficult to establish in practical operating conditions. Operating over a thin strip window of the image across the threshold of the doorway they sought to remove dependency on edge strength (contrast) by identifying edges as maxima in the gradient image and then looking at the orientation of the extracted edges. For their application, the fact that

operation was over a narrow window within which the step edge was being sought as a reference point meant that they found it sufficient to classify edges into one of four directions. This approach was influential in the design of the matching operator used in this work as is described in the next chapter.

Another common approach to signal matching has been to measure correlation over a local neighbourhood of pixels. The main reason standard correlation was unsuitable for this task was its dependence on the level of the two signals to be matched, whereas ideally the match operator needed to be invariant to differences in signal level and scale caused by illumination changes. In addition the operators' multiplicative nature emphasised high contrast signals with respect to those of low contrast and the large number of multiplications required represented a heavy computational load. Normalised correlation methods could have been used to address some of these issues but would have introduced significant amounts of extra computation and required floating-point operations. Most of these shortcomings are overcome by the use of the 'sum of absolute difference' operator which is fast to compute and so has achieved great popularity in vision and image processing work. Its use for this project is described in Chapter 6.

Method	References	Notes
Subtraction	Yin 1996 Lu 1990	Absolute difference measured
Edge strength	Sexton 1994	Subtract and threshold
	Vannoorenberghe 1997	Subtract, square, normalise and threshold
Edge orientation	Bartolini 1994, Regazzoni 1996	Based on Sobel operator
Local uniformity	Sexton 1993	Threshold identifies pixels that differ from a reference by more than the local average difference of its neighbours.
Median Distance	Rossi 1994	Better immunity to impulse type noise than mean square error methods
Sign correlation	Nishihara 1994	The sign of the image's laplacian (second order difference) is used to characterise it. As a binary pattern

Table 10: *Matching Methods*

The importance of selecting background separation methods appropriate to a task is illustrated by work in outdoor conditions by Richards (1995). He admits his use of subtraction followed by fixed threshold binarisation would have problems dealing with the effects of cloud shadows (other shadow effects are removed by constraining operation, in the examples given, to a side-on viewpoint). He indicates that future work will look towards a more complex pedestrian model to address this difficulty. However, even with better modelling, the pre-processing method described for separating foreground from background would be a handicap to system performance as pedestrians in regions of intensity change would be lost. The pre-processing method would also lead to large false positive regions due to global illumination variation (it would also perform poorly for local shadow effects, because of the application of fixed threshold). The important point to make here is that premature loss of information, as was caused in this case by the binarisation step, should be avoided in dealing with the range of contrasts encountered in transportation scenes. This is necessary to avoid the loss of information that might be usefully used in later processing stages in, for example, finding low contrast pedestrians in a scene with

high contrast shadows. This consideration then has to be balanced against the wish to remove as much background information as possible to simplify the spatial modelling task. This realisation was central to the design of detection algorithms that are described in the next chapter.

Although the above methods are of value for compensating against variations in illumination within large and global shadows, they are of no assistance in dealing with the intensity gradient produced at the shadow boundaries which are often of greater contrast than those due to real objects. The large number of small scale shadows due to pedestrians, litter, leaves and passing vehicles mean that these perimeter effects still represent a significant problem even when more stationary background artefacts have been filtered out using methods such as those described above. The next section continues by looking at the use of spatial modelling techniques that might be useful in extracting pedestrian positions from amongst this distracting information.

5.3.2 Prior Art in Spatial Modelling of Pedestrians

Most analysis approaches at some point involve the top down application of a model. These range from 2D clustering of pixels through to deformable 3D volumes. The type of spatial model required for a particular application was found to be to a large extent determined by the nature of the operating environment and what could be assumed of it. It was mentioned above the many workers had been able to work with the camera at a horizontal viewpoint. A related tactic was to operate from a camera placed above the pedestrians so as to obtain a plan view (Zhang 1995, Glachet 1995, Schofield 1994, Schofield 1996, Bartolini 1994, Khoudour 1996, Rossi 1994, Sexton 1993, Sexton 1994, Tang 1991). Both these methods in combination with the camera being positioned at a distance from the pedestrians that was large compared to pedestrian size enabled spatial modelling to be simplified to a 2-D situation. The use of a plan view also had the benefit of removing most of the problems of pedestrians occluding each other from the camera's view.

Some workers also made use of assumptions regarding pedestrian presence in the scene. These included that pedestrian density was low (Okawa 1991) and that exactly one pedestrian was present (Cheng 1997, Guo 1994, Wren 1995, Wren 1997). Yasutomi (1996), for example, segments pedestrians from the scene by obtaining the absolute difference between consecutive image frames. Vertical edge detection is then

applied and thresholded to remove 99% of the signal. This amounts to an implicit assumption that pedestrian density in the scene is low. From this information a pedestrian's feet are isolated and the temporal variation in intensity in this region studied.

In those cases where pedestrians were sufficiently distant for a 2-D model to be applicable many groups have chosen to fit contour models based on standard geometrical forms such as rectangles to model entire pedestrians (Regazzoni 1996, Richards 1995) and circles (Zhang 1995) to model heads. These bounding shapes are positioned so as to best fit the perimeters of clusters of binarised pixels obtained from a pre-processing stage. Assuming the pre-processing binarisation has been accurate and that only the position of pedestrians or pedestrian groups is required, this provides a quick and effective measure of object behaviour. An alternative in this category has been the use of 2-D area models that match image information using measures of centroid and area (Sexton 1994, Wren 1995, Wren 1997).

More complex 2D models have attempted to find the edges of individual parts of a pedestrian's body. Examples include the use of articulated skeleton and ribbon components to represent the limbs and joints of the body (Chen 1992, Guo 1994). A more recent trend has been to use models based on deformable contours to represent a pedestrian's outline. Locally parameterised mathematical curves (such as 'Snakes', 'B-Splines' and 'NURBS') are used to represent the shape of a model according to the positions of a small set of control points. The model is matched to the image by moving the control points and hence deforming the curve until the best match with image data is obtained. Baumberg has taken this furthest in its application to sensing of the human body. He has described a method for automatically generating models of the variation of pedestrian outline shape based on a point distribution model. The method involved analysis of variation in position of a set of key points around a pedestrian's outline as they move. Principal component analysis is then used to extract a basis set to represent the main modes of shape variation as compactly as possible.

Work on these deformable outline models to extract the modes of outline variation also implicitly assumes that variation in the projected shape is due only to the form of the underlying limbs. Extensions to this work have looked at taking better account of the fact that the body is formed from a set of articulated limbs. (Heap 1996.) by

extending the point distribution model to polar co-ordinates specified relative to the joints of rigid components. No account is yet taken of bulky attire. For application to this research the main drawback of this work is that it is currently two dimensional in its modelling and that it operates from a distant side-on viewpoint under weak perspective, in published examples – such that 3D effects are minimised. Although the point distribution method might be extended by training the system to extract the modes of training samples covering from varying viewpoints as well as pose, it is unlikely that it would cope well with viewpoint changes. Furthermore for the present task the problem of variation in, and independent movement of, pedestrian attire would be likely to mean the models were not valid for a sufficiently wide range of conditions. An additional difficulty with all contour type models is that they have problems finding a good match to image data in the presence of complex cluttered background scenes (Wren 1997)

Three-dimensional models have been based on articulated cylinders (Rohr 1994) and applied to a pedestrian crossing a road. The results presented appear promising but are based on a very short data set under unchallenging lighting conditions. It is also considered that it is unlikely that this model would adapt well to dealing with the effects of clothing. More complex 3D models based on superquadric functions have been used but have tended to find application in off-line medical image analysis situations where precise details on limb shape are important and there are no constraints on processing power.

A more recent trend, based on physiological investigations of biological vision systems, has been towards the view-based representation of visual objects. This approach lies somewhere between the 2D and 3D approaches described above. Rather than a full 3D model, an object is represented by a set of views from different orientations with interpolation between them being used to model intermediate views. Pedestrian detection using this approach has been explored using both temporal (Davis 1997) and wavelet templates (Oren 1997). The method is essentially an analysis of the variation of intensity patterns in a training set of 2D pedestrian views taken at a relatively large distance from the camera. As such it is not yet suited to dealing with the range of shape, scale and viewpoint variation required for this task.

On the whole the problem of pedestrian shape variability due to attire (see Chapter 4) has not been addressed by previous model based vision work. Rohr (1994) did

observe that clothing could cause complex illumination phenomena and introduce difficulties in model matching. He also identified that this had led much previous work to be based on the assumption that joint positions were marked e.g. with reflective or luminous disks (e.g. Chen, 1992 and Quian, 1992). He did not however, offer any solution to this problem other than allowing the dimensions of his articulated cylinder models to be slightly expanded from the medical data used as a reference to limb dimensions. For many application areas (e.g. rehabilitation) the problem of attire was circumvented as the subject could be constrained to dress in a helpful way. However many of the studies did address real-world detection and so their results were likely to be limited by the (usually implicit) assumption that attire would closely follow the outline of the body in a consistent manner.

The significant effects of clothing on pedestrian images allied with the need to detect other targets such as wheelchairs and to operate at low computation load meant that for this work spatial models would need to be fairly broadly defined. It is also evident from the last chapter that the proximity of the camera to the scene and its oblique angle mean that some form of 3-D modelling was likely to be necessary. The particular methods developed in this work are described in the next chapter.

This concludes the review of the detection algorithm methods used in previous comparable work. The final sections in this chapter look at some additional issues for the vision system research that were particularly relevant to this project.

5.4 Alternative sensor technologies

5.4.1 Colour

It was apparent from the literature that reductions in the cost of colour cameras were leading to increasing attention being paid to the analysis of colour spaces. The important advantage offered for outdoor applications was that under normal circumstances the colour (but not intensity) of objects should be invariant to illumination effects within shadow bodies. Promising results have recently been reported both for the segmentation of pedestrians from real-world scenes (Heisele 1997, Wren 1995) and for the colour fingerprinting of vehicles (as a means of extracting origin-destination data) also under outdoor conditions (Chachich 1997). As was explained in Chapter 3 it had been decided to explore a monochrome solution to this research's task, however the success of these methods means that the use of

colour may be worthy of further consideration at a later date. The main concern would be that at night-time there may be a lack of colour information due to the more limited spectral content of street lighting compared to daylight. In the above-cited papers, the evaluation described is limited and makes no mention of how their systems would be expected to operate under street lighting.

5.4.2 Infrared (Thermal)

Another spectrally based approach to segmenting pedestrian data from a complex background has been to use an infrared camera to monitor a body's heat emissions. Iwasara (1997) extracted a pedestrian from the background by means of a simple intensity (i.e. temperature) threshold. The environmental conditions were not specified but it is clear from the example given, that the background was cool in comparison to the pedestrian. For the purposes of this work such an approach suffers from two disadvantages; the high (although decreasing) cost of thermal imagers and the expectation that variation in background temperature would be considerable under solar illumination. The latter problem combined with the variation in the degree of insulation offered by pedestrian clothing, means it is unlikely that such an approach would be sufficient for the separation of pedestrian activity from the background - without addressing similar image analysis problems as exist for a sensor at visible wavelengths.

5.4.3 Stereo Vision

Although the majority of work in computer vision is monocular (i.e. using only a single camera) there is also a substantial body of work devoted to the study of multi-ocular and, in particular, stereo vision. The appeal of stereo vision is that it provides a method of extracting information on the range (distance) of image features from the cameras. Stereo vision would appear to be suitable for the current problem, as one of the major distinguishing characteristics between objects and shadows is that shadows usually lie on the pavement surface and are therefore at a different range. Indeed a pedestrian detection system produced by Vision Ltd was based on stereoscopy but unfortunately, commercial constraints meant that no details of its construction and operation were available.

The major problem with the implementation of a stereo vision system is that of finding, for each feature in one image, the corresponding feature in the other image,

such that depth can be inferred from the relative feature positions combined with knowledge of camera separation. This is referred to as the correspondence problem. A consequence is that depth measurements can generally only be obtained where there are image features of high gradient along the x-axis in both images and they can be unambiguously matched. The difficulties in overcoming the correspondence problem are compounded by the fact that features may differ dramatically in intensity due to camera effects, or may even be occluded from one of the viewpoints. The high densities of strong edge information typical of this task particularly where there is background clutter (litter and leaves) would also be likely to lead to frequent ambiguity in correspondence searches and make finding accurate correspondences particularly difficult.

There are also a number of practical problems introduced by the move to using two cameras. Firstly, there is the increased equipment overhead of requiring two cameras and frame-grabbers. Secondly, the various characteristics of the two cameras must be locked together (a feature only available on more expensive cameras intended for machine vision) as must their capture by frame-grabbers. Frame capture synchronisation is also necessary to ensure image differences represent disparities due to object position and not their motion. Common exposure control would also be valuable otherwise features may have very different contrasts in the two images leading to a loss of clear correspondence. Thirdly, there are the computational problems of having double the data rate for transfer and processing. A final point is that when capturing video data for evaluation it is difficult to store the two cameras' video streams, in a synchronised manner, onto two separate video recordings for off-line analysis. Even if this were achieved, the problem of playing the results back in synchronism would remain.

In the light of the above it was decided to continue to concentrate on finding a solution using a monocular method. Nevertheless the potential value of obtaining depth information was such that some work was carried out, in parallel to the main work of this thesis, to try and overcome some of the difficulties mentioned above. Current progress in this respect is reported in Chapter 8.

5.5 Computationally Efficient Vision

In computer vision and image processing the response to persistent limitations in computational power has been to look for computationally efficient methods. Historically this led to processing being partitioned into two distinct stages; pre-processing and high-level processing.

The former is responsible for reducing the huge volumes of data from the sensor by using simple, but fast, operations (which are typically both data and spatially independent) so that the subsequent high-level analysis, which requires more operations per pixel, could be performed within available computing power.

In the past, the pre-processing reduction in data volume was often achieved by reducing the intensity resolution of each pixel, typically to produce a binary image. Use of a fixed empirical binarisation threshold in pedestrian analysis has, for example, often been used (Richards 1995, Rourke 1994) as a means of rapidly reducing the volume of data to be processed. This approach is however only likely to be of value for indoor conditions with diffuse lighting such that any intensity variations are only due to objects and not to environmental factors. Richards (1995), whose application was for outdoor transportation analysis, even pointed out the difficulty in choosing a threshold level as being a balance between missing signal and getting unwanted noise.

From the analysis in the previous chapter it can be seen that such binarisation approaches are not appropriate for this task, due to the low contrast of pedestrians combined with the high contrast of litter and leaves etc., as they would lead to important and unrecoverable information loss. The broader literature in computer vision however describes alternative methods of achieving rapid data reduction that may be better suited to this task. These methods are based on the control of spatial and temporal image resolution, and are discussed further in Chapter 6.

5.6 Evaluation Methods

A final aspect of interest observed in the prior work in this area, and indeed in computer vision as a whole, regards the method and extent of performance evaluation undertaken. In nearly all of the work reviewed, the emphasis of the papers was on the development and justification of the algorithms being presented. Results were usually presented based on the analysis of just a few images. This was often the case even for work addressing practical aspects of gathering transportation information where long-

term tests would properly be required. Even within the limited number of evaluation images shown it was usually the case that conditions were somewhat simplified with respect to the worst case conditions that could be expected to occur. The most common simplifications were for lighting conditions to be overcast (meaning no, or weak, shadows), pedestrian contrast with the background scene to be strong relative to other objects present or for there to be no occlusion. When it is considered that most of this research had already made use of powerful assumptions to reduce the influence of such effects, it would appear that the practical application of most of these systems in general outdoor conditions would be unlikely to produce results of the reliability required for this project.

A notable exception to the above shortage of evaluation data was Gavrila (1996) whose assessment was based on a one-hour sample of video at full video frame-rate. Even this however was a short test period in terms of gaining confidence that a set of algorithms would perform over periods of days, weeks or even years according to the particular application.

The usual means of making the evaluative comparison to test video data was manual assessment. An exception in this case was the work described by Sarma (1996) who considered the problems of evaluating extracted pedestrian trajectories. After initially extracting pedestrian trajectories and comparing them on a frame-by-frame basis to reference data he moved to an event based comparison method triggered by pedestrians passing pre-set landmarks in the scene.

Another interesting approach to the evaluation problem was the use of synthetic data (Attwood 1989, Rohr 1994) generated using computer graphics. This provided an 'ideal' source of reference data on pedestrian position, which would be particularly valuable for methods looking at detailed analysis of pose and gait. Its principle limitations lay in the quality of the models and their rendering and the difficulty in generating synthetic pedestrians that would follow realistic behaviours. In the author's opinion this is an area that it would be interesting to pursue further, particularly for adaptive vision systems which might use the synthetic data as the basis for initial training.

5.7 Conclusion

After looking at a wide range of applications of computer vision techniques to the sensing of the human body, it has become apparent that previous workers have based their work on assumptions that are not consistently valid for the pedestrian detection objectives of this work.

It was discovered that much of the prior art work was more demanding in terms of the level of detail of pedestrian position and activity that was sought (e.g. pose extraction or activity classification). This work would commonly make simplifying assumptions to ease or even remove the difficulties associated with separating pedestrians from the background scene. However it was also found that even in work addressing the extraction of more basic information, constraints on the problem domain were utilised (implicitly as well as explicitly) so as to remove many of the most difficult problems of implementing of a real-world system in a relatively unrestricted environment. The most common constraints used were based on assumptions related to camera position, pedestrian shape and direction of motion, lighting conditions and the absence of distracting background objects of comparable contrast to pedestrians. The use of these constraints, although justifiable by particular applications in most cases, meant that the methods developed were in general not applicable to vision in general outdoor conditions, where reliable levels of performance must be achieved without the freedom to make such assumptions.

This review has nevertheless identified an important point of commonality in many of the systems examined, namely the use of background modelling followed by the application of top down models. It was also determined that, in an environment where pedestrians are often of very low contrast, in the image it is essential to avoid premature information loss in pre-processing stages of an algorithm which might be usefully exploited by a later higher-level analysis stage. This approach provides the framework for the algorithm development work described in the next chapter.

The other main point to take forward from this review chapter is the degree of evaluation used in prior work. The attainment of required performance levels can only be verified by extensive evaluation under a wide a range of test conditions. In the prior work examined however, the degree of evaluation was on the whole very limited, in many cases consisting of the analysis of only a few images. Even where

more extensive evaluation had taken place, this was still often only over a sequence of a few minutes, covering a range of environmental conditions that was far from representative of the full range that practical systems would expect to encounter. Reasons for this are based in the difficulties of performing evaluation in terms of handling large sets of video data and in establishing reference ground truths. These matters are the subject of Chapter 7 which follows a description, in the next chapter, of the development of detection algorithms.

6 Detector Development

6.1 Introduction

In previous chapters the detection task has been defined, knowledge of the problem domain analysed and relevant prior art in pedestrian detection and computer vision examined. In particular the material of Chapter 4 identified the key areas of difficulty and that of Chapter 5 examined methods from the prior art for addressing them. In this chapter the above work is utilised in the design of algorithms for pedestrian detection.

In the earliest stages of this work, certain methods based on computer vision techniques that had been developed to address objectives similar to those of this thesis appeared to warrant further consideration. The first section of this chapter describes the author's exploratory investigation of these techniques. It should be noted that this exploratory work was completed before the full range of difficulties imposed by environmental variation, as described in Chapter 4, was understood and as such were limited in their performance. These investigations were however a valuable part of the research process and the knowledge gained was important in guiding the formulation of the final detection algorithm produced in this work. After the description of exploratory work the author's own final approach is described and justified in a description of the main body of work on algorithm development.

During the course of the algorithm development, it became apparent that a significant amount of work would be required on the creation of a user interface to the detection system, to allow detection parameters to be specified and to assist in calibration. This became of particular importance when the system was independently assessed by the Department of Transport (see Chapter 7), which meant that it had to be put into a form that was useable by non-expert users. The user interface that was developed is described in Appendix B and aspects of it are referred to at relevant points in the

algorithm description that follows. The task of evaluating algorithm performance was also central to the development process but is left as the subject of the next chapter.

6.2 Exploratory Work

6.2.1 Feature Tracking

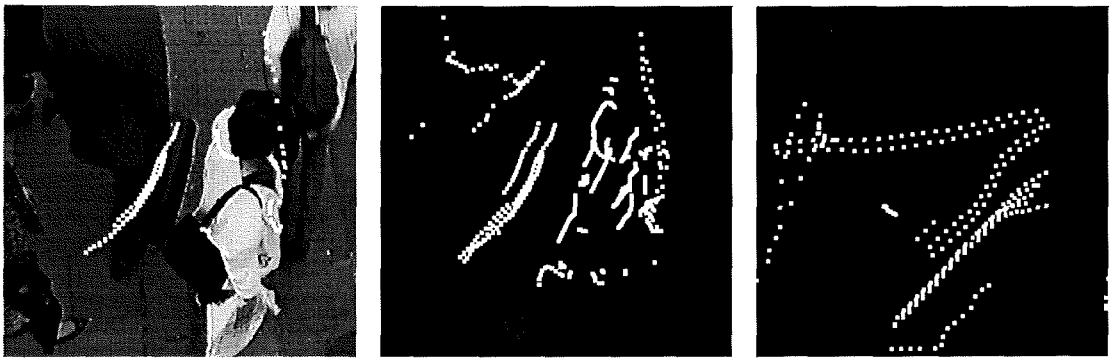
As was described in Chapter 5 a number of researchers have based their work on finding the correspondence between grey-level features in consecutive image frames. For example, Velastin et al (1994) based their analysis on the extraction of an optical flow by taking a pixel and its neighbourhood from a first image and searching for a corresponding pattern in a second image captured a short time later. From this correspondence they deduced a magnitude and direction of movement for each foreground point in the image – the foreground was identified by background subtraction, motion or intensity gradient measures. A drawback of this approach is that the search for feature correspondences is extremely computationally demanding, particularly as a high frame-rate is necessary to minimise inter-frame feature distortion. In Allsop (1997), the number of pixel correspondences to be found was reduced by pre-selecting just those grey-level features that represented corners i.e. regions of high horizontal and vertical gradient. This reduced processing requirements and led to more reliable tracking by removing difficulties due to the aperture problem.

For the purposes of this work, it was decided to investigate the efficacy of this feature tracking approach. It was, however, necessary to modify the above methods, as their implicit assumption that all objects of interest were moving was not valid for his task. As in the prior work, and in keeping with the active vision paradigm, features of interest were identified from the whole image and then only these features were selected for monitoring. However rather than only looking for feature correspondence between a pair of images, each pattern was tracked over a prolonged sequence. The intention being that periods where pedestrians were stationary would be monitored as part of the individual trajectory of each feature - rather than being ignored as areas of zero motion activity, as they would have been in the earlier work described. Once the trajectories had been obtained it was then intended that analysis of their characteristics (e.g. speed, direction and rate of turn) would yield useful information on pedestrian presence and behaviour.

The implementation was based around two operators, an interest operator to pick out the strongest corner features for tracking and a pattern-matching operator to search for features in subsequent image frames. The processing load was reduced by using low-resolution representations of the image to select features of interest and then concentrating the vision system's attention on tracking the selected features at full resolution over a narrow field-of-view. In an attempt to further decrease the scope of the search task (and hence the computational load) the use of frame-to-frame prediction based on assumptions of constant position, constant velocity and constant acceleration was also investigated. The irregular stop/start nature of pedestrian movement meant however that no significant benefit was achieved from the use of predictors.

The software consisted of two important objects. A *periphery-monitoring* object contained the interest operator and detected the strongest features in each incoming, reduced-resolution frame from the camera. The foveal functions of maintaining tracking information and performing pattern matches between frames were embedded into a *tracker* object. In operation the *periphery-monitoring* object then assigned unmonitored features of interest to free members of an array of the *tracker* objects. The *tracker* objects followed their assigned features as long as possible and then registered themselves as free whenever a pattern was lost.

Using this method grey-scale features were tracked successfully over extended sequences. Examples of feature trajectories extracted from video taken from the kerbside detector's viewpoint are shown in Figure 28.



(a) (b) (c)

Figure 28 *Feature Tracking: Frame (a) shows the history of a number of tracks currently being tracked that have been overlaid onto the latest image. Each white*

block represents a feature match and the spacing between the blocks therefore indicates a feature's velocity of movement. Frame (b) shows the tracks in (a) as well as several which were accumulated over a short time previously. In (b) it is apparent that many short tracks start and terminate well within the image, indicating the failure to find a feature until it was in the centre of the image area and the subsequent early loss of the feature for reasons indicated in the main text. Frame (c) shows accumulated tracks for a later period and indicates the main flow of pedestrians is diagonally past the signal pole - which is just off the bottom right corner of the image. In all cases tracking was performed at a rate of 25 frames per second. At the bottom of (c) a weakness of the tracking algorithm is apparent where it has become trapped on the image boundary. This is a consequence of the adaptation mechanism explained in the text.

In practice, this method was found to suffer from several weaknesses in both the feature acquisition and tracking actions of the algorithm.

During the feature acquisition phase by the *periphery-monitoring* object, it was found that there were too many potential features of interest, which were unrelated to pedestrian movement. These included those due to litter and leaves as well as intermittent features from pedestrian and vehicle shadows (see Chapter 4). It was also found that the interest operator was frequently more likely to elect to follow shadow boundaries than pedestrians as these were often of higher contrast under direct sunlight. Furthermore direct sunlight tended to emphasise complex background texture in the pavement to produce multiple, intermittent, high-contrast feature candidates such that the number of features to be tracked became prohibitive. An additional difficulty was the tendency of the interest operator to follow 'phase' points i.e. a projected overlap of two different world features such as the overlap of the woman's hair and coat collar in frame (a) above, which as they moved apart meant the feature suddenly ceased to exist.

During the feature tracking process the main drawback of this method was that feature pattern loss was too frequent due to occlusion and distortion of features resulting from illumination variation, object deformation and rotation. Indeed other researchers have found that rotation by more than 3 degrees or dilation by more than 3% led to loss of correlation peaks (Schalkoff, 1989). These distortion effects were to some extent offset by changing the *tracker* object to update the target feature pattern on each cycle

in which a successful match was found. This adaptive measure improved the lengths of trajectories extracted, however feature loss was still too frequent and the adaptation occasionally led to tracked features merging into background features and becoming permanently trapped, see Figure 28(c).

Computational load was high because of the close proximity of the camera to the scene under observation and the consequent high image velocities of features. This required that processing be carried out at a high frame-rate to keep inter-frame feature distortion sufficiently small as to allow correspondence to be established and also to limit the range over which a search for correspondence had to be carried out.

For the purposes of pedestrian detection, it was decided that this approach did not merit further investigation because of the above problems and difficulties in interpreting the results of the tracking which were in turn due to the many complex motions within each pedestrian from limb and clothing movement. The method was however found to be effective in vehicle tracking where operation was at a longer range from the camera, pavement texture was less apparent and there was no change in object shape.

This work indicated the importance of including a spatial model-based approach to resolve questions that cannot be answered just by examination at the pixel/feature level. It also emphasised the need to develop methods that exhibited some contrast invariance (for detecting faint pedestrians in the presence of strong shadows), and tolerance of a high degree of background edge content (clutter) due to shadows, pavement texture and the presence of litter and leaves.

6.2.2 Relative Intensity based pixel classification

This approach sought to build on previous work by members of the Napier vision group (Darkin 1986) developed for vehicle detection in outdoor conditions. It addressed the problem of shadows by using assumptions about their intensity relative to the background scene. This was combined with a simple spatial model to discriminate between pedestrians and various other aspects of the scene that appeared as lines in the processed image (e.g. the peripheries of building and vehicle shadows). The structure of the algorithm is shown in Figure 29.

A background representation was produced by forming a pixel-wise long-term average of the incoming video stream. A user-defined mask drawn onto an image of

the scene by an operator was used to identify regions in which the detector was to be active. Further explanation of these methods can be found under the main algorithm description, section 6.3.1.3.1.

An assumption was then made that there were bounds on the degree by which a pixel's intensity could vary with respect to the background value and yet still be considered to be the same as it. Two multiplicative factors defined these bounds. The lower bound for each pixel was specified by a fractional factor expressing a proportion of the background value. Similarly the upper bound was specified by a factor greater than unity. Based on this assumption, pixels in incoming images were then classified as being either the same as the background if their intensity lay within the bounds, or as being objects of potential interest if they were outside them. To add stability against the, essentially random, intensity scaling effects of the camera's automatic exposure control mechanism, the bounding values were scaled to compensate for any variation in the global average intensity of the image under analysis relative to that of the background image. The result of this stage was a binary image indicating just those areas where it was believed that pedestrian objects might be present.

The resulting binary image was then analysed in two ways to generate the presence and volumetric detection outputs independently. The volumetric output was obtained by simply counting the number of pixels in the active detection area and applying an empirical scaling factor to relate this to the number of pedestrians present. A linear relationship between activity and volume was assumed, an assumption that had been found reasonable in previous work by another group (Yin, 1996).

The generation of the presence detection output used a simple shape analysis based on filters sensitive to the number of consecutive pixels in a group. Firstly in a 'horizontal run' filter, each scan line of the binary image was examined and all pixels which were not part of a continuous run of at least 'n' pixels were removed. An identical operation was then applied but considering the image data column by column in a 'vertical run' filter. Using the value 'n'=3 in both rows and columns had the effect of removing all pixel groupings that were either too narrow and/or too short to pass through the filter. In practice this was found to be effective at removing point speckle noise and straight lines due to shadow boundaries and pavement features. If after this process any pixel groups remained in the detection regions, a binary detection was

flagged. Finally, consistency of the binary and volumetric outputs was ensured by forcing the volumetric output to zero, if no binary detection had occurred.

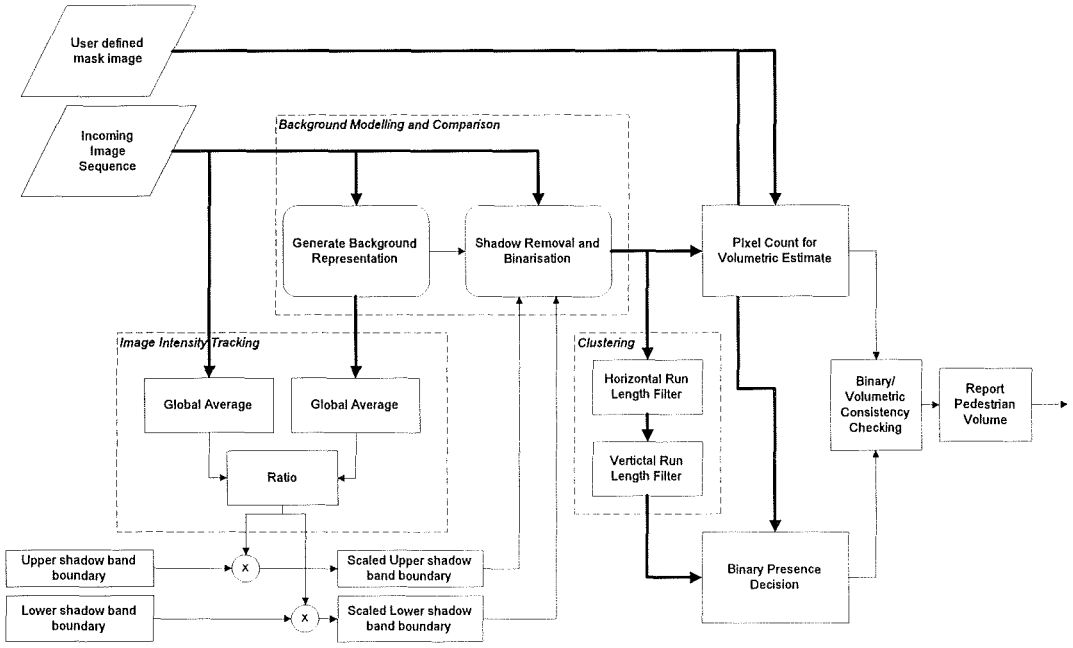


Figure 29 *Pixel Classification Algorithm. Dark and light paths indicate image and parameter flows respectively.*

When implemented, a processing rate of around 6 frames per second was achieved on a development system consisting of an IBM PC compatible with Pentium processor running at 60Mhz. This could however have been significantly improved, as all the software was optimised for development flexibility rather than speed.

The performance of this system was deemed sufficient to justify a quantitative evaluation (Reading, October 1996). The test data used was a set of four half-hour video sequences supplied by the Department of Transport. The results showed an approximately linear relationship between a reference (manually assessed) pedestrian volume and the algorithm-derived estimate. Performance was also analysed in terms of false negative and false positive detections of which 2 and 56 were recorded respectively. Although the false negative rate was low, the number of false positive detections was high due to poor handling of shadows, a problem exacerbated by the reaction of the camera's exposure control mechanism to passing vehicle shadows. The following conclusions were drawn from this piece of work:

- The assumption that shadow intensities lay within a fixed ratio band of a background level, which had been used successfully for traffic analysis, was

found not to be valid in general for the more complex pedestrian situation. The empirical selection of a binarisation threshold to detect low contrast pedestrians, without responding to noise and shadow effect, proved difficult under many lighting conditions.

- The use of binarising thresholds prior to shape analysis resulted in the loss of important information on pedestrians, which were often completely or partially of low-contrast with respect to other components of an image. The imposition of empirical decision thresholds should therefore be avoided until as late as possible in the image analysis.
- Global monitoring and tracking of intensity variation was not appropriate for outdoor vision tasks, as much of the actual variation that occurred was due to localised illumination effects (shadows and highlights). Global analysis is not appropriate for outdoor, shadowed scenes as confirmed by work of Velastin, 1982.
- The use of shape information rather than contrast will be needed to distinguish shadows from real objects. There will inevitably be transient signals that pre-processing stages will not eliminate and which will need to be discriminated against by spatial models. Basic spatial modelling by run-length constraints was effective in discriminating against noise and showed value of combining pre-processing with a top-down spatial model based approach.
- The method as described was reliant on empirical parameters at the shape filtering stage and also for scaling of pixel activity to obtain a volumetric estimate. These parameters were site-specific and so had to be adjusted by hand. Furthermore no account was taken of the effects of perspective in scaling projected pedestrian size. To deal with the larger variation in projected pedestrian viewpoint and scale with video taken from a lower camera mount, it would be necessary to introduce some calibration knowledge of the three dimensional structure of the scene for a useful level of accuracy in pedestrian counting to be achieved.

The next section describes the design and development of a detection algorithm designed to overcome the deficiencies identified in this exploratory work.

6.3 Model Based Algorithm

The algorithm presented here was based on observations of the task analysis (Chapter 4), combined with knowledge of the effectiveness of various vision methods as explored in the previous sections. The resulting algorithm consisted of the four stages indicated in Figure 30.

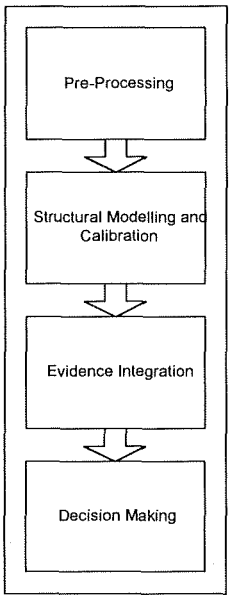


Figure 30 *Overview of algorithm structure*

For the purposes of this work, operations in the pre-processing section were those that involved the spatially invariant, pixel-wise processing of images without the use of knowledge of higher-level (object) structure. Each pixel of the image was processed identically using methods aimed to remove information likely to relate to non-pedestrian objects and scene artefacts, whilst minimising any loss of pedestrian signal. The methods used were related to those described in the exploratory work and can be viewed as improving the signal to noise ratio.

Information on the pre-processing methods developed can be found below. In the subsequent section models of pedestrian shape are developed along with the necessary user interaction and calibration processes, to establish their relationship to the world scene. Thereafter an evidence integration stage is described which applied the spatial models to the output of the pre-processing stage, to find candidate pedestrian positions. Finally, a decision making process is described which determined whether there was sufficient evidence for each candidate position to consider it as being due to a pedestrian.

6.3.1 Pre-processing

A breakdown of the components of the pre-processing into sub-stages can be seen in Figure 31 below.

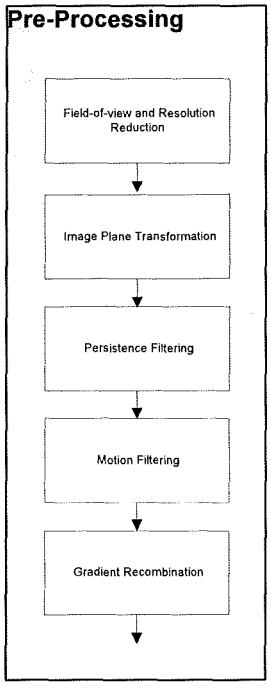


Figure 31 *Image Pre-processing Operations*

The details of each of these sub-stages, which constitute the pre-processing, are now described in the following sections before moving onto structural modelling issues.

6.3.1.1 Field-of-view and Resolution Reduction

A major operating constraint for the detector was that it should be able to run on low-cost hardware. As the various stages of a computer vision system are computationally intensive, it has been identified that it is efficient to reduce the volume of image data to be processed as early as possible in a system (Burt, 1991) to avoid any unnecessary processing of image information. Other workers addressing this problem have developed a large body of work based on the active control of operating resolution and field-of-view. Much of this work was inspired by observations of the human eye's use of a combination of a low-resolution (pre-attentive) peripheral vision system with a high-resolution (attentive) foveal vision system.

Methods known as coarse-to-fine analyses have been developed which operate by identifying features of interest at coarse resolutions, to reduce the computational cost of the initial search, and then use the results to spatially constrain the search for more

detail at progressively higher resolutions. An important prerequisite for the use of these methods is that features exist across the full range of scales and in particular that they must be discernible at the lowest resolution representation of the scene which is used for the initial search. In addition a means must be identified of generating the reduced resolution images, which is both computationally efficient and doesn't introduce any shape distortion of imaged objects such that their identification would be prevented.

In this work it was decided that rather than working within a multiple resolution hierarchy a simple dual resolution sensing model would suffice. Firstly, a peripheral vision system (PVS) would be designed to monitor the scene at a coarse level of resolution and provide pre-attentive information on potential pedestrian positions. It was expected that the result of the peripheral vision process would suffice for the task of binary and volumetric detection and indeed this is the subject of the work described below. However it was envisaged that a foveal vision system (FVS) operating at full sensor resolution, could be added later to perform higher-level functions such as tracking of individual pedestrians based on the cues provided by the PVS.

The first measure of data reduction used was to control the active field-of-view by the restriction of all processing, to just the active detection zone specified by the installer. Typically this resulted in about half of the image requiring no processing. The specification of the detection zone by the user is discussed in section 6.3.2.3.1 and examples can be found in Appendix B.

Secondly a stage was added to allow the spatial resolution to be chosen to suit the task rather than, as has often previously been the case, being determined by the resolution of the video camera. The limit on the degree of resolution reduction that could be used for the PVS was dependent on the size of pedestrians (when projected into the image) and the method of reduction.

A variety of methods have been described in the literature for achieving resolution reduction. The simplest of these is to sub-sample by taking every *n*th pixel to achieve a reduction by the required factor. It has the advantage of being very fast to implement, however considerable aliasing occurs, distorting the appearance of the image content. Methods designed to minimise this distortion at the expense of more processing, include the use of averaging over blocks of *nxn* pixels and the hierarchical

application of gaussian filters (Burt, 1981) known as the gaussian pyramid. The block averaging and gaussian methods both use low-pass filtering to reduce the shape distortion introduced by aliasing. Block averaging can be seen as the application of a low-pass filter of uniform weight, whilst the gaussian pyramid uses a local average weighted by coefficients of gaussian form so as to minimise aliasing and offer better preservation of shape.

For this task, however, it was decided that using the simpler sub-sampling scheme would be preferable to low-pass filtering methods because:

- The presence of a pedestrian over the background can be considered as a step function, in signal processing terms, and the removal of pixels during sub-sampling will not lead to loss of the object but only to its contraction, provided it is sufficiently large in the field of view.
- The large degree of variation in pedestrian form meant that structural models of pedestrian shape will not be very detailed and so the shape distortion of the pedestrian images shouldn't be a handicap to further processing.
- Smaller objects, such as litter, which only occupy a few pixels may well be removed or usefully decreased in their influence on subsequent processing.
- Pedestrians although large in the field of view, are often of much lower contrast than much smaller distracting objects such as litter and reflections. The use of filters with low-pass characteristics (i.e. averaging or gaussian) tends therefore to emphasise the unwanted signal in favour of the pedestrian signal, by spreading the signal energy around the pixel's locality. Support for this point-of-view comes from the work of Pye, 1994, who advocates resolution reduction methods based on non-linear functions (e.g. median filtering) in order to achieve better preservation of edge features.
- Even using specially designed fast implementations of the low-pass reduction methods, there is considerably greater computational overhead - particularly as the frame-grabber used incorporated a sub-sampled scaling function which could be used to perform all the reduction in hardware.

Taking the above into consideration and as the detection objective is only concerned with determining pedestrian presence, rather than identification, a sub-sampled

resolution reduction stage was included in the software. The option of reduction by integer factors between 1 and 8 relative to a captured resolution of 256 by 256 pixels was included with a typical operating factor of 4 being used to reduce to 64 by 64 pixels. Combining the effect of the field-of-view and resolution reduction measures therefore resulted in a reduction of the volume of data to be processed in each video image from 64Kb to about 2Kb.

The final data reduction method incorporated was a temporal sub-sampling which made use of the fact that the required detector response time was 500ms (Chapter 2). This meant that only two images per second needed to be processed resulting in a data flow to subsequent processing stages of 4Kb per second.

6.3.1.2 Image Plane Transformation

As the camera used was a standard mass produced device, it produced an image in accordance with standard television standards and hence with an aspect of 4:3 between the horizontal and vertical dimensions. Given that field-of-view was already stretched to cope with the size of detection area required at such close range, it was important to make the best use of the available imaged area. For this reason the camera was used on its side (rotated 90 degrees about its z-axis) to get the best match between the aspect ratio of the camera and that of the detection area which is longer than it is wide. However this action left pedestrians oriented horizontally in the output video image. Although as far as the computer vision algorithms are concerned the image could have been processed directly in this form it was convenient to have pedestrians vertical in the image during the installation process when the detection zone was being specified. It also made human assessment of the scene much easier during evaluation and development.

To correct this a transformation was used to rotate the image such that pedestrians were vertically orientated with their feet towards the bottom of the image. This was achieved with little computational load by re-mapping the pixels from a captured input image $I_i(i,j)$ into an output image $I_o(i,j)$ by 90 degrees according to:

$I_o(i,j) = I_i(j_m - j, i)$ to rotate clockwise and

$I_o(i,j) = I_i(j, i_{\max} - i)$ to rotate anticlockwise.

The two alternatives were implemented to allow for the two possible orientations of the camera with respect to its casing.

6.3.1.3 Persistence Filtering

The objective here was to find a means of discriminating between objects on the basis of their time of presence at a fixed position in the scene, a parameter that will be referred to as persistence. From observation of the crossing environment it was clear that using this measure the effects of many permanent and semi-permanent (i.e. non-pedestrian) objects might be distinguished from the more transient pedestrian activity. To avoid the loss of important low-contrast pedestrian information that might be exploited by later stages of the algorithm it was decided that no decision-making (threshold type) operations should be used in this process. The output required was a pixel-by-pixel measure of the likelihood (as opposed to a binary decision) that a given pixel represented a short duration object - and might therefore be due to a pedestrian.

For this task, the mounting of the camera was effectively rigid as the only likely source of camera motion was vibration and this had been found to be negligible in practice (see Chapter 4). It was therefore valid to assume that the camera was effectively fixed with respect to the scene and the background modelling techniques, as reviewed in Chapter 5, should be applicable to this task

Background modelling is a powerful technique commonly used in vision systems in which the camera operates in a fixed position relative to the scene. The technique involves the generation of a representation of permanent scene content, referred to as the background. Transitory objects of interest are extracted from the current video image by comparing it against this background using a matching operator.

For this work it was decided that the requirements of the background representation were that it should:

- Absorb newly arrived objects into the background after a pre-settable time of presence.
- Be quick to compute and amenable to hardware implementation. In practice, this meant that it should be based on integer arithmetic and able to be calculated in a fixed-point accumulator of limited numerical precision.

- Initialise to a useable state quickly from start-up to allow the detector to be operational within a short period of start-up. This required some thought due to the fact that background representations are formed by assessing scene content over a necessarily prolonged period.

The numerous approaches to the use of background modelling that have been used by previous researchers were described in the review of Chapter 5. Based on those discussions the development of the methods used in this work is described below. The next section covers the means used to generate the background representation whilst the matching method for comparison of the background to the current image is dealt with in section 6.3.1.3.2.

6.3.1.3.1 Generation of Background Representation

Having recognised that instantaneous lighting and shadow variation would mean there was no perfect background modelling method for this application (see Chapter 5), it was decided to avoid the more complex non-linear background modelling processes that some workers have used. It was preferred instead to base the background modelling on a low-pass IIR filter to gain a stable reference to the relative intensity of those image artefacts that are temporally separable from pedestrian objects, with the intention of removing other effects at a later point in processing. The stability of his method would also make it more suitable for secondary functions such as alignment monitoring.

In previous work (Seed 1988) it had been discovered that use of this method with integer data storage of limited precision could cause the result to appear artificially quantised. The section below describes the author's formulation of an IIR filter to operate effectively in integer arithmetic (which is valuable for hardware implementation and the reduction of computational load) and its adaptation to allow rapid initialisation.

The basic filtering process is described by:

$$B_{i,j,t} = \frac{(\tau - 1)}{\tau} B_{i,j,t-1} + C_{i,j,t}$$

Where B represents the accumulated background image and C the current image from the framegrabber at time t. The parameter τ is an integer time-constant that defines the low-pass characteristics of the filter. As there is no neighbourhood

interaction between pixel values, we need only consider the response b at a sample background pixel from B to a pedestrian arrival modelled as a step change in intensity in the current image from zero to c. From the above equation therefore:

$$b_t = \frac{(\tau - 1)}{\tau} b_{t-1} + c$$

Substituting a for $\frac{(\tau - 1)}{\tau}$ to simplify notation and expanding we get:

$$b_t = c.\left(1 + a + a^2 + a^3 + \ldots a^{t-1}\right)$$

The maximum value that will be held in the background accumulator must be limited so as not to overflow the data type used to hold it (e.g. an unsigned16 bit integer). Therefore summing the geometric progression in brackets to infinity we find that the final value will tend to:

$$b_\infty = \frac{c}{1 - a} = \tau . c$$

Using this result, plus knowledge of the maximum input step size that will be encountered (256 grey levels) and the maximum value that can be held by the accumulator's data type we can set a maximum value for τ. The working value used was $\tau = 255$ with a 16 bit unsigned integer accumulator to hold B thus making full use of the available numerical precision in the accumulator.

From this equation we can also see how to obtain a filtered output from the accumulator b of the same scale (255 grey levels) as the input τ as:

$$c = \frac{b_\infty}{\tau}$$

As mentioned earlier, it was desirable to speed-up the early stages of background image formation during system initialisation or realignment. The simplest methods considered were to simply initialise to an, essentially randomly, captured image at start-up or alternatively to modulate the time constant to get a faster initial response. These methods however would have put an unwanted weighting on objects that happened to be in the scene during initialisation and this influence would have taken several minutes to decay. It was a concern that the errors so produced would be a source of confusion for an installer. The method used therefore was to derive a time-

dependent pre-scaling factor to amplify the output of the background filter for the time period immediately after an initialisation.

This factor was found by summing the first part of the geometric progression to find the changing response of the background filter during the initial period following the application of the input step, giving:

$$b_t = c.\left[\frac{1-a^t}{1-a}\right] = c.\tau.\left[1-\left(\frac{\tau-1}{\tau}\right)^t\right]$$

An output, o_t scaled to the range of the input pixel intensities was therefore obtained from:

$$o_t = \frac{b_t}{\tau}.\left[\frac{1}{1-\left(\frac{\tau-1}{\tau}\right)^t}\right]$$

This is the same as the previous result with the addition of the multiplying pre-scale factor in square brackets. The pre-scaling term will tend to τ as the number of samples t increases after initialisation. In fact to reduce computational load the pre-scaling was turned off after $3.\tau$ cycles at which point the result scaled by τ had already reached 95% of its final value.

The response of this background method is shown graphically in the figure below.

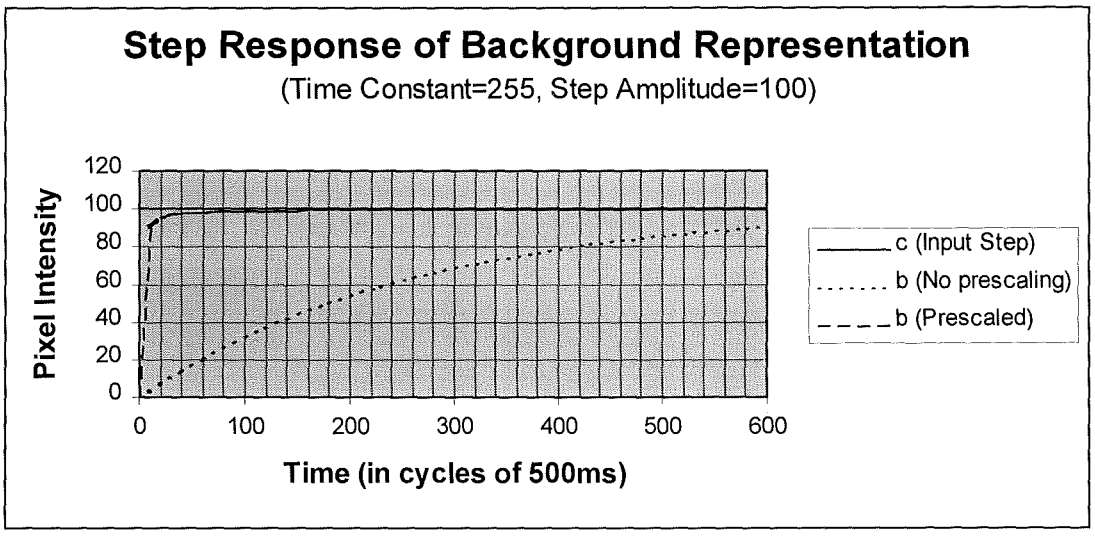


Figure 32: *The response of the background filter to a step input is shown along with the pre-scaled version used to accelerate the initialisation process.*

In the experimental work described below a time constant of 255 was used. For the frame rate of about 2 frames a second, this corresponded to around 2 minutes of real-time. This was a compromise between achieving negligible absorption over the maximum possible period of stationary pedestrian presence, whilst minimising the effect of background changes that could produce false positive signals until they were absorbed. This value of τ was sufficiently large that absorption into the background over the maximum time a pedestrian was likely to remain still was small (around 7% over 10 seconds) and equated to about 90% absorption after 5 minutes for objects of greater persistence.

The background generation functions were implemented in software as a standalone MISA class that, so far as the main algorithm was concerned, transparently handled the image management and calculation requirements. As the Image class (see Chapter 3) only supported 8-bit unsigned pixel data, the 16-bit accumulator was implemented using a pair of 8-bit images, splitting the 16-bit data between them as high-byte and low-byte. The power calculation for pre-scaling required floating point arithmetic but was only performed once per update of the entire background image, as the resulting factor was common to all pixels. Once calculated it was stored in scaled form as an integer for application on a pixel by pixel basis.

6.3.1.3.2 Matching Methods

The persistence filter worked by identifying features that were present in the current image but not in the background. The goal was to find temporary objects and remove

permanent ones. To perform this comparison a matching operator was required to measure the degree of similarity between these two images.

A major complication in performing matching was that the background intensity would always, by necessity, lag behind the variation in that of the current image as it tracked illumination changes. For this work it was accepted that it was not feasible to maintain a background that followed background intensity variations closely. Therefore a method was sought to match background and current images which offered some invariance to the expected differences.

This required a matching operator that would minimise the effects of changes due to variation in illumination level but identify changes that were due to a fundamental difference in the underlying image patterns, rather than the effects of shadow body or illumination change. Of necessity such a pattern analysis needed to examine a neighbourhood of pixels to establish the form of the pattern and exploit any directional information in the local image gradients. The use of directional information was found to be essential for distinguishing the presence of new patterns when they lay over background areas of high edge density, as magnitude-only based comparisons led to false matches.

In deriving a suitable comparison operator, the image formation process was considered in terms of the incident light intensity arriving at the pixel sensors and the conversion of this incident light into electrical signals within the camera.

The formation of an image by a sensor's conversion of 3D radiometric information (light) in the scene to image brightness can be modelled by the radiometric model described in Schalkoff, 1989. Image formation is modelled as a non-linear, multiplicative process given by:

$$f(\underline{x}) = e(\underline{x}).r(\underline{x})$$

Which says that the resulting image function, $f(\underline{x})$ is the product of the incident surface illumination, $e(\underline{x})$ and a parameter of the object called its reflectivity $r(\underline{x})$. The derivation of this model assumes that there are no light sources visible and that all surfaces in the scene are Lambertian reflectors i.e. they are ideal diffusers of light, such that emitted radiance is independent of viewing angle. Even for non-lambertian

reflectors the fixed relationship between the camera and background objects means that the model should still be valid. Constancy of illumination direction is also required but again can be assumed for the most significant light source, the sun, so long as the effects of its movement are slow with respect to the background time constant. The above also assumes the camera's conversion of incident intensity to output response is linear.

It was assumed that averaging over a period of a few minutes was sufficient to remove the effects of transient objects and rapid illumination fluctuations. The low-pass filtered background was therefore interpreted as a map of the reflectivity of all the permanent objects in the scene, weighted by a factor related to the average recent illumination. The average recent illumination factor had components due to both global and local characteristics associated with changes in overall illumination levels as well as those due to local variations. It was therefore not valid to assume a single global gain factor could be found that would bring the images into registration. In modelling the variation in image-to-image global intensity it was also essential to consider the camera's characteristics. As was discussed in Chapter 3, global changes to the exposure control were automatically used by the camera's internal control mechanisms to allow it operate over a wide range of illumination levels. This variation was beyond the control of the system for the camera in use. It was therefore treated as an additional unknown signal gain factor.

The matching operator was then derived in the form of a test of the hypothesis that corresponding regions of the current and background images contained the same pattern differing only by a scale factor. The operator used is derived below. It should properly be described as a measure of *mismatch* as a perfect pattern match resulted in an output of zero.

Firstly the image formation factors were treated as simple gains and combined (assuming they are constant over the spatially localised area of the matching operation) into a single multiplicative factor. To avoid these variations affecting the vision system, invariance to differing gains was required. This was achieved by scaling the current image pattern of scope defined by k,l about a pixel in the current image, $C_{i,j}$ to the level of the corresponding background pattern $B_{i,j}$, on the basis of

the brightness of the central pixels, to produce $C'_{i,j}$. The background was used as the reference to which the current image was scaled as it was the more stable, given its slow time constant, and also as it tended to sit in the middle of the grey-scale range. The resulting scaling function is expressed by the equation below. The variable $a_{i,j}$ is included to represent the effect of an object's presence at $C_{i,j}$ such that it is zero wherever the scene is empty of new objects.

$$C'_{i+j,k+l} = \left[\frac{B_{i,j}}{C_{i,j}} \cdot \left(C_{i+k,j+l} + a_{i+k,j+l} \right) \right]$$

If all the above assumptions were valid and no non-background objects were present, the background and the scaled foreground would be expected to be identical. A concern was that as this scaling was based on the two central pixels of the local pattern it would be susceptible to the effects of impulse noise. This could have been countered, at some computational expense, by the alternative use of a local average or median measure but such measures have so far been found to be unnecessary.

Now that image formation effects had been removed, a metric was needed to compare the background and scaled foreground patterns on a pixel-by-pixel basis. As discussed above it was considered important to build-in sensitivity to edge direction information. Knowledge of what the edge directions actually were was however not required so rather than losing useful information by calculating the gradient first, it was decided to a make a direct pixel-by-pixel comparison which would implicitly take account of differences in edge orientation. The absolute difference operator was chosen for this role as it was computationally faster than squared error measures and didn't weight the result towards high-contrast over low-contrast image features.

As described thus far the matching operation scaled the signal due to a pedestrian according to the relative intensity of the current and background images. A problem with this was a weakening of low amplitude signal within areas that were bright relative to the background. To ensure that a pedestrian intensity signal of given amplitude should result in the same output signal strength independently of the brightness of the background region it happened to have fallen on, the match signal

was therefore post-scaled by the reciprocal of the ratio of foreground to background intensity. This resulted in an entire matching process represented by:

$$M_{i,j} = \frac{C_{i,j}}{B_{i,j}} \cdot \sum_{k,l=-n}^{n} \left| \left[\frac{B_{i,j}}{C_{i,j}} \cdot \left(C_{i+k,j+l} + a_{i+k,j+l} \right) \right] - B_{i+k,j+l} \right|$$

It was found in practice that low value divisors needed to be dealt with as a special case. Accordingly an empirically selected threshold was introduced below which a mismatch signal of zero was returned. This measure was justifiable, given that the camera's automatic exposure control mechanism tended to keep image brightness around mid-grey level under most conditions.

The results obtained using the persistence filter operation combining the background extraction and matching operations are illustrated in the figures below. Operation under diffuse illumination conditions is shown in Figure 33. It can be seen how strong pavement edge features due to the white line along the roadside, paving detail and the traffic cone (front right) have been removed successfully. The value of the direction sensitivity of the matching operator is demonstrated by the signal obtained from the head of the right hand pedestrian in front of strong contrast white line at the road's edge (compare (d) and (e)). It can also be seen that the angle of the scattergraph (background (b) on y-axis and current image (a) on x-axis) is slightly less than 45 degrees due to the lag of the background behind current image intensity.

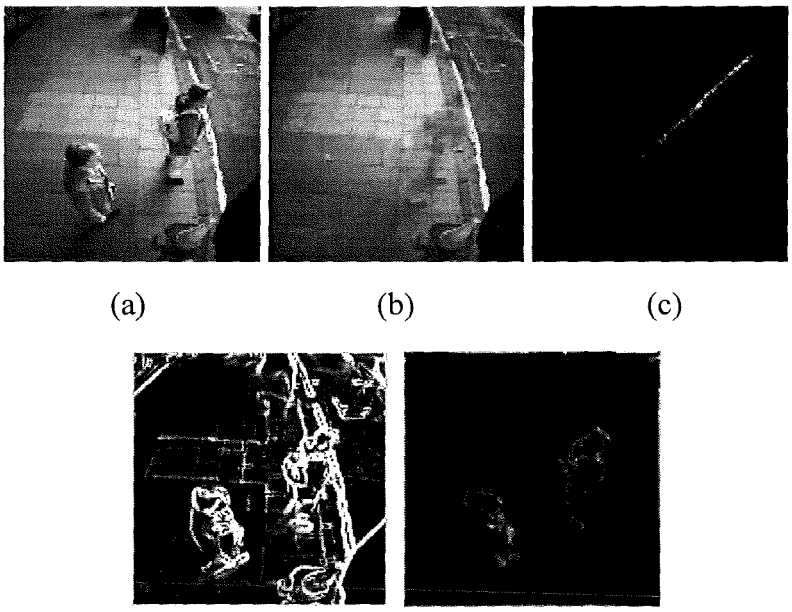


Figure 33: *Performance of the persistence filter under diffuse lighting conditions. The relationship between the current image (a) and background (b) is indicated in the scattergraph (c). The selection of features of interest from the gradient features in the current image (d) by the matching operation is shown in (e).*

The spread around the main diagonal on the scattergraph is due to pedestrian signal. It might appear that identification of the main diagonal would be a possible means of separating out the pedestrian signal from that of the background. If all changes were global to the image this might be the case, however when operating under more difficult conditions of bright sunlight with strong shadows, as shown in Figure 34 below, it can be seen that the scattergraph splits into multiple regions. These correspond to regional intensity variations in the image that could be due to pedestrian or illumination and shadow effects. The lower region, below 45 degrees, represents areas where the current image is brighter than the background, i.e. most of the image, and the more diffuse region at a steeper angle to the horizontal is principally due to the area covered by the shadow of a passing bus.

Figure 34 shows the filter's performance when different areas of the image are both above and below the levels in the background image in a situation produced by intermittent periods of bright sunlight. Strong background features on the footway and road surface are still removed without diminishing pedestrian signal. At the same time the signals from both pedestrians and shadow boundaries were, as expected, largely

unaffected – consequently a distinction between them had to be made between them at a later stage of the algorithm.

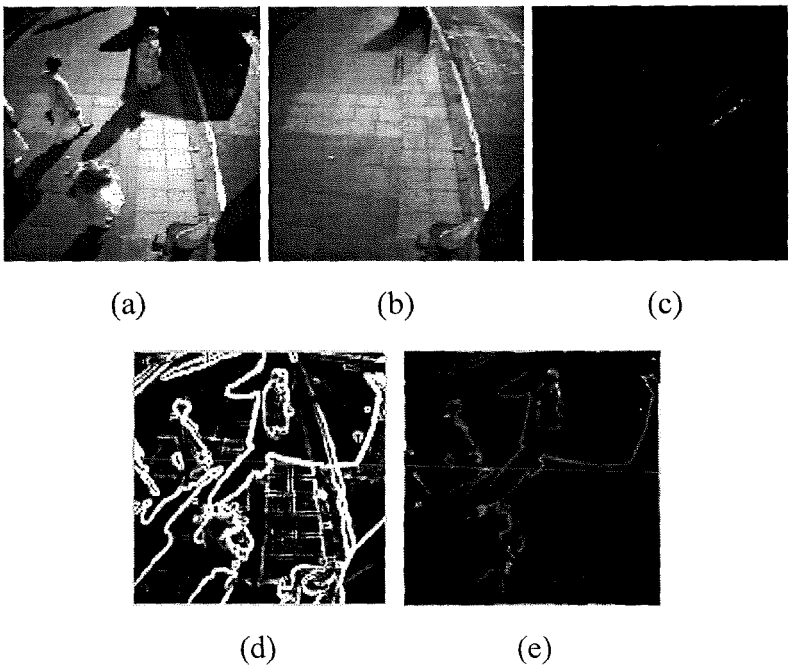


Figure 34: *Performance of the persistence filter under combination of bright sunlight and strong shadows.*

6.3.1.4 Motion Filtering and Noise Reduction

The objective here was to find a simple operator, of low-computational complexity, to remove the effects of fast moving patterns due to vehicle shadows whilst causing minimal disturbance to slower pedestrian related signals. By thus exploiting the difference between the speed of passing traffic (of up to 40 mph) and the maximum velocity of pedestrian movement (around 5mph) it was hoped to simplify the task of later processing stages. Although the fast moving effects were mainly due to vehicle shadows falling on the pavement they could also be due to the shadows of pedestrians and crossing infrastructure thrown by passing car headlights.

A cut-off velocity was defined, above which patterns were to be removed on the basis that they are moving too quickly to be pedestrians. In setting this cut-off velocity it was important to err on the low side to be sure to detect pedestrians who moved around whilst waiting.

The use of motion analysis methods often implies the need to process images at a sufficiently high rate so as to limit object movement between images. This is

exemplified in the use of optical flow techniques where the change between consecutive images needs to be restricted to the order of a single pixel. The consequent need for a high processing rate therefore tends to lead to a requirement for powerful processing systems, which were not available given the constraints on this work.

To overcome this problem a method was developed which operated by sampling the video stream from the camera whereby pairs of images, of known temporal separation, were captured for processing at each cycle of the detection process. This method of temporally inhomogeneous sequence sampling allowed the system to make use of the full temporal resolution of the sensor, whilst keeping the average data rate to be processed low, due to the relatively long interval between the capture of image pairs. Rather than requiring a set of direction vectors such as generated by an optical flow algorithm, the task here was simply to provide a low-pass filter on the velocity of object movement and so computational requirements were expected to be low.

The filter was implemented by the enforcement of a maximum allowable movement between the pair of captured frames. This was achieved by extracting the intersection of features in the image pairs (after persistence filtering) such that only those features that had remained effectively stationary over the intervening time period were retained. This ANDing operation was implemented by taking the minimum of the corresponding pixel values from each pair of the images. An additional benefit of this comparison of the image pairs was a reduction in the effects of random noise, as any noise signal had to be present in both images to have an effect in the output image.

A complication with this method was that the movement of an object between the two frames was that of its projection into the image plane and was therefore dependent not only on its actual world velocity but also, due to perspective, upon its position in the scene. It was important therefore to look at the relationship between the motion filtering algorithm's characteristics and the world velocity of imaged objects with respect to the system field of view and resolution.

As the most critical concern was not missing pedestrians the most important constraint was that the filter did not remove objects moving at the maximum likely pedestrian velocity. This matter was examined by considering the region of the image where the camera was most sensitive to movement i.e. the nearest point to the camera

thus corresponding to the highest world resolution. The closest a pedestrian would get to the camera would be a function of the difference between the minimum camera height $H_{c\min}$ and the maximum pedestrian height $H_{p\max}$. As the bottom edge of the camera image was aligned so as to image a point on a pedestrian standing vertically below it, the smallest world dimension of a pixel in the image plane, $dL_{\min}$ was calculated from:

$$dL_{\min} = 2.(H_{c\min} - H_{p\max})Tan\left(\frac{b}{r}\right)$$

where r is the working resolution over a field of view covering b degrees.

The larger the proportion of a pixel that a feature moved between the pair of frames to be compared, the weaker the combined signal would be and the greater the filtering effect. The amount of movement was parameterised as p and expressed as a proportion of a pixel. The time between captured frames taken from the frame-grabber was controllable, via hardware registers, to the nearest video frame (1/25) of a second. The delay separating the two images to be compared was accordingly described in multiples of 1/25 of a second by a parameter t_{sf}.

Combining the above, a cut-off velocity v_c was defined as:

$$v_c = \frac{p.dL_{\min}.25}{t_{sf}} \text{ m/s}$$

Expressing this result entirely in terms of measurable and controllable parameters and scaling pedestrian and vehicle velocities to their most usually quoted units of miles per hour gave:

$$V_c = \frac{3600}{1000}.\frac{5}{8}.p.\frac{25}{t_{sf}}.2.(H_{c\min} - H_{p\max})Tan\left(\frac{b}{r}\right) \text{ mph}$$

The choice of the value used for the working resolution parameter in this equation was complicated by the fact that sub-sampled images were being used. The problem was that even though a 64 by 64 image, say, was being processed a single pixel movement at full sensor resolution could result in a whole pixel movement at the reduced resolution. This effective amplification of the pedestrian movement meant that, when using sub-sampled resolution reduction, the calculation had to be based on

the full underlying sensor resolution r_s. A further consideration was that even when sub-sampling, the effect of the background matching operator, which preceded this filtering stage for both images and which operated over a three pixel neighbourhood, was such that the effect of a single pixel change was spread over its nearest neighbours. Considering all the above, a value for the effective resolution of $r_s/3$ pixels was used as the basis for calculation. The figure below demonstrates how, for worst case conditions, v_c is less than 1mph even for the minimum inter-frame interval allowed by the equipment.

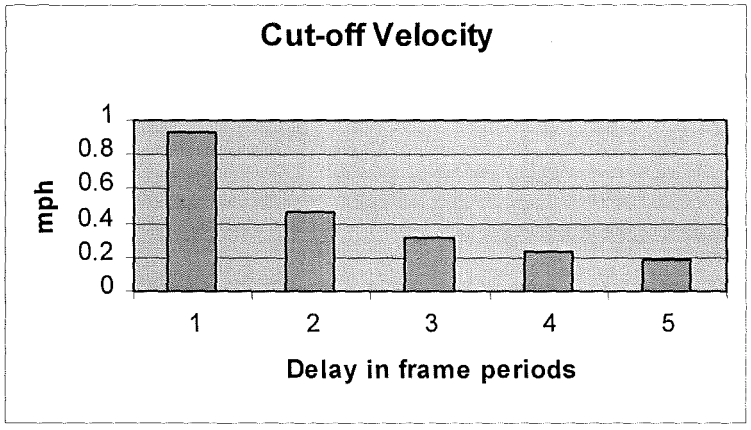


Figure 35: *Variation in cut-off velocity for equipment used in this work* ($p = 0.5$, $H_{c\min} = 3.0m$, $H_{p\max} 2.0m$).

The typical movements of a vehicle shadow through the scene were shown in Chapter 4. It can be seen that some discrimination based on the amount of motion between frames should be a possibility for the vision system. However, the figure also shows a shortcoming of this approach in that some vehicle shadow boundaries, particularly those due to flat vehicle roofs, move parallel to the pavement resulting in zero image velocity. If the camera is also aligned parallel to the pavement then the shadows will move along the vertical axis of the image, such that there is no apparent movement of shadows between the two frames shown. This problem is known as the aperture effect. A separate approach to counter the effect of these shadow boundaries, which appear vertically in the image, is however described in section 6.3.3.1.

In use the motion filter was initially found to be effective in reducing false positive detections from passing vehicle traffic. The main shortcoming was that it was only applicable when the required temporal separation between image pairs was greater

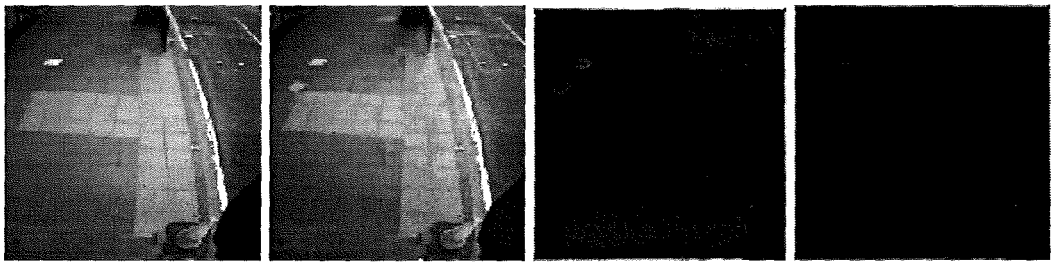
than one video frame period. In general pedestrian velocities were found to be such that this was not the case. Even though waiting pedestrians would often be moving sufficiently slowly, the detector specification did not allow the assumption to be made that fast moving pedestrians were not waiting and so could be ignored.

As the maximum pedestrian velocity had to be preserved and was high enough that there could be significant change between the image pairs it was decided that this method could not be usefully employed at the standard video rates available from the camera mounting used. If at a future point a lower cut-off for pedestrian velocity could be assumed (e.g. to select only those moving slowly enough to be definitely waiting), a plan view could be used or higher frame rates are available, this method would be worth reconsideration. For now however, to avoid the risk of false negative detections due to fast moving pedestrians near the camera, this pre-processing stage was disabled in the final version of the algorithm that was evaluated.

6.3.1.5 Gradient Recombination

The pre-processing stages described so far resulted in a pixel-by-pixel measure of the dissimilarity between the current scene and the long-term background.

Although background features were successfully removed and temporary foreground features correctly retained, the result also contained some unwanted background features. This occurred for two reasons. Firstly the matching operation caused background features to appear in its output when they coincided with areas of the current image, which contained little gradient information e.g. the body areas of pedestrians. This effect resulted in the appearance of false information from underlying background areas and distorted the shape of pedestrian outlines. Secondly, a difficulty with the average based background model was that if it ever wrongly absorbed image features (e.g. due to litter or pedestrians that stayed still for a prolonged period) then they would take a long time to decay away after the causative object had moved. This also produced unwanted false positive signal from the match operator, see Figure 36 below.



(a) (b) (c) (d)

Figure 36: *The current image (a) shows how some litter has recently moved in the wind leaving an unwanted artefact in the background (b). The result of the matching operation (c) erroneously produces signal due to the old position of the litter that is successfully removed by the gradient recombination operation (d).*

To remove these artefacts a recombination stage was devised which combined the output of the background matching stage with a gradient image of the current image (generated using the magnitude of the Sobel edge detection operator). The combination was achieved by ANDing the two representations, by multiplication. This recombination stage was effective in reducing both types of unwanted false positive signal by removing any features that were not still visible in the current, live image.

6.3.2 Structural Modelling and Calibration

6.3.2.1 Introduction

The exploratory work had indicated the value of the top-down application of, even very simple, structural models. This section describes the extension of that work to take account of the variability of pedestrian shape and the relationship between pedestrian appearance in the image and their position in the world co-ordinates of the scene.

Compared to the majority of vision work which had addressed the detection of rigid artificial objects, the problem of defining a pedestrian model for this task was considerably more difficult. Much previous vision work on pedestrian detection had approached the problem of shape variability by using complex shape models based around information on individual limb positions. From the point of view of computational load the large number of degrees of freedom within these more specific models would have made fitting them to image data demanding. In addition it was

hard to see how these methods would cope with some of the more extreme effects of clothing movement, due to inertia and wind buffeting, for which no adequately detailed models existed.

Others have worked with simpler models based on the detection of moving 'blobs' of pixels moving as a group. In such work camera position and distance with respect to the scene have been arranged such that variation of model size and shape with pedestrian world position could be neglected. The 'blob' analysis process was also often aided by the fact that a simplified background scene could be assumed allowing easier segmentation of pedestrians as areas, rather than the outlines to which this work was limited by the pre-processing stages.

In this work it was concluded that the variability in parameters such as pedestrian shape, size, and attire meant that a very general model would be required for it to be valid for all possibilities. This decision was considered justifiable given the objective was to extract information on pedestrian presence and position rather than on their detailed pose. It was also compatible with the shape distortions introduced by the resolution reduction operator chosen to minimise computational load. The use of this simple model also made it faster to compute.

Given the closeness of the camera to the scene it was further decided that a calibration process would be required to relate the changes of the model's appearance in the image plane to changes in scale and viewpoint due to their world position. The first such model was two-dimensional and is described immediately below. The extension from this to a three-dimensional model then follows in section 6.3.2.3

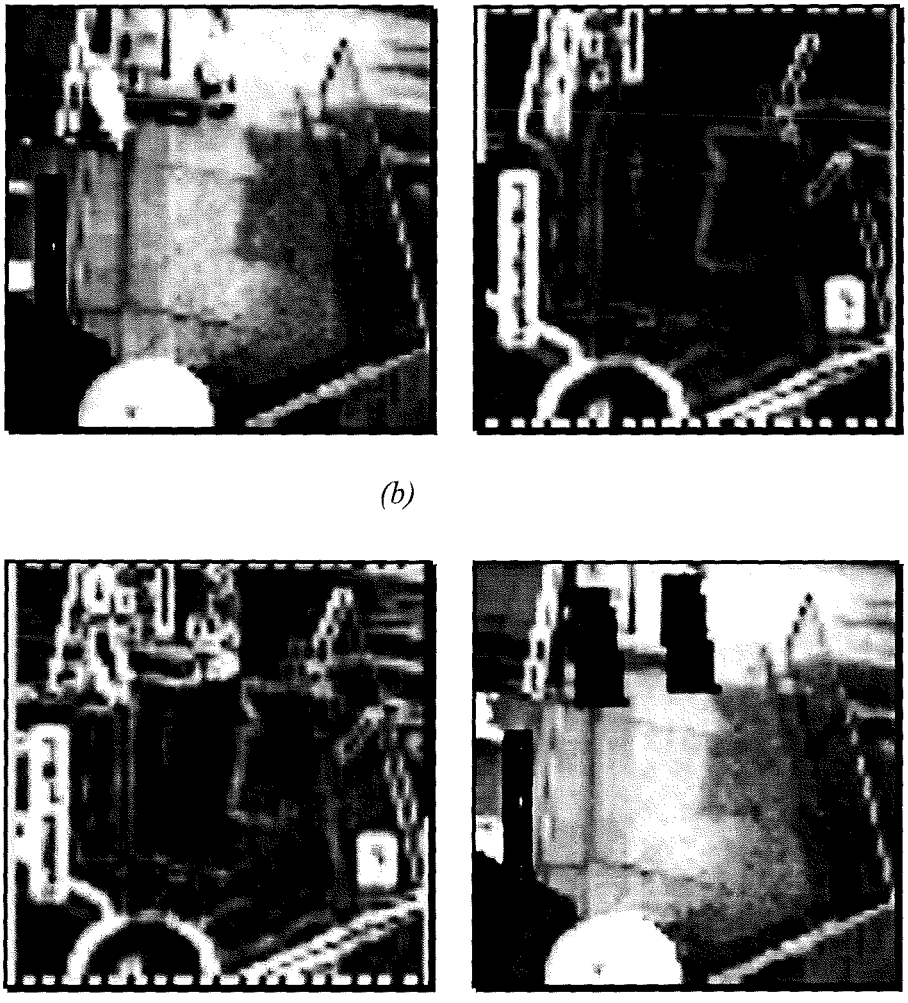
6.3.2.2 Two Dimensional Structural Model

This model consisted of a bounding quadrilateral, which the installer manually specified by moving the vertices to surround an average-sized pedestrian standing in the centre of the detection zone. Although no explicit calibration was performed, this process implicitly incorporated information on the camera parameters and its position and orientation with respect to the scene. The resulting quadrilateral was stored as a bitmap and used as a template area over which to integrate evidence (from the pixel classification stage) of pedestrian presence, see section 6.3.3 below.

The most important assumptions implicit in this method were that the range from camera to pedestrian was as large as possible and that the active detection region

within the field of view was as small as possible. This meant that the variation in a pedestrian's projected shape due to perspective distortion and changing viewpoint with different waiting positions within the scope of the detection zone was minimised. A further assumption was that the orientation of the camera with respect to the pavement was known (i.e. horizontal or vertical) and that pedestrians remained standing essentially vertically, although some bending of the body was often evident.

The evaluation and results of an algorithm based on this model were published in Reading (1996) and a sample of the results obtained is shown in Figure 37.



(a) *(b)*

(c) *(d)*

Figure 37: *Starting from the original scene (a) edges were extracted (b) and compared with those in a frame taken 100ms later and those of a background to identify moving and temporary edges (c). The pedestrian model was then used to accumulate this information and identify the position of individual pedestrians (d).*

Note that the strong central shadow is ignored successfully and the lower pedestrian is detected despite being in front of strong background edge features.

The results obtained were much improved over pixel classification method with the algorithm proving effective at discriminating between pedestrians and their shadows in many circumstances.

The main limitations of this method were that:

- The template was increasingly inappropriate as it was applied further away from the point at which it was originally specified. One way around this would have been to extend the model by interpolating between templates drawn at front and back of zone but this would have been overly reliant on the user and would not have correctly taken account of viewpoint and perspective effects.
- As an installation procedure it was too complex to be practical and too loosely defined, being reliant on a consistent judgement of a typical pedestrian by a human operator.
- There was no means of extending the description to a world co-ordinate based description of the scene activity.

In consideration of the above and bearing in mind the need to operate from lower camera heights than had been used for this evaluation, it was decided that there was a need to explicitly model viewpoint and perspective changes in a three-dimensional model. The quadrilateral bounding model was therefore extended into a three-dimensional bounding box model as described in the next section.

6.3.2.3 Three Dimensional Structural Modelling

Pedestrians were now modelled by a 3D bounding box with a reference point at the centre of the base of the pedestrian, which was assumed to lie on the ground plane. It was also assumed, as with the two-dimensional model, that pedestrians would be vertically orientated in the scene. The nature of the model, world scene and their co-ordinate systems are illustrated in the figure below.

For the pedestrian model to be of use, it was necessary to be able to relate its position in the image plane to a real world position by projecting it into the scene where it could be matched against the image data. Furthermore in order to be able to evaluate a hypothesis about a pedestrian being present at a particular pixel position in the image,

it was also necessary to be able to perform the inverse of this projection process. Accordingly, the necessary mathematical methods to support a three-dimensional modelling of pedestrians and their perspective projection into the scene were developed (see Appendix E).

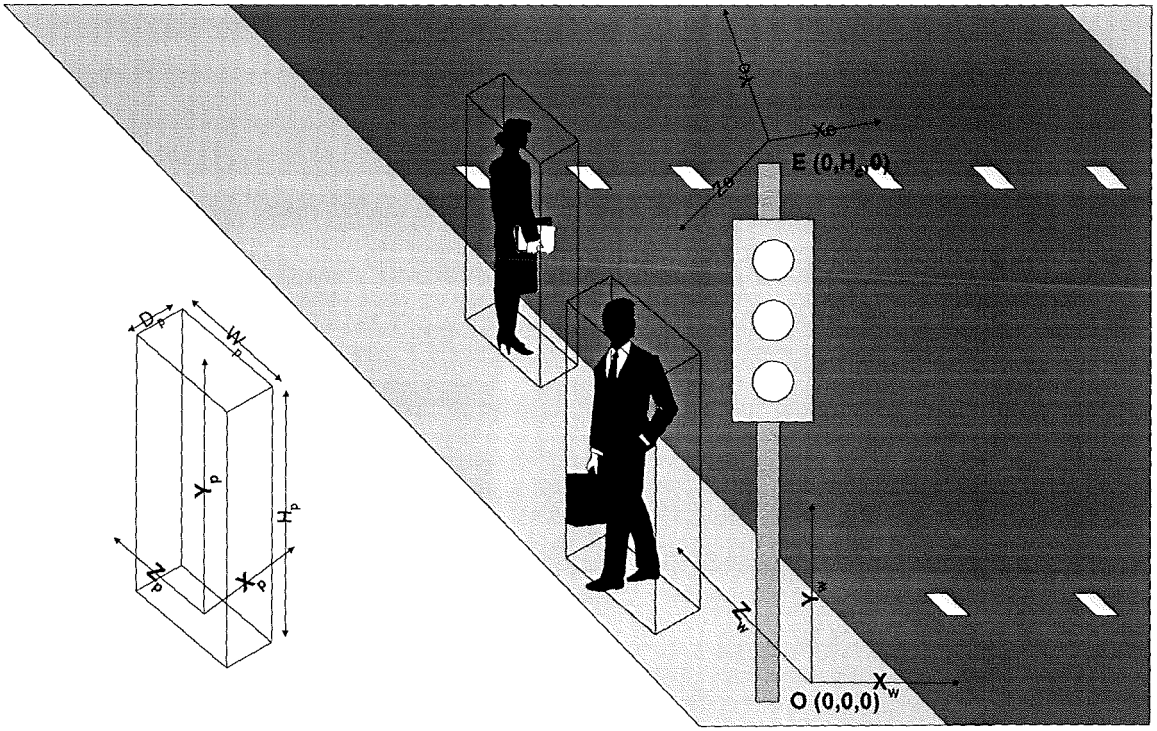


Figure 38: *Three Dimensional Pedestrian Model. The diagram shows the relationship of the world co-ordinate system (defined with respect to an origin at the base of the signal pole, with the z axis along the surface of the pavement) to the camera co-ordinate system defined from an origin at the camera position with the z axis looking down at the centre of the detection zone. Both world and eye co-ordinate systems are left-handed.*

The model was eventually simplified even further to assist with the evidence integration process described in section 6.3.3. This process required scanning over each visible plane of the projected model to assess it with respect to image content. As this part of the code was in the most frequently executed section of the program loop it was seen that considerable computational effort could be saved by reducing the model to a single diagonal plane. In addition, a lack of reliable assumptions regarding pedestrian rotation about their vertical axes meant that in practice the model width and depth were made equal.

The model projection equations and their inverses required the availability of a small set of calibration measurements from the installation process. The process by which the set-up of the camera and its calibration was performed by an operator is briefly described in the next section after which this chapter returns to details of the use of the model in evidence integration stage of the algorithm.

6.3.2.3.1 Calibration for 3D modelling

This section summarises the calibration method more details of which can be found in Appendix B. Given the nature of the application it was considered reasonable to expect some basic measurements to be contributed by the installer, thus obviating the need for a fully automated calibration process. The installation measurements needed however to be clearly defined linear measurements from which a useful level of accuracy could be expected. As is described below the required installer input was reduced to four measurements. First however it is appropriate to describe the degrees of freedom associated with the mechanical mounting of the camera and its alignment.

The camera's mount was designed to permit yaw (i.e. rotation about its vertical, y-axis). It was rotated to point towards a, user-selected, centre point of the detection area that was generally further from the kerbside than base of the signal pole ($D_{mk} > D_{ck}$ in Figure 39 below). Incorporating yaw into the calculation allowed the co-ordinate system to be related to the natural co-ordinate axis choice of the kerbside edge. It was assumed that the kerbside edge was a straight line over the region of the waiting zone - this was valid for all the crossing sites examined. A modification of the calibration procedure would be required to deal with non-linear kerbside edges to define an alternative datum - otherwise the operation of the system should be unaffected.

Camera pitch (i.e. rotation about its x-axis) could vary over a maximum range of 90 degrees between the horizontal and pointing vertically down at the pavement. The signal pole was assumed to be at right angles to the pavement surface, accordingly there was no provision for camera roll (i.e. rotation about its z-axis).

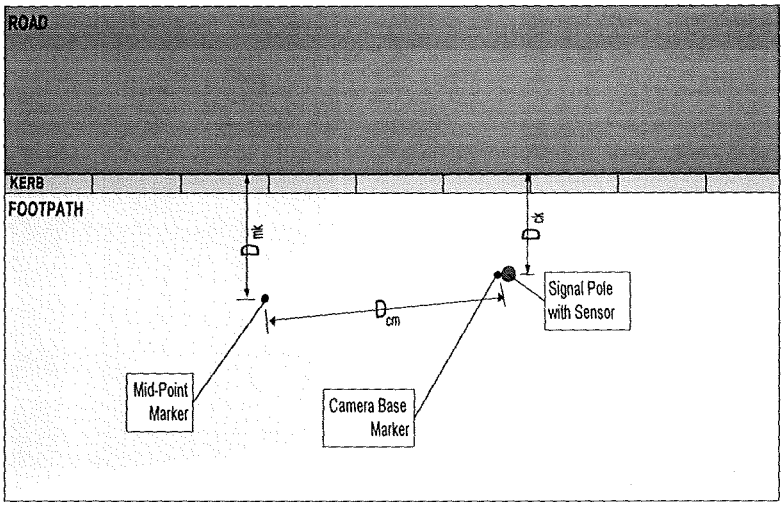


Figure 39: *Plan view of detection area showing installation parameters*

Symbol	Parameter
D_{cm}	Distance between the base point vertically below the camera (near to the base of the signal pole) and the mid-point.
D_{mk}	Distance between the mid-point and the outside edge of the kerb.
H_c	Vertical distance from pavement surface to centre of the camera's face plate.
D_{ck}	Distance between the base point vertically below the camera and the outside edge of kerb.

Table 11: *Installer Measured Parameters*

The camera was attached to the pole such that it was horizontal with respect to the pavement and the height of the camera above street level was measured. The bottom edge of the image was then aligned with a base marker, which had been placed vertically below the camera, by pitching the camera forward about its x-axis. The installer was then asked to identify the mid-point of the image, with the aid of an alignment program (see Appendix B).

A mid-point marker was then placed in the centre of the camera's field-of-view after which the three required measurements D_{cm}, D_{mk} and D_{ck} could then be made with a tape measure on the pavement surface.

To assist the installer in confirming the correctness of the calibration process a 1-metre grid was overlaid onto the live video image, see Appendix B. A facility was then added whereby any errors could be corrected by the user interactively altering calibration parameters and viewing the effect on the 1 metre grid with respect to the scene.

As is the case with all image processing, a method of dealing with boundaries of the finite image had to be found. When processing operators that span several pixels are used there is always a lack of information when they are applied near the image boundaries. For this work, as it had been decided to impose the limitation that a full view of pedestrians would be required for reliable detection, this in turn imposed bounds on the usable field-of-view. Finding the position of these bounds was complicated by the fact that the operators to be applied were based on the projected model and so were of variable size at different points in the scene.

The determination of where the usable bounds should lie was achieved by a search based on the iterative process of inverse projection of a test pixel to find its world position, then projection of the pedestrian model to this position and checking if all vertices fell within the image bounds. The test pixel was initially in a corner of the image and then as the search proceeded was moved diagonally towards the image centre until a projection of the model could be made which fell entirely in the image. This process was repeated starting from each corner of the image.

The result of this process was a set of four vertices defining a bounding quadrilateral for the usable detection area. The vertices were connected to form a trapezoid that was, as would be expected, broader towards the back of the scene (where the model projection is smaller). This limiting trapezoid was displayed to the user for guidance and used to ensure that the specified detection zone stayed within bounds by clipping the detection mask to it. Examples of the resulting quadrilateral and its use to restrict the active detection area that a user could specify can be found in appendix B.

6.3.3 Evidence Integration

The pre-processing stages were essentially a bottom-up process which produced an output that indicated the likelihood of a particular image pixel representing part of a pedestrian in the scene.

This section describes a method that examines this evidence in a top-down fashion, in the light of prior information encapsulated in the pedestrian model, to determine the likelihood that the base of a pedestrian is present at each possible position in the scene/image.

The components of this part of the algorithm are illustrated in more detail in Figure 40 below.

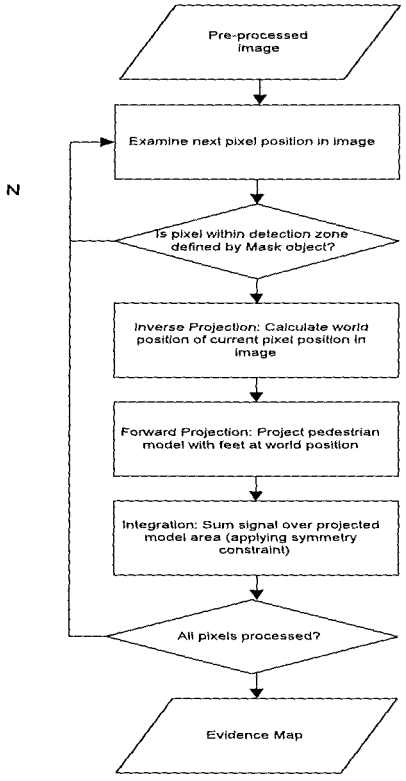


Figure 40 *Evidence Integration*

In this evidence integration stage each pixel within the active detection zone, specified by the user, was examined sequentially. Inverse projection (based on the assumption that feet are positioned on a horizontal ground plane) was used to find the world co-ordinates on the ground plane to which each pixel corresponded. The pedestrian model was then forward projected into the scene at this point to determine the expected boundaries, in the image plane, of a pedestrian standing at this world position.

Initially all the pixels contained within this boundary were scanned and the total strength of pre-processed signal summed. However further experience showed there was a need to discriminate between the strong signals from items such as litter, leaves

and shadows which passed through the pre-processing stage due to their movement. Litter and leaf effects were of small spatial extent but of high contrast and so could often contribute the same integrated signal strength as an entire pedestrian (of low contrast) if a simple integration of match signal was performed. As discussed in Chapter 4 the effects of shadows, after pre-processing, were predominantly due to their boundaries (intensity edge) which resulted in high-contrast lines which were problematic for the same reason.

A method was therefore sought to preferentially select objects with signal distributed over the projected model area and to discriminate against the effects of small high-contrast objects and strong shadow boundaries. The relative contrast of the pedestrian signal to those due to shadows and other distractions meant that threshold-based methods, even those that adapted to local image conditions, were unsuitable as they were likely to lead to the loss of important pedestrian information.

6.3.3.1 Horizontal Symmetry Constraint

The approach devised was to look for symmetry across the model area. This had the advantages that:

- It could be applied without prior binarisation of the image and made no assumptions about the strength of the signal.
- Signal concentrated in a small part of the projected area would be filtered out, as it would be unlikely to have matching features in the symmetrical position.
- Signal due to (most) single lines passing through the area would be filtered out.
- It could be implemented with little computational cost.

An operator was defined to respond to horizontal symmetry (i.e. about the vertical axis). Two implementations of the symmetry test were evaluated including pairing of matching pixels (currently used) and half model-width pairing of horizontal line segments. The combinations were formed by an ANDing operation that was initially implemented by multiplication and subsequently by minimum operator.

The effects of shadows after pre-processing were mainly limited to their boundary intensity edges. Most of these were steeply angled away from the vertical (according to sun and casting object position) and they often occurred singly. It was rare that a pedestrian shadow had similar orientation and size to a genuine pedestrian. The

symmetry operator was effective in filtering out such single, non-vertical lines. However, a weakness discovered with this approach was that it ignored lines that happen to be aligned with the vertical axis of the projected model - as was frequently the case due to the shadows of vehicles passing parallel to the pavement (see Chapter 4). A modification of the symmetry operator was therefore made to eliminate these cases in which a central void was made in the projected area examined effectively imposing a minimum on the width of vertical line that would be detected as self-symmetrical.

A symmetry measure about the horizontal axis was also considered but would have run the risk of missing people if the model height had not matched the actual object height well. The use of the symmetry about a vertical axis was nevertheless considered safe in that it was unlikely to make the detector more unlikely to miss wheel/pushchairs and young people. Indeed pedestrians, wheelchairs and pushchairs tended to have a high level of internal detail in comparison to shadows (where internal detail would be that of the background which is removed during pre-processing) and so produced a stronger response to a pixel pairing symmetry operator. As long as the width of the projected model was greater than any pedestrian encountered then action of the operator caused it to self-centre on the pedestrian's vertical axis.

6.3.3.2 Pedestrian base position constraint

After applying the symmetry operator it was found that the strongest signal could occur in the middle of a small group of, vertically distributed, overlapping (in the image plane) pedestrians. To encourage the algorithm to produce the strongest response for pedestrians at the front of a group (essential to correct estimation of scale and pedestrian volume) a further measure was taken. A row of pixels below the base of the projected model area was examined for lack of signal by forming the average of the complement of the pre-processed signal.

Consideration was also given to applying a similar constraint for left-hand side position as a means of preventing the algorithm selecting the centre of groups of horizontally overlapping pedestrians. However this was not used as it made the horizontal symmetry operator too sensitive to how well the model width matched the width of the object under examination. The same argument applied for not looking at top and right hand sides using similar measures.

6.3.3.3 *Normalisation*

The results of the above-described integrated measures were normalised with respect to the total projected area of the model. As the decision threshold was set globally this was necessary to avoid any bias towards measurements at the front of the image where pedestrians appeared larger due to perspective effects.

6.3.3.4 *Combining constraints*

It was required that both the horizontal symmetry and base position measures were as strong as possible and so the results were ANDed by multiplication (or taking the minimum) to get a combined measure of evidence. Use of multiplication was slower (especially when it is considered that any final implementation may be on a processor without floating point unit) and the result weighted towards stronger signals over weaker signals. The use of the minimum was therefore favoured, as it was quick to calculate and preserved the original signal levels.

6.3.3.5 *Output*

The result of processing thus far is an image map whose intensity is related to the likelihood that a pedestrian is standing with their feet at the corresponding point in the image.

(a) (b) (c)

Figure 41: *The result of the evidence integration process (c) shown with the detected pedestrians in the original image (a) and the match image (b).*

An example is shown in Figure 41 where two peaks in the evidence map (c) can be seen to correspond with the base positions of the pedestrians. A three-dimensional presentation of the map as a surface of height and intensity proportional to the intensity of corresponding pixels in the evidence map is given in Figure 42. This rendering makes it easier to see the shape of the intensity profile forming the expected peak.

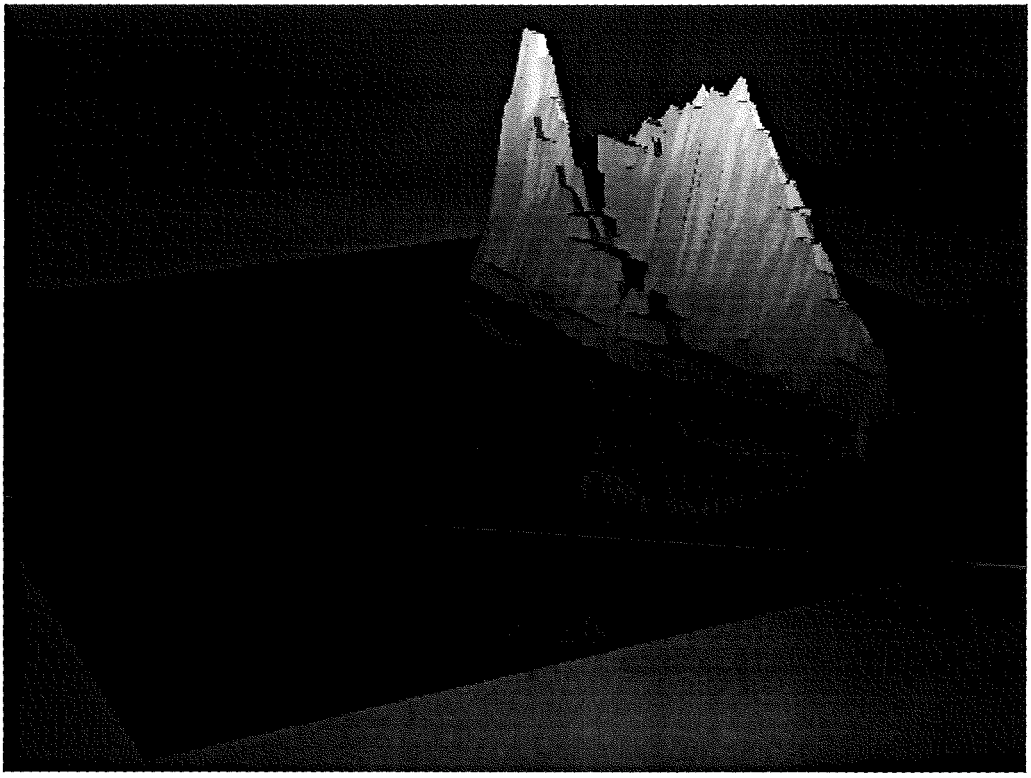


Figure 42: *Rendering of the evidence map of* **Figure 41***(c). The viewpoint is from slightly above the bottom left of the plane of the original image. Height and intensity of the surface correspond to the intensity of pixels in the evidence map. Lighting has been applied form the left to enable the separation between the two peaks to be more easily visualised.*

This map was used as the basis for deciding whether and where pedestrians were positioned in the scene, as is discussed below.

6.3.4 Decision Making

6.3.4.1 Interpretation of evidence.

So long as the fit of the model to pedestrian size was good and in the case that there was no occlusion, then the surface of the evidence map should have comprised a set of unimodal peaks corresponding to each pedestrian present.

Use of this surface to find and count the pedestrians, presented two main problems namely, the choice of decision threshold levels and dealing with the effects of occlusion. An overview of this stage of the algorithm is given in Figure 43, the components of which are discussed in the following sections.

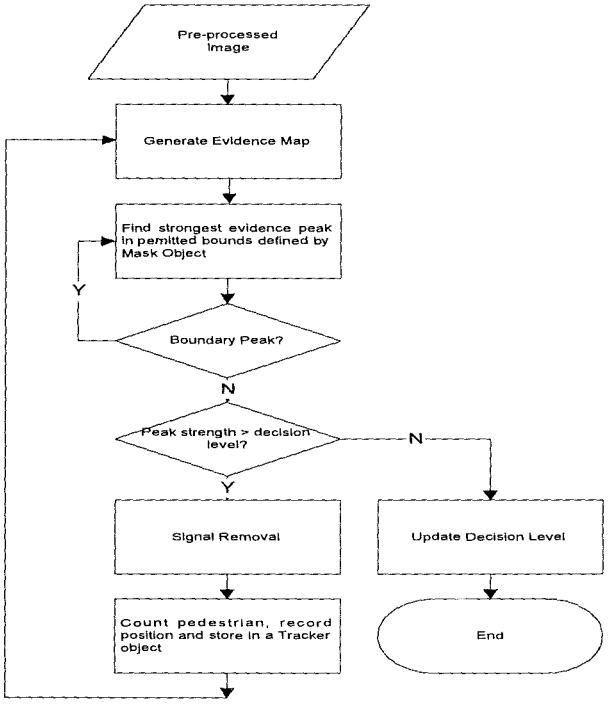
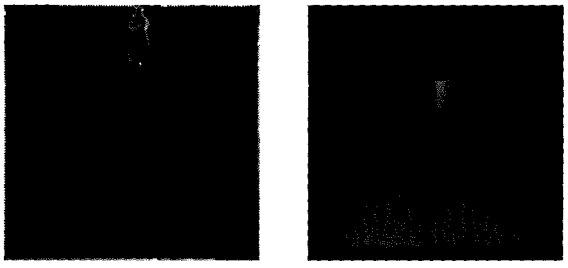


Figure 43: *Decision making process*

6.3.4.2 Boundary Peak Elimination

A first measure found to be necessary was to filter out any peak that was at the boundary of the active detection area. This avoided false positive detections based on partial pedestrians (or other objects) positioned beyond the detection zone, as illustrated in Figure 44.



(a) (b)

Figure 44: *Example of the evidence map (b) for a pedestrian just above the active detection zone (a).*

In these cases the true peak would actually lie outside the detection zone and any detected peak would have been artificially created by the truncation of processing at the boundary of the active detection zone.

6.3.4.3 *Decision threshold choice*

Up to this point the use of decision thresholds had been avoided to minimise the risk of introducing processing stages that might remove information pertaining to low contrast pedestrians. Having reached this point in the algorithm the hope was that the measures already used to process the image signal would have effectively increased the signal to noise ratio to the point that a decision threshold could unambiguously be applied to separate-out pedestrians from other objects and background noise. This section discusses whether the basis for this separation had been achieved, the factors in choosing a threshold level and the means of applying it to find pedestrian positions.

The figure below shows histograms of the maximum integrated signal for a 10,000-frame test sequence (see Chapter 7 on evaluation). Along with a histogram of all the maximum evidence values, there are histograms for just those frames that contained pedestrians, and for those that did not (as judged by the manual reference). If, as hoped, the point had been reached where a simple threshold could be used for decision making then the top histogram would have been split into two parts representing frames with and without pedestrians. The figure shows however that although some separation had been achieved it was insufficient to allow a perfect detection threshold to be identified. In fact it can be seen from the middle histogram that to avoid false negative detections a threshold would have to have been set to detect all peaks over strength 14. Looking at the bottom histogram this would have meant an enormous number of false positives with a detection being registered 90% of the time.

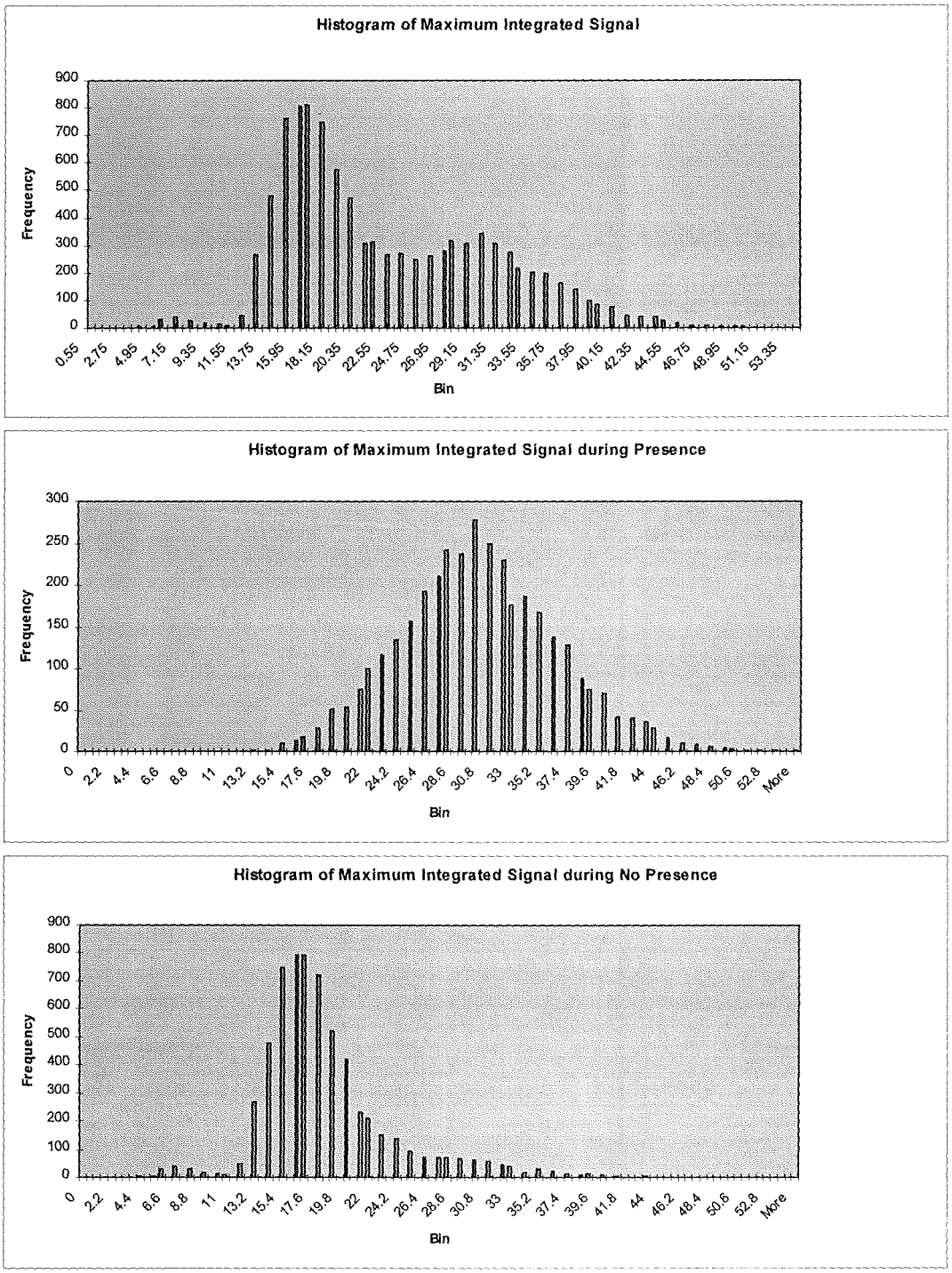


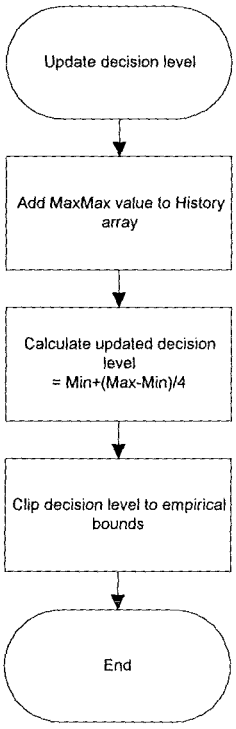
Figure 45: *Distributions of Maximum Integrated Signal from the Evidence Map monitored over a 10,000 frame test sequence.*

The conclusion was that the previous processing stages needed to be improved such that a separation could be achieved that was immune to the effects of all environmental variations that could occur. Note that this would have required the chosen decision value to also be in the corresponding 'threshold' gap for all the test sequences. Rather than pursue this course of action it was decided to investigate the

possibility of using a temporally adapting threshold to achieve the required separation. The following discussion looks at the work undertaken on the design of this adaptive threshold algorithm.

The first adaptive tracking method used was based on the maximum of the long-term average of the peak values occurring at the output of the evaluation phase. The use of long-term average tracking was basically a low-pass filter and therefore its use and time constant made implicit assumptions about the nature of pedestrian activity. If the level of pedestrian activity stayed very high for a long period with respect to the time constant, then the threshold level rose and could start to miss single pedestrians. The time constant of the average was therefore set to be long compared to the maximum time of presence of any individual pedestrian (a value that is related to the cycle time of the crossing). This amounted to an implicit assumption that the waiting area was usually empty. Conversely if pedestrian activity was very low for a long period with respect to the time constant then the threshold level would drift down to the point where it may have started to cause false positive detections due to noise effects of above average amplitude in the maximum. This should be rare due to the threshold being based upon the behaviour of the maximum level.

An improvement was to set the threshold as a function of minimum and maximum values. A proportion (0.25) of the difference between the minimum and maximum values that had occurred in the last n (typically 100) seconds was used as a decision threshold. The advantage of this method was that as the base level of the signal decreased (due say to leaf/litter movement to which the background hadn't yet had time to adapt) the decision level reacted immediately to the decreasing minimum thus avoiding false negative detections. An increase in base level however was not responded to until the n seconds had expired, so false positives could occur. This asymmetry worked in the system's favour as false negatives were not acceptable whereas false positives, whilst inconvenient, were not a safety hazard.



In addition to the adaptive process described above a safety factor was used to improve reliability whereby empirical knowledge was used to place bounds on the range over which the threshold level could vary. This was to ensure that periods of high pedestrian activity wouldn't lead to false negatives when interspersed with weaker signals (a small child of relatively weak signal would be particularly susceptible to this effect) and also to remove false positive responses to background noise-levels, particularly during initialisation.

6.3.4.4 Occlusion and signal removal

To cope with the occlusion problem the strategy used was to extract pedestrian positions sequentially taking the highest peak first. Assuming the peak was high enough to pass the decision threshold (see section 6.3.4.3) the signal in the area of the pre-processed image corresponding to this pedestrian was then removed (i.e. set to zero) and the evidence integration process repeated to produce a new version of the evidence map. The highest peak in the resulting map was then extracted as the next pedestrian. Applying this process iteratively each pedestrian was extracted in turn. The iteration ceased when the highest peak available dropped below the decision threshold at which point all pedestrians were deemed to have been found.

The step of removing signal from the pre-processed images and re-integrating was required to allow the algorithm to find all pedestrians even in the presence of stronger signal due to other pedestrians. Although it may appear simpler and more efficient to have removed signal directly from the evidence surface, this approach caused numerous false positive detections and consequently inaccurate volumetric estimates. This occurred as the energy of the detected pedestrian was spread over an area significantly larger (dependent on the projected area of the model) than that actually occupied by the pedestrian. Simply removing the signal produced discontinuities in this surface that invalidated the operation of the peak detection process, as all points along the discontinuity became peaks. Removal of signal from the pre-processed image followed by the reintegration stage was required to produce a smooth likelihood surface.

6.3.4.5 Conclusion

The result of this decision making process represents the end of the peripheral vision process wherein initial, fast estimates of pedestrian position(s) within the scene had been made. A sample output screen captured from the software in operation is given below (in landscape format for printing clarity). For this work it was now necessary to assess whether these measurements were sufficient to address the tasks of determining pedestrian occupation of the waiting zone. It could however also provide a basis for future work involving the further refinement of knowledge of the scene by the detailed investigation of each pedestrian by a foveal processing stage. This stage could start to relate pedestrians found, to those identified in an earlier image of the sequence and would be the route to extracting detailed behavioural understanding of the scene and its occupants.

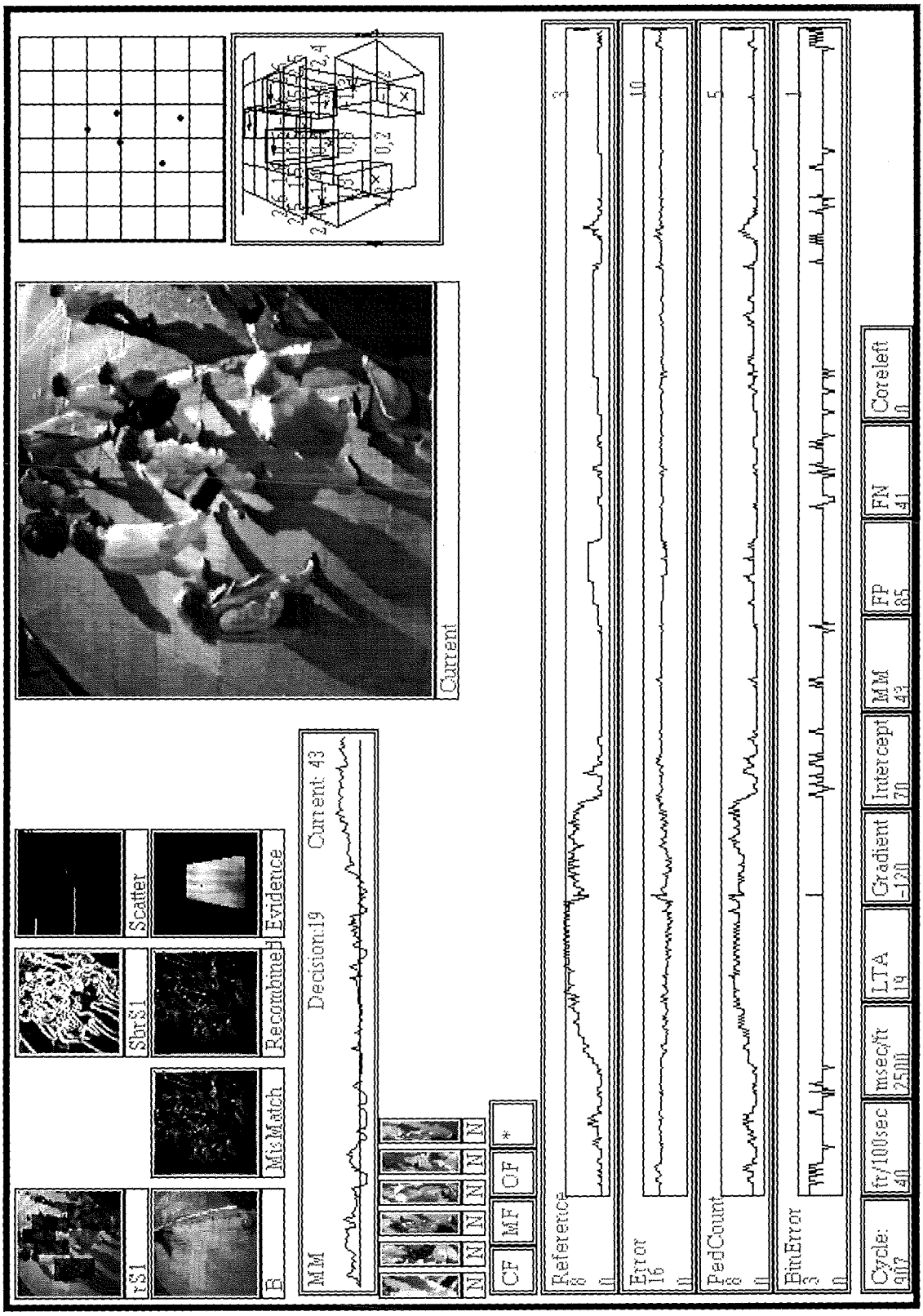


Figure 46: *Output of the pedestrian detection algorithm described in this chapter. A busy scene is shown where the occlusion handling mechanism has limited the volumetric estimate such that it is too low. Underneath the processed images, the peak signal levels are graphed along with the decision making level. Below this is a*

set of small windows, corresponding to tracking software objects, that show each of the extracted pedestrians. On the far right alternative representations of the scene are given based on the three-dimensional information that has been extracted – a plan view and a virtual view. Finally the graphs at the bottom relate to the evaluation process which is described in the next chapter.

6.4 Conclusion

The exploratory work resulted in algorithms that provided a useful level of performance for the pedestrian sensing task but only under a restricted range of conditions. This work was extended resulting in an algorithm which combined bottom-up processing of pixel data with a top-down model based approach to finding pedestrian positions. The level of use of model information was such that the system was not so specific as to miss objects such as wheelchairs and pushchairs yet that it was sufficiently discriminatory to minimise the effects of shadows and other distracting phenomena that might have led to false detections.

Algorithm components were designed so that the processing requirements were within scope of low-cost hardware and so suitable for real-time embedded implementation. In addition hardware image acquisition requirements were kept within the capabilities of commercial CCD cameras and frame-grabbers, which are now available at low cost.

As the above algorithm design progressed and improved in performance, it became apparent that a major barrier to further progress lay in the lack of an objective evaluation system that would allow the benefits of changes in various algorithm stages and/or their parameters to be quantified.

Accordingly the next phase of the work was concerned with the design of an evaluation system to meet these needs. The development of the evaluation system and results obtained are the subject of the next chapter.

7 Evaluation and Results

7.1 Introduction

The central objective of this work was the achievement of reliable pedestrian detection operation over a wide range of operating conditions. During the algorithm development process it became clear that an extensive evaluation system would be necessary not only to gain confidence of reliable on-street operation by the detector but also as means of making comparisons between variations in the analysis algorithm. The complexity of the algorithm meant it had become difficult to judge the effect of each stage (let alone that of each of their embedded parameters) on the final results of processing.

This chapter looks at the methods devised to make these comparative evaluations. As all the video data gathered for these assessments was taken from natural scenes (previous work has used synthetic data generated by CAD systems e.g. Attwood, 1989) the only source of an 'ideal' reference performance was one derived from the opinions of a human operator. Such manual assessment is a labour intensive process and imposes a practical limit on the extent of evaluation that can be performed. The objective in designing the evaluation processes was therefore to allow the comparison of algorithm performance against a manually derived reference performance over test sequences that were as long as possible whilst keeping the amount of human input required within realistic bounds. During the course of this work several methods were used and these are described in more detail in the sections below. The chapter concludes by presenting results obtained by applying the final evaluation method to the algorithm described in the last chapter.

7.2 Evaluation Methods

During the earliest stages of algorithm development, it was adequate to perform assessments, by eye, of performance using short samples (each about 10 minutes long) of video footage representing a range of both typical and difficult situations. At this stage, errors were frequent enough to allow quick feedback into algorithm design.

As performance levels improved longer, busier test samples were required and real-time assessment of performance by a human operator became impractical. An evaluation method was therefore developed in which the algorithm was set to analyse a test video playing on a first VCR whilst the algorithm's output to the computer screen was recorded (via a scan converter) to a second VCR. The resulting recording was then manually analysed, off-line, on a frame-by-frame basis. This method removed the real-time constraint and was used for long-term confirmatory testing (Reading, 1995) but was of limited use in contributing to algorithm development. Its main flaws were the intensive manual input required for each evaluation and that it didn't allow a precise comparison between algorithms to be made. The problem with making comparisons was due to the fact that it was not possible to always synchronise the start of algorithm operation to the same fixed start point on the tape. Even if a means had been developed to make this alignment possible, the variable execution period of each cycle of the algorithm (as the volumetric algorithm involved data-dependent execution times) meant that synchronisation would have quickly drifted such that a meaningful comparison between two algorithms could not have been made.

To address these difficulties a semi-automated evaluation system was produced by working from a time-coded video and comparing a one-off manual analysis, which had been recorded to file as a list of events (i.e. changes in pedestrian volume), against an algorithm's output which was also referenced to the same time-code. Outputs from the computer analysis were then individually assessed against the (temporally) nearest previous event from the manual analysis (Reading, 1996). This method made it easier to home in on situations where an algorithm was having particular difficulty and to identify exactly where they occurred on the test tape. There was however still a significant amount of manual work involved. In particular defining the exact moments of pedestrian arrival and departure and consistently judging boundaries of the detection zone, were found to produce inconsistencies that

required frequent human adjudication. The other problem with this method was that, although an improvement of the previous system, the data was not exactly repeatable due to temporal jitter between the algorithm and reference opinions as well as variations in noise and distortion of the video playback between test runs.

At this point the evaluation requirements were reconsidered in the light of the experience that had been gained using the methods described above. It was identified that there were three key characteristics required of an evaluation system, namely that it should be:

Representative: To be confident of the results of the evaluation, the test data must realistically represent the conditions under which the vision system will have to operate in on-street conditions. It was therefore important that the data be extensive, be taken from actual crossing sites and incorporate the range of variability that occurs in actual operation.

Repeatable: If the effect of small variations in a vision algorithm were to be measured then it was important that they are all assessed under identical conditions, such that two separate evaluations of an algorithm on the test data produce identical results. This is important for two reasons. Firstly it allows accurate assessment of the effects of changes in the algorithm. Secondly, it allows investigation into the root cause of a problem that may have occurred at any stage of a complex algorithm, despite the fact that evaluation is only applied to the end decision.

Automated: When test sets are of significant duration (several hours) it is advantageous for practical purposes that manual involvement in the assessment of performance is minimised. The evaluation methods discussed above, demand a considerable amount of manual input with the operator requiring several times real-time to accurately evaluate the results from videotape. Aside from the inconvenience this process was also found to be prone to errors, presumably due to boredom, by the human assessor. Increasing the degree of automation would also make parametric search of algorithm parameters a possibility and could even provide the performance feedback necessary for the use of adaptive methods such as genetic algorithms to allow the system to self-optimise.

This understanding was the basis for the development of the final evaluation method, described below, which was designed to meet these needs. The means used to obtain

repeatability and automation of analysis are covered in the next section whilst the gathering of suitably representative test video is described in sections 7.4and 7.5.

7.3 Software Development for Digitised Sequence Based Evaluation

An evaluation method based on pre-captured digitised test sequences was developed to address the goals of repeatability and automation. Use of these stored sequences solved the above-mentioned problem of synchronisation and variation in image quality between test runs. It also offered independence from real-time constraints for the evaluation of relatively slow algorithms that could later be speeded-up if their performance was found to justify the effort. This latter aspect was particularly valuable for the later stages of this work, where the algorithm used evidence integration and shape analysis stages that were relatively computationally intensive.

The evaluation method developed can be considered as consisting of the five phases of: sequence capture, manual analysis, format conversion, automated evaluation and review. Software tools developed to support these phases and are described in more detail below.

7.3.1 Sequence Capture

The capture system used an audio time-code dubbed onto the test videos that could then be read into a PC through the parallel port. The parallel port data then provided an absolute reference to the current frame in the video signal.

The sequences were first captured to hard disk, as the CD writing system used required that all data be written in a single session. Once the sequence has been correctly captured to hard disk, it was then permanently written onto a CD.

Software, based on the DOS32 implementation of the MISA ImageStream classes (see Chapter 3), was written to enable the capture of sequences from a user-selected starting point on the tape at a user-selected interval (specified in units of 25ths of a second corresponding to the nearest video frame). A first pass captured as many of the desired images as possible at a rate limited by the hard disk access time. All missed frames were recorded by the program. At the end of this pass, the user was then prompted to rewind the tape such that the missed frames could be acquired. Further iterations were made as required until all frames for a sequence had been successfully captured.

A dedicated 1Gb hard disk was used as the capture buffer. Initial rates of writing images to disk were high enough to sustain writing every 12th image without error. However earlier attempts to capture sequences based on purely software timing had discovered that, as the disk filled-up, the write-rate decreased significantly eventually taking several seconds to write each image. The progressive slowing of write time of an image to disk was so severe that even 2/3 of the way through a sequence capture the write time was up to 6 seconds. This fact made the use of a time-code system and digitisation in multiple passes of the videotape essential. Even using the time-code the large degree of slow down was a considerable inconvenience as several passes were found to be necessary. This problem was diagnosed as being because each image in the sequence was being stored as a separate file. This meant that each image-write required the disk filing system to search the disk directory for space and then to write the image into the space found. As the sequence was captured in multiple passes the filing system became progressively more fragmented with each pass increasing the search time

A further problem with storage as individual files was that each file took more space than necessary, due to the minimum block allocation size of 16kb on the hard disk, thus limiting the length of sequence that could be stored in a given disk space. Once a header had been added each image was 66kb in size thus requiring 80kb of disk space when rounded up to the nearest 16kb boundary and reducing the useful capacity of a given amount of disk space by around 20%.

The solution was to add a new stream type to the MISA ImageStream object to support storage of a sequence in a single large file. This approach eliminated the slow down problems mentioned above, allowing sequences to be captured in only two passes of the videotape. It also used the disk space to an efficiency of the nearest byte. Corresponding advantages were also gained in access speed for the reading process. The only disadvantage was the loss of convenient access to any particular image from a sequence into third party programs. The hard disk was reformatted between capture sessions to avoid any fragmentation of the single large capture file.

The MISA environments ImageStream class was then modified to allow the digitised sequences to be read from disk in a manner identical to that used to access live video input, but with the constraint for real-time execution times removed. To retain the time reference information the associated time-code for each grabbed image was

converted into bytes and stored in the top left corner of each image. Being at the image boundary this had no affect on the computer vision algorithms.

The spatial resolution of image capture was chosen as 256 by 256 pixels from which any lower resolutions could be generated in software. This maximum resolution was a compromise between the number of images that could be fitted onto a CD-ROM (and hence the maximum length of the test sequence) and the highest level of detail that it was estimated might be required by developments of the computer vision algorithms.

The temporal resolution chosen was a compromise between the period of time to be covered by the sequence, given the limit in the total number of images that could be stored in the CD's capacity. Given the specification for the detector was for a 500ms response time, this was chosen as the desired interval between captured images in order to maximise the total period of time covered by the sequence. This precise interval was inconvenient as video was produced at 25 frames a second by the camera and so a capture rate of one frame every 12/25 of a second was actually used.

The result was that each test sequence consisted of 10,000 frames covering 80 minutes of real-time captured in the 640Mb capacity of a CDROM at a rate of one frame every 12/25 of a second. These figures are based on a test sequence occupying a single CD. Obviously, sequences could have spanned several CDs. It was decided however that the time span held on a single CD was sufficient for algorithm test particularly as it had to be balanced against the fact that longer sequences would have required more effort in performing the manual reference analysis (described below).

7.3.2 Manual Analysis

Having captured the evaluation sequences it was necessary to generate an 'ideal' reference performance by having a human operator assess the position of all pedestrians in each frame of each sequence. To this end, a program was written to enable an operator to do this as quickly and accurately as possible.

Using the analysis program, the operator is stepped forward frame-by-frame through the test sequence. As each new frame is analysed, they are shown this current frame alongside the previously analysed frame. They are then sequentially prompted with the position of all known pedestrians, which have been carried over from analysis of the previous frame. In response, the operator uses a mouse-controlled cursor either to indicate the new position of each pedestrian or to click in a 'gone' box if they have

left the scene. When all known pedestrian have been processed, an opportunity is given to mark the positions of arrival of any new pedestrians. Various program features were included to minimise the number of errors, such as the requirement to alternately confirm different actions using keyboard and mouse buttons and the sounding of audible 'beeps' to confirm each operator action. Interruptions in the analysis process were handled by dumping the configuration of the analysis to file such that re-running the program automatically restarted the analysis at the next frame.

The resulting data set contained the time of capture, the total number of pedestrians present and the position of each pedestrian, for all of the video frames. It was stored to a file indexed by the position in the sequence of the frame under analysis. Each pedestrian traced was identified by a unique index number to allow later analysis of the data set on a per pedestrian basis. This file was stored in binary format for compactness, which was also convenient for subsequent use in the evaluation program. An option was also included to allow this file to be converted to text form suitable for import to spreadsheets for human assessment.

The fact that the analysis process associated each pedestrian's position with their position from the previous frame, along with the allocation of a unique identity code, meant their trajectory could be deduced by analysis of the data set. In conjunction with the calibrated three-dimensional model of the scene, this allowed the extraction of behavioural information such as maximum velocity and waiting time. The identity numbers also left open the option of performing a later expansion of the manual data set by prompting an operator to enter information on, for example, the height, width, sex, age and behaviour (i.e. waiting, passing-by or crossing) of each individual.

In performing this manual analysis, it was necessary to decide how to deal with certain categories of object. The following policies were followed:

- Pushchairs and wheelchairs: Although they are always attached to an adult, pushchairs need to be counted as separate entities as they may be located within the detection zone when the accompanying adult is not. Wheelchairs clearly needed to be counted but presented the problem of defining a reference point. For both pushchairs and wheelchairs the reference point was chosen as an estimate of the centre base of the chair.

- Umbrellas: It was often impossible to be exactly sure of where a pedestrian is positioned under their umbrella. Under these circumstances, a best guess had to be made as to pedestrian foot position.
- Bicycles: Bicycles were treated in the same way as pushchairs as a separate entity (when being pushed) with the reference point chosen beneath the pedals.

Using the above-described tools the manual effort involved was about five man-days to analyse each sequence of 10,000 frames. Obviously, this varied significantly according to the level of pedestrian flow and the consequent number of pedestrian positions to be registered.

7.3.3 Format Conversion

The manual analysis data generated, as described in the last section, was indexed by the frame of the test sequence in which it occurred. A conversion program was written to reformat this data from a frame-based to a pedestrian-based representation indexed by the pedestrian identity numbers. Each entry in the converted output gave the number of frames that the pedestrian was present and their position in each frame. This reformatting of the data provided a better basis from which information on aspects of pedestrian behaviour such as velocity of movement and time of presence could be extracted.

7.3.4 Automated Evaluation

Having obtained the manual reference data set (generated according to the last section) this could now be automatically compared with an algorithmic analysis under program control. There was however a decision to be made as to how to best compare the manual reference data to the opinions generated by the algorithm. One possibility would have been to use the algorithm to generate a frame-by-frame output file which could then have been compared off-line with the reference data file using a spreadsheet. It was, however, decided for the following reasons, that a better alternative would be to perform the comparison as the algorithm executed:

- The manual data required spatial filtering to eliminate pedestrians outside the particular detection zone active during a test.
- Tolerance on the moment of entry/exit from the active detection zone (i.e. implementation of the 'may detect' area) could be programmed by applying a

dilation operator to expand the bitmap that was used to store the definition of the zone below.

- A live display with audible warning of error events could be given as the algorithm ran and so quickly alert the operator to events of interest.
- Conversion of image co-ordinates into 3D world co-ordinates could be conveniently applied on the fly using the 3D object model which was already a part of the algorithm program.

The algorithm program was therefore extended to accept the manual reference file as input and to make performance comparisons as it ran (see bottom of Figure 46). An output text file was generated tabulating the opinions of the manual reference against those of the algorithm to provide a permanent record. This file was formatted such that it could be read into a standard spreadsheet package and its tools used for presenting the results. Once in the spreadsheet, areas of the sequence which were problematic could be identified for more detailed assessment using the review program.

7.3.5 Review

A review program was written as a tool to allow the user to view captured sequences in a manner analogous to the use of a video cassette recorder. A sequence can be stepped through in either direction at variable speed. The embedded time code (see section on capture above) is extracted from the image data and displayed to the user. This program was used for the identification and examination of interesting events such as those presented in Chapter 4, for checking the reference data generated by the manual analysis, for investigating the conditions leading to any detection errors and for checking captured sequences for errors. If capture errors were found, due to glitches in tape playback or the time-code equipment, the exact frames in error could be identified by the time-code and recaptured.

7.4 Evaluation Test Sites

As the system under development needed to work at any pedestrian crossing, it was considered important not to dedicate too much work to a single environment, as there was a danger of inadvertently incorporating site specific assumptions into algorithm structure. Site specific factors include pavement colouration and texturing (amount of

background edge information), compass direction (hence shadow direction and angle), pavement slope, levels of pedestrian flow and available detector mounting positions. A practical compromise was to use three outdoor test-sites along with a laboratory set-up for tests under controlled conditions and to confirm calibration. Two sides of a north-south oriented crossing in Bracknell were used in conjunction with a Department of Transport trial. A camera on the west side was positioned facing north and one on the east faced south. Flows of pedestrians were medium to low but included many examples of pushchairs and children. An east-west oriented site was set-up in Princess Street, Edinburgh with the assistance of the City of Edinburgh Council. It allowed conditions of heavy vehicle and pedestrian flow to be captured, and being local was used to look at night-time operation. More detail on the test sites can be found in Appendix D.

7.5 Evaluation Data Set

From the outdoor test sites a total of 60 hours of video footage was gathered. This included a period of three consecutive days during which 1.5-hour recordings were taken at 3-hour intervals between 6am and 11pm from the Princes Street site. Video to represent variation in weather, pedestrian and traffic flow was taken opportunistically when suitable conditions arose.

From this video a set of sequences were chosen for digitisation so as to cover the most important conditions and transitions between them (e.g. from sunshine to overcast, light to dark. This resulted in a set of five CD-ROMs (labelled as BC4_1, BC5_1, WEPS4_1, WEPS6_1 and WEPS10_1) each containing an 80-minute test sequence. The characteristics of these sequences are given in Table 12 below.

Sequence Code	Total Peds	Time of Capture	Environmental Conditions
BC4_1	1497	11:30am	Objects: Pushchairs Weather/Lighting: Mainly bright sunlight. A large building shadow traverses the scene. Interactions between the camera's automatic exposure control, and varying lighting and image content are evident. Light wind.
BC5_1	1035	11:20am	Objects: Leaves, Pushchairs, Umbrellas, Wheelchairs, Bikes. Weather/Lighting: Initially raining and overcast. Transition to bright sunlight and strong shadows occasionally interrupted by passing cloud cover. Wet leaves on pavement dry out and blow about in strong wind.
WEPS4_1	1838	8:00am	Objects: Vehicle shadows on pavement Weather/Lighting: Initially overcast with transition to bright sunlight. Strong shadows from pedestrians and passing vehicles. Some litter. Dry throughout.
WEPS6_1	1334	6:00pm	Objects: Examples of pedestrian illuminated by headlights. Weather/Lighting: Dusk period. Illumination initially solar then transition to street lighting, shop display lighting and car headlights. Moving highlights cast over waiting area by passing vehicle headlights
WEPS10_1	113	6:00pm	Objects: Pedestrian flow and contrast are very low. Weather/Lighting: Night-time. Illumination from street lighting, shop lights and car headlights. Initially dry then transition to heavy rain. Specular reflections.

Table 12: *Summary of digitised test sequences and their characteristics*

It was considered that the above test sequences, whilst not being completely comprehensive, represented a sufficiently wide range and testing set of conditions to avoid the adoption of an overly simplistic solution (e.g. one that operates successfully only in overcast conditions). Conditions which made the vision task simpler e.g. overcast conditions with no variation in illumination, no shadows and no distracting

objects such as litter and pushchairs were not represented within the test data. Indeed all sequences were chosen precisely because of the difficulties they introduced.

7.6 Results

The results given in this section illustrate the use of the evaluation system to assess detector performance using the five test sequences described above. In total, this amounted to an evaluation based on 50,000 images sampled every 12/25^{ths} of a second over a total period of 6 hours 40 minutes.

7.6.1 Binary Detection Results

In evaluating binary detection performance the safety of users of the crossing must be of primary importance. A pedestrian ignored by the crossing detector is liable to become frustrated with increasing delay and is more likely to cross when it is unsafe. Pedestrian safety is therefore most likely to relate to the number of false negative detections (FNs), where a waiting pedestrian is missed by the binary detector. For efficiency of vehicle flow it is also important to minimise the number of false positive detections (FPs) where a presence signal is given in the absence of waiting pedestrians thus leading to unnecessary stoppage of traffic.

During the evaluation a practical problem was the range of variation in manual judgement of pedestrian position. This caused detection errors to be flagged frequently when pedestrians were waiting around the boundaries of the active detection zone. To counter this effect a 'may detect' area was introduced in line with the approach taken in the Department of Transport specification for Puffin pedestrian detectors (DOT, 1997). This area around the borders of the detection zone allowed pedestrians just outside the zone to be counted as being inside for the purposes of false negative judgements and conversely, those just inside to be eliminated from causing false negatives. This region was given a value sufficient to allow for the likely magnitude of errors in pedestrian position judgement at the front of the scene.

A possible concern was that as the evaluation is frame-based, some false negatives might fail to be picked up because a different pedestrian was detected in compensation. However, although this may have happened on occasion, the extensive test sets meant that such a weakness in an algorithm was unlikely to remain undetected.

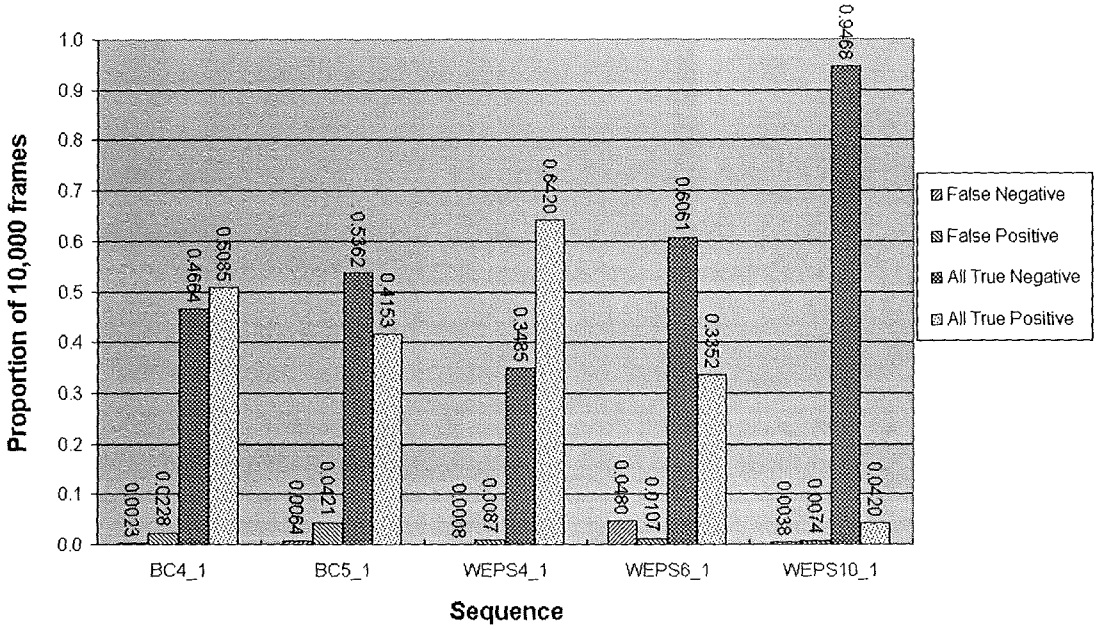


Figure 47: *Binary detection performance of the algorithm on a frame-by-frame analysis of five test sequences. The algorithm's output is classified according to whether it is correct (True/False) and whether there are pedestrians present (Positive/Negative) as a proportion of the 10,000 frames analysed.*

Figure 47 shows the results obtained from analysis of each of the test sequences using the 'may detect' allowance. As expected there are more false positives than false negatives. This is a consequence of elements in the algorithm design aimed at the avoidance of the more safety critical factor. A further point of interest is the relatively high number of FNs for sequence WEPS6_1. Investigation showed this to be a consequence of decision threshold choice being adversely affected by the combination of low-contrast conditions (night-time) coupled with occasional very strong signals when a pedestrian is illuminated by the headlights of a passing car. Examination of the false positives showed that they occurred in a few short bursts that were due to the time-delay in the algorithm adapting to a change in conditions of the background scene.

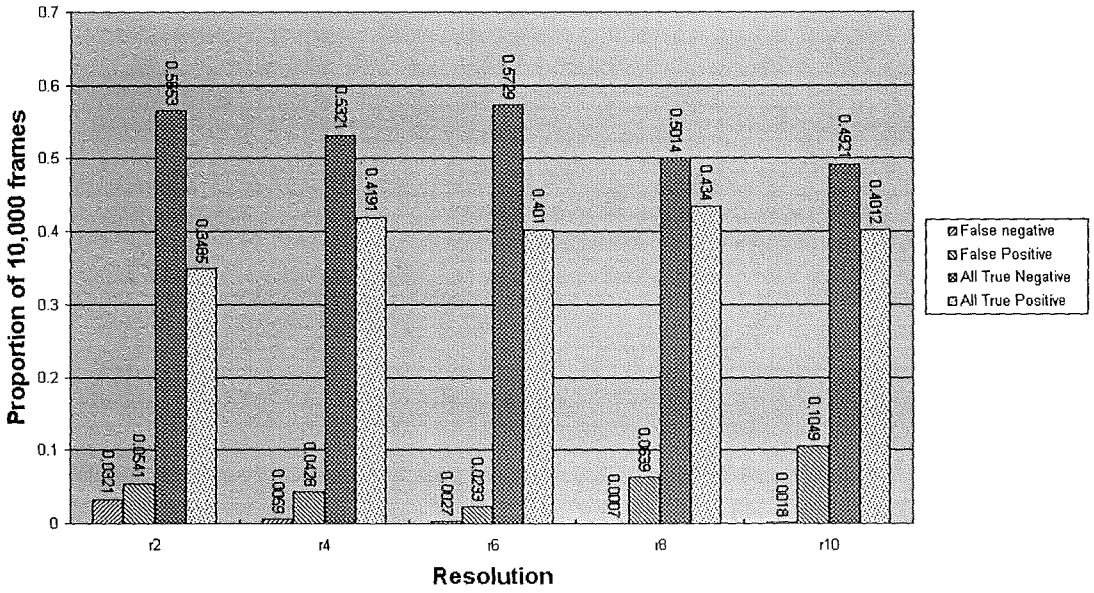


Figure 48: *Binary detection performance of the algorithm as a function of operating resolution on analysis of sequence BC5_1. The algorithm's output is classified according to whether it is correct (True/False) and whether there are pedestrians present (Positive/Negative). Resolution is expressed as a fraction of 256 pixels (e.g. r4 =256/4).*

The detection algorithm was designed to operate on a range of resolutions to allow a compromise whereby operating resolution was traded off against response time when operating on less powerful hardware. This feature was used to demonstrate the use of the evaluation system to perform a parametric search examining the relationship between resolution and detection performance. The results can be seen in Figure 48. The resolutions indicated are of the form r*n* where *n* is an integer factor by which the resolution of the test sequence images (256 by 256 pixels) had been reduced. It can be seen that a minimum in the number of FPs occurred at resolution r6, and in FNs at resolution r8. It is believed that the reason for this is that pedestrian shape becomes simpler when examined at lower resolutions thus decreasing the effect of factors such as moving arms, clothing and the forward lean of a pedestrian's body whilst walking.

It should be noted that the errors in the above results represent a worst case being based on a frame-by-frame analysis. The error rates shown will only rarely result in FNs when considered on a pedestrian-by-pedestrian basis given that failsafe protection against single frame errors is easily added at the output stage.

7.6.2 Volumetric Detection Results

Figure 49 shows a graph of reference volume against the average opinion of the detection algorithm. It is apparent that there is a simple, but non-linear, relationship between measured and actual volume, which is probably a consequence of pedestrians occluding each other. The shape of the curve however indicates that a calibration curve could correct for this effect - a quadratic fit is shown in the figure

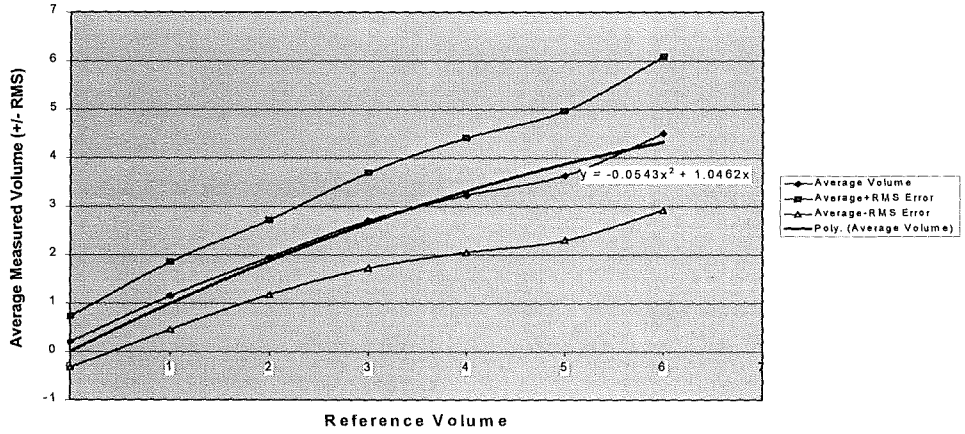


Figure 49: *Graph of Actual Pedestrian Volume (from reference data) against Average Volume output from the detection algorithms. The upper and lower traces show the RMS deviation of the measured values from the reference.*

The parameters of the volume curve vary between test sites due to differences in viewing height and angle, which have the effect of varying the degree of occlusion. The spread in measurements shown is largely a consequence of the 'may detect' problem mentioned earlier.

7.7 Independent Evaluation

The performance of the pedestrian detection algorithms has also been independently assessed by traffic consultants acting on behalf of the Department of Transport's Highways' Agency. Their project aimed to look at the viability of Volumetric Puffin crossings and so required appropriate detectors.

Assessment took place in three phases. An initial laboratory demonstration (from videotape) led to an invitation to participate in controlled outdoor trials. Success at these trials meant that an embedded system implementation of the algorithms was purchased from the University for long-term trials (over 2 months) at a test site in Bracknell, Berkshire.

The assessment was competitive in that manufacturers/universities were invited to put forward detection equipment for examination by the consultants. Around fourteen detectors were proposed, of which six were invited to the controlled outdoor trial and of these only two were purchased for the long-term trial.

After assessment of the two systems, the MoniPed system was chosen for live connection to the traffic signal controller during the study of Volumetric Puffin operation carried out by the Transport Research Laboratory. The results of the TRL study were published in (Crabtree 1997) and are somewhat inconclusive as to the value of volumetric detection, but mention no problems with the performance of the detector. Indeed the system was accidentally left operating the crossing for a period of one month after the trial had been completed with no complaints being received from the public – the usual method of feedback on equipment performance for local councils.

7.8 Conclusions

This chapter has described the requirements and development of an evaluation system for the pedestrian detection algorithm. Although various evaluation methods working directly from videotape were sufficient during the initial stages of the work, they were eventually found to be limiting particularly in terms of the degree of manual input required. Further consideration of the evaluation task led to the identification of three key evaluation requirements for an evaluation system, which were that it be representative, repeatable and automated.

An off-line evaluation method based on digitised stored sequences was developed to meet the need for repeatability and automation. In order to ensure the test data was as representative as possible of the actual range of conditions under which the sensor would have to operate, it was desirable to test on as much video as possible. This however had to be balanced against the degree of manual labour required. In compromise a test set, consisting of five, 80 minute test sequences was selected, from a total of 60 hours of video, to cover the most important conditions.

The main problem encountered with the automation of the evaluation process, was the sharp definition of the boundary of the active detection area specified by the user. The fact that pedestrians occupied many pixels in the image meant it was fairly frequently the case that the manual data put a pedestrian just within the detection zone whilst the

algorithm put them just outside. This scenario produced a false negative detection error. Similarly false positive detection errors were produced when the situation was reversed. This problem was particularly prevalent when a pedestrian was walking at speed making consistent manual judgement of the foot position difficult. Methods were developed to filter out these cases automatically when they only occurred for a single detection cycle. Any remaining errors reported by the evaluation usually required further investigation and were generally found to be due to genuine weaknesses of the detection algorithms.

The real test of the above-described work, particularly from the point of view of quality of representation, will be to have the results confirmed by extensive on-street testing. So far as this is concerned the results of the detector's participation in the Department of Transport's independent trial process are encouraging.

It will be noted that all the laboratory-based evaluation described was derived from video recordings rather than directly from a camera. This was found necessary to enable repeated test under the same conditions. Compared to direct operation from a camera the effect of video recording was the degradation of signal quality (with additional noise) and the occasional capture of incorrectly framed video images due to loss of tracking. To minimise the degradation wherever possible video capture was performed using 8mm video recording and playback equipment in favour of the standard VHS system that offered poorer quality. It was assumed however, that if algorithms could be designed to operate successfully on this degraded image information that they would also operate successfully on live input direct from a camera. The independent trials described above, which were performed with the system operating directly off a camera, would seem to confirm that this was the case.

Results were given demonstrating the use of the evaluation system over the digitised test data set. They show how relatively rare error conditions can be identified even when these are as few as 8 in 10,000 cycles. The repeatability of the system allowed the causes of these events to be investigated. The benefits of an automated system were demonstrated by results from a parametric test of operating resolution. These results indicated that operation at lower resolutions can be better than those at high levels of detail and that an optimum operating resolution could be identified. The results further suggest that despite the well-known high processing demands of implementing vision systems, a low-cost solution to the pedestrian detection task

should be practical. Although the results obtained were promising there are still several areas that require further work. Most important are the better understanding of the effect of the 'may'detect zone on evaluation accuracy and the introduction of methods to detect items such as umbrellas that do not fit the shape model (and associated assumptions) well.

8 Conclusions and Future Work

8.1 Introduction

This work set out to determine whether the detection, by computer vision, of pedestrians waiting at road crossings could be made sufficiently reliable and affordable, using currently available technology, to be suitable for widespread use in traffic control systems.

It has resulted in the development of a vision system, as presented in this thesis, which has been shown to attain a useful level of performance under a wide range of environmental and transportation conditions. This was achieved, in real-time, using low-cost processing and sensor components so as to demonstrate the viability of developing the results of this work into a practical detector. The quality of the resulting system can be judged from the fact that it was selected by the Department of Transport for on-street evaluation and that industrial partners and the Teaching Company Directorate adopted it as the basis for a Teaching Company Scheme program to develop the research system into a commercial product.

It should be noted however that a significant amount of work will still be required to take the work that has been completed so far to the point that its performance and reliability are sufficient for permanent on-street installation. Aside from the need to carry out evaluation under an even wider range of environmental conditions, the performance in low light conditions will need attention and the pedestrian model used will need to be extended/supplemented to cope with umbrellas. These problems should not be insurmountable though, and their solution should be significantly accelerated by the development and evaluation tools produced during the course of this work.

The next section reviews the content of the chapters of this thesis in more detail with the goal of identifying the most important facts that have been learnt concerning the nature of the pedestrian detection task and what has been achieved in this work aimed at its solution. This is followed by a consideration of the future direction of the work and closing comments on what has been achieved.

8.2 Review of Chapter Contents

An examination of the trends in traffic control highlighted the increasing demand for more sophisticated detectors to improve the level of service to traffic in terms of safety and flow rates. It was also seen that the historical emphasis on the optimisation of vehicle rather than pedestrian movement is now being redressed by government initiatives encouraging pedestrian-responsive road crossings. A critical component of these new pedestrian-responsive crossings is the inclusion of pedestrian detectors to provide the basis for decision making by the crossing's control algorithm. It was also observed that factors such as the variability in pedestrian behaviour and shape make pedestrian detection significantly more difficult than vehicle detection. This was borne out by a review of the literature in this area which indicated the limitations of the detection technologies that had previously been applied to detecting pedestrians. Results obtained from monitoring the performance of the best currently available detector type at a trial crossing were also reported. The problems found with these detectors combined with the anticipated future need for more flexible detection features provided the justification for the remainder of this work to investigate the use of computer vision to perform the pedestrian detection task. It was decided to limit the detection objectives for this research to the determination of pedestrian presence and volume, in line with the requirements of the Department of Transport's Puffin and Volumetric Puffin crossing types. It was also decided that the practical aspects of a detector necessary to allow its installation and alignment for on-street trials would be addressed in this work but that operational features such as the automatic monitoring of failure conditions would not.

The requisite equipment and software were then selected to permit the development and on-street trial of computer vision algorithms. An important part of this was the consideration of technological and economic constraints on system component choice that identified limitations in the performance of the image sensor and in available

processing power as important issues. The sensor's limitations were consequences of its dynamic range being too narrow to cope with the required range of illumination (from sunlight to street-lighting) and the global, uncontrolled nature of its exposure control mechanism. Awareness of the constraints on available processing power became an important guiding factor in the design of detection algorithms.

Throughout the completion of this work a considerable amount of equipment and especially software was developed. Most significantly a software development environment, the Multi-resolution Image Sequence Analyser (MISA), was programmed in object oriented C++. The flexibility of the design of MISA as a set of software objects that could transparently handle image manipulation, capture and storage proved invaluable throughout all of this work.

In the analysis chapter the objective was to obtain a better understanding of the nature of the vision task by considering the characteristics of the crossing infrastructure, environmental variation and pedestrians. On the basis of video footage from real-world test sites the problems likely to face a vision system were identified as a guide for the later assessment of the relevant prior art from the literature and for algorithm development.

It was seen that the camera being mounted on a traffic signal pole and looking across the monitored area meant that there was significant variation in viewpoint and scale of pedestrians throughout the camera's field of view and also that occlusion of pedestrians, by each other, was commonplace. Looking at previous work in computer vision it was found that these problems had generally been avoided by positioning the camera at a distance from the pedestrians and/or observing them from a plan viewpoint.

The most significant effects of the environment were a consequence of distracting objects in the background and illumination affecting image formation. Leaves and litter were identified as a particular difficulty as their high contrast, high density and occasional motion could resemble the characteristics of pedestrians in the image. No mention of these difficulties was found in previous comparable work directed at outdoor computer vision. A further source of high contrast moving edges (intensity gradients) was the sun throwing shadows off pedestrians, vehicles and buildings. It was evident for this task that shadows would have to be considered as having two

components - perimeter and body - which presented very different problems to vision algorithms. Most previous work concerned with shadows addressed the problem of achieving invariance to the uniform darkening within a shadow body. However, no previous work was found which acknowledged that, especially when large numbers of relatively small objects are moving freely around a scene, it is the shadow boundaries that present most of the problems to interpretation of the image. As most algorithms described in the literature were intentionally sensitive to such intensity gradients, to help eliminate the effects of brightness changes caused by variation in illumination level, it is likely that they would have had difficulty operating under these conditions

The main difficulties associated with the nature of pedestrian behaviour were that there could be any number of them present in the scene at any time and that each might be either static or moving. Modelling of pedestrians' form was made difficult by their deformable shape, use of wheelchairs, use of carried items such as bags and variation in factors such as their attire and size. In seeking to tolerate this range of variation this work has dealt with the task of detecting pedestrians with less restriction on their attire and behaviour than any previous study.

The conclusion therefore of the review of relevant prior art in computer vision for pedestrian detection was that previous work was inapplicable due to its adoption of constraints that were too restrictive for detection at road crossings.

An algorithm was developed to cope with these difficulties whilst minimising the need for processing power. Inspired by the approaches used in reviewed work it consisted of a pre-processing stage to remove relatively permanent background features from the image followed by a spatial modelling stage to identify pedestrian positions.

To reduce computation, resolution reduction methods were applied to reduce volume of data in the incoming video stream and a means for the integer calculation of a recursive-average based representation of the background scene was developed. A matching operation was derived to give sensitivity to edge direction at minimal computational load. The pedestrian spatial model needed to be sufficiently generic to cope with high variability of pedestrian form. The final version utilised the projection of a three-dimensional 'box' model into the scene to cope with variation in viewpoint and perspective. The model was reduced to four vertices to allow a good

approximation to the variation in pedestrian scale and viewpoint to be modelled whilst avoiding heavy computational load. To provide the required discrimination between the effects of pedestrians and of moving edge features due to moving shadows and distracting objects constraints were imposed during the process of matching the model to the pre-processed data. These included measures of symmetry about the vertical axis and filters to eliminate single lines that might be due to vehicle shadows.

A further distinctive feature of the algorithm was that the use of empirical thresholds was avoided so as to minimise the likelihood of failing to detect pedestrians of low contrast. This avoidance of thresholds was utilised during both the pre-processing and modelling stages to avoid unreliable assumptions about the relative contrast of object and background features. Consequently in situations where thresholds would have been applied in previous work the problem was recast as a test of to what degree a set of two of more conditions could be said to be true. The test was then implemented without binarisation of the input data by means of an ANDing operation. This approach was applied in several contexts namely the velocity filtering method, the recombination method and during the symmetry measurements of the evidence integration stage.

To support an independent evaluation of the detection system by the Department of Transport a user interface was developed along with a calibration process to allow the relationship between the camera image and the 3-D scene to be established. This calibration process was designed to require only a simple set of linear measurements so as to be suitable for on-street implementation. A test site was established in Princes Street, Edinburgh with the co-operation of the City of Edinburgh Council. Along with the test site used by the Department of Transport trial this provided the video data upon which the evaluations were based.

Once the algorithm had attained a useful level of performance it was found necessary to develop an evaluation system. It was realised that to meet the objectives of this work it would be essential to find a way of repeatably evaluating algorithm performance. Accordingly a considerable amount of the effort was spent on the development of tools to support the evaluation process. A repeatable evaluation system based on the automatic comparison of stored digitised image sequences against a one-off manual reference opinion was developed.

The result of using the evaluation system was that repeatable tests on extended test video sequences could be applied. The use of this facility was demonstrated and results given for performance variation over a set of sequences and also over a range of operating resolutions. Various practical issues in making the automated comparison of algorithm generated opinions against reference opinions from a human operator were also addressed. The ability to examine the results off-line to home in on problem areas and the ability to repeatably return to these situations to investigate their causes without the constraint of real-time operation meant that difficulties with various stages of the algorithm could be identified. The other important aspect of the evaluation was to ensure that the test data was sufficiently representative of the range of real-world conditions that a detector would encounter so as to be sufficient to give confidence that algorithms tested on them would be useful on the street. The volume of manual labour required to prepare these test sequences was considerable and so for the purposes of this work five, eighty-minute sequences were used.

8.3 Future direction of work

Given the scope of this work there are a number of areas that might benefit from further investigation. They are discussed below starting with image sensing techniques and then looking at ways in which the algorithm and its evaluation might be improved.

8.3.1 Image Acquisition

The point was made in Chapter 3 that the quality of image delivered by the image sensor was critical to the system's performance. Indeed many of the difficulties faced in algorithm design can be traced back to the limitations of the image sensor in terms of its dynamic range and gain control mechanisms.

Practical experience of operating the detector has indicated that sensor sensitivity is barely sufficient to form a useable image at night-time (particularly as filters were required to deal with peak solar illumination). As the night-time test site used in this work had nearby shop window lighting in addition to street lighting even more difficulty would be expected when other test sites are added to the data set. The simplest solution to this difficulty is probably to add a source of near infrared illumination that can be activated when required.

The importance of obtaining good quality input images was such that developments in image sensing technology were kept under review. It is of interest that several research groups are working on devices that are intended to avoid some of the exposure control problems highlighted in Chapter 3. One approach being followed at the University of Leuven is to make the sensitivity of each pixel logarithmic allowing operation over a far wider range of intensity levels without saturation occurring. Another alternative under development is a sensor with locally adaptive exposure control so that each part of the image can be correctly adjusted for the light levels from the parts of the scene it is exposed to.

In Chapter 5 it was mentioned that stereo vision offers the major benefit of range measurement allowing separation of surface features (shadow, pavement texture) from objects above the plane of the pavement. It therefore has the potential, particularly when combined with the analysis methods developed thus far, to provide a more reliable solution.

The result of the earlier consideration of stereo was a decision that the capabilities of a monocular system should be explored first particularly in view of the practical difficulties in capturing, storing and processing the two synchronised data streams. As the work progressed however further knowledge of hardware options (from the Teaching Company collaboration) and decreasing sensor cost meant a re-evaluation of the use of stereoscopic system was justified.

The main difficulties in using stereo aside from the additional cost of a second sensor were:

- There is the increased equipment overhead of requiring two frame-grabbers.
- Frame capture synchronisation is necessary to ensure image disparities are due to position and not to motion.
- Common exposure control is valuable otherwise the contrast of features imaged by one camera may differ from those in the other and lead to a loss of clear correspondence.
- There is double the data flow for transfer and processing. Evaluation requires that repeatable recordings of pedestrian scenes can be recorded.

- It is difficult to take two separate, synchronised video recordings for off-line analysis. Even if this is achieved, there is the problem of playing the results back in synchronism.

To address these problems an experimental stereo vision system has been built in which the video streams from the two cameras are multiplexed onto a single output video stream, either on a line-by-line or a frame-by-frame basis. Using the line-by line system the temporal disparity is sufficiently small (64 microseconds) that any disparity due to the speed of motion of objects could be neglected. Using this system algorithms will be able to be developed on a conventional single frame-grabber system with street-gathered test data that has been recorded onto a single conventional video recorder - without the synchronisation problems mentioned above.

This development equipment was relatively expensive (the two synchronised cameras used on the test system cost £600.00 each). However work on a custom hardware design has indicated that grabbing images from two cameras simultaneously can be achieved with little hardware overhead by concatenating the two, 8-bit bytes (one per video stream) into a single 16-bit word. This can then be read in a single operation onto the 32bit bus of a digital signal processor. The costs of mass-produced CCD sensors have now dropped sufficiently that this is also no longer a barrier to such an approach.

8.3.2 Algorithm Development

Important weaknesses of the system as it stands include performance at low light levels, due to loss of contrast in the image from the image sensor, and the failure to detect umbrellas, as they do not fit the 'box' shaped pedestrian model. The availability of increased processing power will allow these problems to be addressed. With sufficient processing capability, for example, multiple algorithms could contribute opinions on pedestrian presence in parallel and a combination, such as a majority vote, used to determine the current output condition. This would offer a way of coping with the problem of detecting umbrellas. An additional evidence integration process (pre-processing would be common to the pedestrian model activity) could examine the scene for evidence of conic sections that fit the range of scale and shape distortions that occur when an umbrella is projected into the image plane. The same model could be used, with changes to the model constraints, to detect bicycle wheels

to extend the detector to operate at Toucan crossings. The approach of combining independent opinions might also be used to make use of the strengths of other sensor types for increased reliability. For example, simple passive infrared detectors are now available at low cost and may be useful as a means of reducing any risk of false negative detection - but only when the subject is moving. This secondary binary output could be used to override a negative binary opinion from the vision algorithm when necessary.

A further simple measure to increase robustness using faster hardware would be to analyse images at a frame-rate greater than the specified minimum of two frames per second, then the output from the analysis of several frames could be combined. A majority voting system could then decrease the chance of transitory single cycle false negative errors being output from the system.

A concern that should be addressed relates to the detection of children. Currently the algorithm's spatial model, which is used for detection during evidence integration, is based on the typical pedestrian model dimensions at all times. There is therefore a danger that if the first pedestrian of a group to arrive were a child then they might be missed. An improvement would be to extract the first pedestrian using the minimum pedestrian model, to reduce chance of a false negative, and then to switch to the typical model for the rest of processing so as to minimise volumetric measurement errors.

The other important aspect of improving performance would be to continue the process of optimising the existing algorithm using feedback from the evaluation system as is discussed below.

8.3.3 Evaluation System

Having developed a multi-stage, multi-parameter algorithm a major problem is finding the best adjustment of the parameters and relating them to detection performance. It was demonstrated in Chapter 6 that the evaluation system could be used to perform a parametric search relating the resolution parameter to performance. It was the author's intention that the continuation of the work on the evaluation system would close the detection-evaluation loop such that performance data would be fed directly back to the algorithm. This would allow the algorithm's parameters (and/or structure) to be automatically optimised by adaptive means such as genetic

search. The large hard disks now available mean that all the test sequences used to date could easily be put together to allow the rapid automated test of an algorithm over all test data in a couple of hours.

It is also important that care be taken to keep test set broad enough to avoid overly simplistic solutions appearing to work. The evaluation system could therefore be improved by an extension of the data set to cover a wider range of sites and environmental conditions e.g. to cover snow and fog. It would also be beneficial if the time span of individual test sequences were longer. As each frame is analysed largely on an individual basis a time-lapse video recorder could be used to capture sessions of up to 40 hours which would give valuable information on the system's adaptation properties throughout the day-night cycle. It would also be of value if the degree of automation could be extended to shorten the time required to process the results - this is currently performed by hand using a spreadsheet.

8.3.4 System Aspects

The system developed thus far can be considered as being analogous to the peripheral vision of the human visual system in that it provides coarse cues as to the positions of pedestrian in the scene. It could now form a suitable starting point for a stage analogous to the H.V.S.'s foveal system to look for more detail on object shape perhaps by using more localised models for body parts such as heads and feet.

Whilst this step is probably not necessary for satisfying the specification of the binary and volumetric pedestrian detector, it will be required if the effects of occlusions are to be resolved in detail and if the extraction of behavioural information, based on the analysis of pedestrians as uniquely identified individuals, is required.

With this in mind the manual analysis of the test sequences used in the evaluation system was designed so that individual pedestrian trajectories could be extracted. This data is therefore ready to form the evaluative basis of future work on the tracking of individual pedestrians.

8.4 Wider Achievements

8.4.1 Hardware and Software Equipment Developed

A small run of the frame-grabber design (see Chapter 3) has been produced and used along with a version of the MISA software in the image-processing teaching

laboratory of the university's Electrical Engineering department. It has also been used as a platform for hardware based student projects on real-time image processing. The MISA software has been used as basis for the research of other workers in the university.

8.4.2 Independent Evaluation of Performance

The detection methods described in this document were selected, on a competitive basis, from about 14 other systems, by the Department of Transport for a funded, independent trial of volumetric pedestrian detectors. An embedded system implementation of the detector was submitted for the on-street trials, which were carried out by the Atkins Wootton-Jeffries consultancy acting behalf of the Department of Transport. A paper on the results of this trial, published by the Transport Research Laboratory (Crabtree, 1997), reported no operational problems with the detector. An additional positive outcome of this trial was that although the system was accidentally left in operation as an active part of the crossing control system for a period of one month, no public complaints were generated. According to the contacts at the City of Edinburgh Council, members of the public are quick to complain about problematic equipment such that this has become a normal performance metric for local councils to assess the acceptance of equipment.

8.4.3 Research Funding

The studies of crossing behaviour described in Chapter 2 contributed to a successful EPSRC grant application (held by Professor Dickinson and Dr. Wan) to look at pedestrian detection which funded the author for most of this work.

Towards the end of this work the government's Teaching Company Directorate awarded a technology transfer grant to enable the results of this work to be developed into an industrial product in conjunction with a local manufacturer of transportation equipment. The project was to run over a three-year period employing two associates for two years each. After running for 18 months the viability of meeting the hardware specification at the required cost had been demonstrated and a prototype system constructed. Unfortunately the industrial partners withdrew from the scheme at this point due to financial considerations.

8.5 Closing Remarks

The main achievement of this work is that a practical solution to an outdoor vision task has been demonstrated without imposing the constraints typically applied by the prior art concerning pedestrian characteristics and environmental factors. The development of the system has also taken place within the constraint of limited financial resources.

It is noteworthy that the use of empirical thresholds has been minimised to allow tolerance of a wide range of signal contrast in combination with the requirement for very high degree of reliability. This was particularly critical during pre-processing stages where methods used in the prior art (particularly binarisation) might have irreversibly removed information that was important for detection. An empirical threshold is first applied only at the end of the entire signal processing cascade so as to allow even the weakest of signal information to be exploited in making the best detection decisions. To avoid the use of early thresholds, ANDing operations have been derived to allow the logical combination of pre-processing representations of the images without resorting to a binary representation. This was performed by either a multiplication' or a 'minimum' operator in combining edge representations and in checking symmetry.

This work has also recognised the central importance of evaluation as part of the algorithm development process. An evaluation system was created to permit repeatable, semi-automated, off-line algorithm assessment that has allowed extensive testing on image sequences which were chosen to be representative of a wide range of difficult conditions. Previous work in vision was found to be lacking in this respect.

The application oriented nature of this work and its interdisciplinary position between transportation and computer vision meant that a wide range of skills was required. Consequently its undertaking has provided the author with valuable experience in a variety of areas. Management and supervisory experience have come from the organisation and direction of work for the on-street tests as well as from involvement in the Teaching Company Scheme. The latter also provided valuable awareness of the commercial realities of implementing vision systems. The majority of technical work was centred on the understanding and development of image processing and computer vision methods. In addition however a range of related practical skills have been

developed. These include C and C++ programming, hardware design for video signal capture, field programmable gate array design, hardware synthesis from a hardware description language, PCB layout and debugging as well as the use of embedded systems.

9 References

Akita K. "Image Sequence Analysis of Real World Human Motion" *Pattern Recognition* 1984; **17:** 73-83

Allsop R.A., Smith S.M. "Image Processing for the Analysis of Pedestrian Behaviour" *British Crown Copyright 1997, DERA, Farnborough, Hampshire, GU14 6TD 1997.*

Anderson S., Bruce W.H., Denyer P.B., Renshaw D., Wang G. "A single chip sensor and image processor for fingerprint verification" *IEEE International Conference CICC San Diego*, 1991.

Ando K., Ota H., Oki T. "Forecasting the Flow of People". (Japanese*) Railway Research Review*, 1988; **45**(8): 8-14

Attwood C.I., Sullivan G.D. "Model-based Recognition of Human Posture using Single Synthetic Images" *Fifth Alvey Vision Conference*, 1989, 25-30.

Azarbayejani A., Wren C., Pentland A "Real-time 3-D Tracking of the Human Body" *M.I.T. Media Laboratory Perceptual Computing Section Technical Report, published in Proceedings of IMAGECOM* 1996; **374**: 1-6

Bartolini F., Cappellini V. "Counting people getting in and out of a bus by real-time image-sequence" *Image and Vision Computing* 1994; **12**(12): 36-41.

Baumberg A,M., Hogg D.C. "Learning Flexible Models from Image Sequences".*European Conference on Computer Vision (ECCV)* 1994; **1**: 299-230.

Baumberg A., Hogg D. ""An Adaptive Eigenshape Model" *British Machine Vision Conference* 1995

Baumberg A., Hogg D. "Generating spatiotemporal models from examples" *Image and Vision Computing* 1996: **14**(8); 525-553.

Baumberg A.M., Hogg D.C. "An Efficient Method for Contour Tracking using Active Vision Shape Models" *University of Leeds, School of Computer Studies, Research Report Series* 1994

Baumberg A.M., Hogg D.C. "An Efficient Method for Contour Tracking using Active Shape Models" *IEEE Workshop on Motion of Non-Rigid and Articulated Objects* 1994; **5** (94):194-19

Baumberg A.M., Hogg D.C. "Learning Spatiotemporal Models from Training Examples" *University of Leeds, School of Computer Studies, Research Report Series* 1995

Binne T.D., Reading I.A.D. "Image Capture Board for the PC" *IJEEE* 1995; **32**(3); 235-255.

Blake A.,Yuile A. "Active Vision" Book from M.I.T. Press 1992.

Boyd G. and Barker D. J. (1995) Pedestrian non-compliance with pedestrians at signalised junctions. Working Paper, Department of Civil and Transportation Engineering, Napier University (Internal).

Burt P.J. "Fast filter transforms for image processing", *Computer Graphics and Image Processing* 1981; **(16):** 20-51.

Burt P.J. "Multi-resolution techniques for image representation, analysis, and 'smart' transmission" *SPIE* 1989; **1199**:1-15.

Burt P.J. "Image Motion Analysis Made Simple and Fast One Component at a Time" *BMVC* 1991;1-8.

Cai Q.,Mitiche A., Aggarval J.K. "Tracking Human Motion in an Indoor Environment" *IEEE International Conference on Image Processing* 1995; 215-18.

Cameron M.H. and Milne P.W. "Pedestrian exposure and risk in New South Wales" *Proceedings Joint ARRB/DOT Conference, Sydney.* 1978.

Campbell L.W., Bobick A.F. "Recognition of Human Body Motion Using Phase Space Constraints" *IEEE International Conference on Computer Vision* 1995; **8**(95): 624-63.

Chachich A, Pan A "Traffic sensor using a color vision method" *SPIE* 1997; **2902**: 156-165.

Chelette T.L., Repperger D.W., Repperger D.W., Phillips C.A. "Pattern Recognition of Spastic Motion" *Journal of Intelligent and Fuzzy Systems* 1996; B(2): 141-16.

Chen Z., Lee H.J. "Knowledge-guided Visual Perception of 3D Human Gait from a Single Image Sequence" *IEEE Transactions on Systems, Man and Cybernetics* 1992; **22**(2): 336-34.

Cheng J-C., Moura J.M.F. "Tracking Human Walking in Dynamic Scenes" *IEEE International Conference on Image Processing* 1997; **I**:137

Crabtree M, 1997, "A Study of Four Styles of Pedestrian Crossing" *PTRC* **session K**:171-182.

Crisman J.D. "Color Region Tracking for Vehicle Guidance" *Active Vision, Edited by Blake A. & Yuille A., published by M.I.T. Press* 1992.

Darkin C.G.. "Investigation Into Video Image Processing Applied to Traffic Detection" *Final Report to TRRL Research Contract No. TRR 842/408,* 1986.

Davies H.E.H. "The Puffin Pedestrian Crossing: Experience With the first experimental sites", *Transport Research Laboratory, Department of Transport, Research Report 364*, 1992.

Davis J.W., Bobick A., "The representation and recognition of human movement using temporal templates" *IEEE International Conference on Computer Vision and Pattern Recognition*, 1997: 928-934.

Department Of Transport "Specifications for Kerbside Detection Systems for use at Puffin Crossings" *Traffic Systems and Signing Division* TR2182, January 1997.

Department Of Transport "Requirements for new pedestrian crossings strategy (Puffin)" *Network Management and Driver Information Division.* September 1993.

Department Of Transport "Volumetric Puffin Specification" *Network Management and Driver Information Division* 1996.

Department Of Transport Urban and Local Transport Directorate Research Committee (ULTDRC), research project UG54.

Department Of Transport Urban and Local Transport Directorate Research Committee (ULTDRC), research project N107.

Ekman L., Sherbourne D. "Microwave Detection of Pedestrians in England and Sweden", *Proceedings IEE 6th International Conference on Road Traffic Monitoring and Control* 1992: 210-213.

Fathy M., Siyal M.Y. "A Combined edge detection and background differencing image processing" *Road and Transport Research* 1995; **4**(3): 112-114

Garder P. "Pedestrian safety at traffic signals" *Accident analysis and prevention* 1989; **21**(5): 435-444.

Gavrila D.M., Davis L.S "3D model-based tracking of humans in action: a multi-view approach" *IEEE International Conference on Computer Vision and Pattern Recognition* 1996;73-80.

Glachet R., Bouzar S., Lenoir F., Blosseville J-M "Counting pedestrians in the subway corridors using image processing" *SPIE* 1995; **2564:**261-27

Gu H., Shirai Y., Asada "MDL-Based Segmentation and Motion Modeling in a Long Image Sequence of scene with multiple independently moving objects" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1996: **18**(1): 58-64

Guo Y., Xu G., Tsuji S. "Understanding Human Motion Patterns" *IEEE International Conference on Pattern Recognition* 1994; **1**(94): 25-32

Heap T., Hogg D. "Extending the Point Distribution model using polar coordinates." *Image and Vision Computing* 1996; **14**(8): 589-59

Heisele B., Kressel U, Ritter W. "Tracking Non-Rigid Moving Objects Based on Colour Cluster Flow" *IEEE International Conference on Computer Vision and Pattern Recognition* 1997; 257-26

Hogg D. "Model-based vision: a program to see a walking person" *Image and Vision Computing* 1983 **1**(1): 5-20

Hunt J.G. "Pelican Crossing Operation in Areas under UTC" *Proceedings IEE 3rd International Conference on Road Traffic Monitoring and Control* 1990; 129-133

Hunt J.G., Chik A. A. "Midblock Signalled Pedestrian Crossings - Alternative Operating Strategies" *IEE 8th International Conference on Road Traffic Monitoring and Control, 23-25 April* 1996: **Conference Publication No. 422**: 126-130.

Hunt J.G., Lyons G.D. "The Operation of Pelican Crossings - Is Pedestrian Detection

Worthwhile?" *Proceedings IEE International Conference on Road Traffic Monitoring and Control 28-30 April* 1992; 48-52

Hwang B.W., Takaba S. "Real-time measurement of pedestrian flow using processing of ITV images" *Systems Computer Controls* 1983; **14**(4): 46-55.

Iwasawa S., Ebihara K., Ohya J., Morishima S. "Real-Time Estimation of Human Body Posture from Monocular Thermal" *IEEE International Conference on Computer Vision and Pattern Recognition* 1997; 15-20

Kakadiaris I.A., Metaxas D. "3D Human Body Model Acquisition from Multiple Views" *IEEE International Conference on Computer Vision* 1995; **8**(95): 618-62

Khoudour L., Duvieubourg L., Deparis J-P "Real-Time Pedestrian counting by Active Linear Cameras" *Journal of Electronic Imaging* 1996; **5**(4): 452-45

Kinzel W. "Pedestrian Recognition by Modelling their Shapes and Movements" *International Conference on Image Analysis and Processing* 1994; 547-55

Lerasle F., Rives G. "Human Body Tracking using Monocular Vision" *European Conference on Computer Vision (ECCV)* 1996; 518-52

Leung M.K., Yang Y.H. "Human Body Motion Segmentation in a Complex Scene" *Pattern Recognition* 1987; **20**(1): 55-64.

Leung M.K., Yang Y.H. "First Sight: A human-body outline labelling system" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1995;**17**(4): 359-37

Levelt P.B.M. "Pussycats: New Pedestrian Facilities: Technique Observations and Opinions" *Proceedings of Conference on Strategic Highway Research Program and Traffic Safety on Two Continents*, 1994: **A**(2): 144-157.

Long W., Yang Y-H. "Log-Tracker: An Attribute-Based Approach to Tracking Human Body Motion" *International Journal of Pattern Recognition and Artifcial Intelligence* 1991; **5**(3): 439-45.

Lu Y.J., Yuan X., Yan C.D "Image Sequence Analysis for Measurement of Pedestrian Flows" *Journal of the Transportation Research Forum* 1992; **32**(2): 399-40.

Lu Y-J, Tang Y-Y, Pirard P., Hsu Y-H., Cheng H-D "Measurement of Pedestrian flow data using image analysis techniques" *Transportation Research Record* 1990;

1281: 87-96.

Marr D "Vision" *W.H. Freeman and Company, New York* 1982

Meyer D., Denzler J., Niemann H. "Model Based Extraction of Articulated Objects in Image Sequences for Gait" *IEEE International Conference on Image Processing* 1997; **III:** 78-81

Microsense Limited, Meon House, 10 Barnes Wallis Road, Segensworth, Fareham, Hampshire, PO15 5TT. "Above Ground Kerbside Detector AKD-R-24".

Michalopoulos PG "Vehicle detection video through image processing: the autoscope system *IEEE Transactions on Vehicular Technology* 1991; **40**(1): 21-29.

Mohr R, Triggs B., "Projective Geometry for Image Analysis" *ISPRS* Tutorial Session 1996.

Mori H., Charkari N.M., Matsushita T "On-Line Vehicle and Pedestrian Detections Based on Sign Pattern" *IEEE Transactions on Industrial Electronics* 1994; **41**(4): 384-39

Nishihara H.K., Thomas H.J., "Real-time tracking of people using stereo and motion" *S.P.I.E.* 1994 ; **2183**: 266-273.

Okawa Y., Hanatani. "Determination of a pedestrian who does a prespecified body motion in a structured environment" *IEEE International Conference on Industrial Electronics, Control and Instrumentation* 1991; **2591**: 2343-2.

Oren M., Papageorgiou C., Sinha P., Osuna E, Poggio T. "A Trainable System for People Detection". *Proceedings of DARPA Image Understanding Workshop May 11-14* 1997; **1**: 207-21.

Oren M., Papageorgiou C., Sinha P., Osuna E., Poggio T. "Pedestrian Detection Using Wavelet Templates" *IEEE International Conference on Computer Vision and Pattern Recognition, June 17-79.* 1997.

O'Rourke J., Badler N.I. "Model-Based Analysis of Human Motion using Constraint Propagation" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1980; **2**(6): 522-53.

Papanikolopoulos N.P, Richards C.A. "Detection and tracking of traffic objects in IVHS vision sensing modalities" *Proceedings of 1995 Annual Meeting of ITS,*

America 1995; 453-461.

Polana R., Nelson R. "Low Level Recognition of Human Motion" *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects* 1994; 77-82

Pye C.J., Bangham J.A. "2D Pattern Recognition using 1D Sieves based on alternating sequential, root" *EUSIPCO* 1994: 856-85.

Quian R.J., Huang T.S., "Motion Analysis of Human Ambulatory Patterns" *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1992; **1**:220-223.

Reading I.A.D., Dickinson K.W.D., Barker D. J. "The Puffin pedestrian crossing pedestrian-behavioural study" *Traffic Engineering and Control* 1995; **36**(9): 472-478.

Reading I.A.D., Wan C.L., Dickinson K.W.D. "Pedestrian Detection at Road Crossings by Computer Vision" *IEE International Conference on Road Traffic Monitoring and Control* 1996

Reading I.A.D., Wan C.L., Dickinson K.W.D., "Developments in Pedestrian Detection" *Traffic Engineering and Control* 1995; **36(**10): 538-541.

Regazzoni C.S., Tesei A. "Distributed data fusion for real-time crowding estimation" *Signal Processing* 1996; **53:** 47-63

Richards C.A., Smith C.E. *Proceedings of Annual Meeting of I.T.S. America.* 1995: 453-56

Richards C.A., Smith C.E., Papanikolopoulos N.P. "Vision-Based Intelligent Control of Transportation Systems" *IEEE International Symposium on Intelligent Control of Transportation.* 1995; **5**(95): 519-52.

Rohr K. "Incremental Recognition of Pedestrians from Image Sequences" *IEEE International Conference on Computer Vision and Pattern Recognition* 1993; 8-13.

Rohr K. "Towards model-based recognition of human movements in image sequences" *Computer Vision Graphics and Image Processing* 1994; **59**(1): 94-115

Rossi M., Bozzoli A. "Tracking and counting moving people" *IEEE International Conference on Image Processing* 1994; **3**: 212-21

Rothengatter J.A., Sherborne D. J. "Responsive Signal Settings for Pedestrians in Urban Areas" *Proceedings of 1st World Congress on Applications of Transport Telematics and Intelligent Vehicle Highway Systems* 1994; **2**: 477-478.

Rourke A., Bell M.G.H. "An image-processing system for pedestrian data collection" *IEE International Conference on Road Traffic Monitoring and Contro*l 1994; **391:** 123-12.

Rourke A., Bell M.G.H. "Wide area pedestrian monitoring using video image processing" *IEE International Conference on Image Processing and its Applications* 1992; 563-56.

Sarma P., Seken J. "Performance Evaluation of a People Tracking System" *Proceedings of 3rd IEEE Workshop on the Applications of Computer Vision* 1996: 33-38.

Sato A.,Mase K., Tomono A., Ishii K. "Pedestrian counting system robust against illumination changes" *SPIE* 1993; **2094**(3): 1259-1.

van Schagen I.N.C., Sherbourne D. J. "Evaluation of Microwave Detection Equipment in terms of Reliability, Durability and Pedestrian Safety" *DRIVE Conference project* 1991; **1031**(11): 1526-1538.

Schalkoff R.J. "Digital Image Processing and Computer Vision" *John Wiley and Sons Inc.* 1989.

Schofield A.J.,Mehta P.A.,Stonham T.J. "A system for counting people in video images using neural networks to identify the background scene" *Pattern Recognition* 1996; **29**(8): 1421-1

Schofield A.J., Stonham T.J. "Automated people counting using image processing and neural network techniques" *The 3rd International Confernece on Automation, Robotics and Computer Vision (ICARCV 94)* 1994; **TP1.6**: 903-906.

Seed N.L., Houghton A.D. "Background updating for real-time image processing at TV rates" *SPIE* 1988; **901**: 73-81.

Sexton G., Zhang X. "Suppression of Shadows for Improved Object Discriminaton" *IEE Colloquium on Image Processing for Transport Applications* 1993.

Sexton G., Zhang X. "Automated Counting of Pedestrians" *SPIE* 1994; **230**(2): 830-

83.

Sexton G., Zhang X., Redpath G., Greaves D. "Advances in Automated Pedestrian Counting" *European Convention on Security and Detection* 1995; **408:** 106-11.

Smith R.A."Volume Flow Rates of Densely Packed Crowds" *Engineering for Crowd Safety, Elsevier Science Publishers,* 1993: 313-319.

Sullivan M.J. Richards C.A., Smith C.E., Masoud O. "Pedestrian Tracking from a Stationary Camera using Active Deformable Models" *IEEE Conference on Intelligent Vehicles* 1995; **95th**: 90-95.

Tang Y.Y., Lu Y.J., Suen C.Y "A robot vision system for pedestrian-flow detection" *SPIE* 1991; **1611**: 630-64.

Tarko A., Tracz M. "Accident prediction models for signalized crosswalks" *Safety Science* 1995; **19**: 109-111.

Tesei A., Foresti G.L., Regazzoni C.S. "Human Body Modelling for People Localisation and Tracking from Real Image" *IEE International Conference on Image Processing and its Applications* 1995; **410**: 806-80.

Traffic Control Users Group "Traffic Signals Survey 1996" *County Surveyors' Society, Traffic and Safety Committee, Traffic Management Working Group* June 1997: T&S/3-97.

Traffic Control Users Group. "Traffic Signals Survey 1994" *County Surveyors' Society, Environment Committee, Traffic Management Working Group*, December 1994: ENV /3-94.

Tsuchikawa ,M., Sato A., Koike H., Tomono A "A Moving Object Extraction Method Robust Against Illumination Level Changes for a Pedestrian Counting System" *IEEE Symposium on Computer Vision* 1995; 563-56.

Tsukuyama T., Shirai Y. "Detection on the movements of persons from a sparse sequence of TV images" Pattern Recognition 1985;**18**;207-213.

Vannoorenberghe P, Motamed C "Crowd monitoring using image sequence processing" *PTRC* 1997;**Session K**: 197-206.

Velastin S. "Software Development for a Vision System" *PhD thesis Victoria University of Manchester* 1982.

Velastin S.A., Yin J.H. "Analysis of crowd movements and densities in built-up environments using image processing" *IEE Colloquium on Image processing for Transport Applications* 1993; **digest no 236**: 8/1-8/6.

Velastin S.A., Yin J.H. "Automated measurement of crowd density and motion using image processing" *7th IEE Int Conf on Road Traffic Monitoring and Control* 26-28 April 1994, London, UK. **29-29A**

Velastin S.A., Yin J.H. "Image processing for on-line analysis of crowds in public areas" *7th Int Symp. On Transportation Systems (IFAC/IFORS 'TS94)* 24-26 August 1994, Tianjin, China.

Waterfall R.C., "An Automated Data-Capture System for CAAD" *PhD Thesis submitted to Victoria University of Manchester* 1981.

Wren C., Azarbayejani A., Darrell T., Pentland A. "Pfinder: Real-time tracking of the Human Body" M.I.T. *Media Laboratory Perceptual Computing Section Technial Report No.353* 1995.

Wren C., Azarbayejani A., Darrell T., Pentland A. "Pfinder: Real-time tracking of the Human Body" *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 1997; **19**(7): 780-78.

Xidong Y, Yeau-Jye L, Sarraf S "A Computer Vision System for the Measurement of Pedestrian Volume" *TENCON* 1993; **2**: 1046-1.

Yasutomi S.,Mori H., Kotani S. "Finding pedestrians by estimating temporal-frequency and spatial-period of the moving objects" *Robotics and Autonomous Systems* 1996; **17**: 25-34.

Yin J.H., Zhang X. "Incident Detection in Pedestrian Traffic using Image Processing" *Road Traffic Monitoring and Control* 1996; 422 (conference publication).

Zhang X., Sexton G. "A new method for pedestrian counting" *IEE International Conference on Image Processing and its Applications* 1995; **410**: 208-21

10 Appendix A: Pelican/Puffin Crossing Operation

The signal aspects and typical time settings of both the Pelican and Puffin crossings are given below. The actual values used are based on those that were in operation during the before and after phases of the study described by Reading et al, 1995 for the Pelican and Puffin respectively.

A Pelican crossing installation operated as shown in Figure 50 with the settings shown in Table 13 offers a fixed time available to pedestrians. This period is the period between the start of green to pedestrians, labelled period (d), and the end of flashing green to pedestrians, period (e) making a total fixed time of 17 seconds. The total cycle time is 45 seconds.

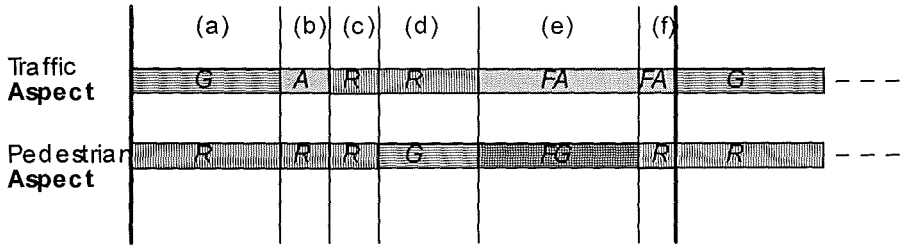


Figure 50: PELICAN Crossing Cycle

Period	Traffic Aspect	Pedestrian Aspect	Duration	Class
(a)	Green	Red	20s	Programmable
(b)	Amber	Red	3s	Fixed
(c)	Red	Red	3s	Programmable
(d)	Red	Green	6s	Programmable
(e)	Flashing Amber	Flashing Green	11s	Programmable
(f)	Flashing Amber	Red	2s	Fixed

Table 13: PELICAN Crossing Timings

A Puffin crossing, see Figure 51 and Table 14, includes a green man time of 6 seconds, a fixed pedestrian clearance period of 3 seconds, a variable extension of the clearance period of between 0 and 13 seconds, and a possible additional 3 seconds clearance time following maximum extension giving, in total, a minimum of 9 seconds and a maximum of 25 seconds of pedestrian priority. The length of the extension time is governed by the output of an on-crossing detector which determines if any moving pedestrians are still moving on the crossing area. Other important differences to note with respect to the Pelican are that the state of priority is clearly either with pedestrians or vehicles at any given point in the cycle and that for all but 6 seconds of a possible 25 seconds of pedestrian priority a red man (which is now situated on the near side of the crossing) is shown thus enforcing the role of this signal as invitation only to **start** crossing. The cycle time of the crossing now ranges from 47 seconds (assuming minimum extension time) to 63 seconds (assuming maximum extension time).

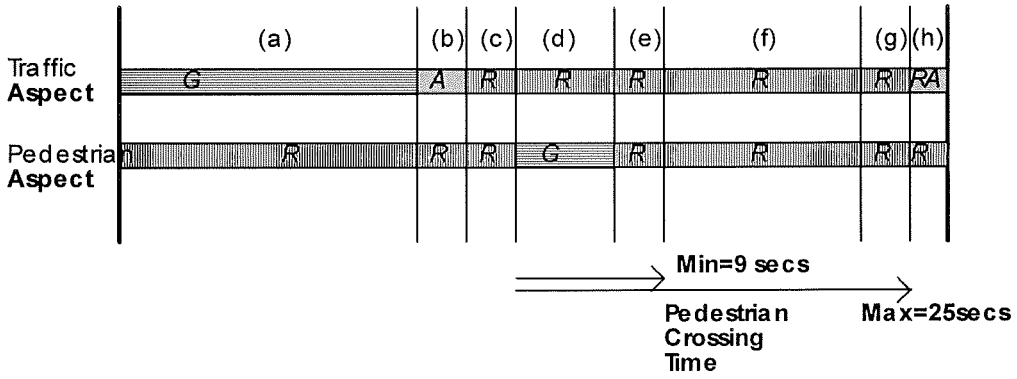


Figure 51: PUFFIN crossing cycle

Period	Aspect		Duration	Class	Description
	Traffic	Pedestrian			
(a)	Green	Red	30s	Variable	Minimum of Traffic Priority
(b)	Amber	Red	3s	Fixed	
(c)	Red	Red	3s	Programmable	
(d)	Red	Green	6s	Programmable	Green Man
(e)	Red	Red	3s	Fixed	Pedestrian Clearance
(f)	Red	Red	0-13s	Variable	Clearance Extension
(g)	Red	Red	3s	Fixed	Additional Clearance on Max. Extension
(h)	Red + Amber	Red	2s	Fixed	

Table 14: Key to Figure 51

More detailed information on the Puffin crossing can be found in the specification (Department of Transport 1993).

11 Appendix B: MoniPed Operator's Manual

MoniPed©

v 1.0

USER GUIDE TO SOFTWARE SET-UP AND INSTALLATION

Dept. Civil & Transportation Engineering

NAPIER UNIVERSITY

Edinburgh

Table of Contents

The function of the MoniPed© pedestrian detection system is to detect pedestrians waiting on the footway, at a road crossing. Outputs are provided that indicate both the presence or absence, and the number of pedestrians in the monitored area. These outputs are appropriate for the requirements of Puffin crossing controllers.

The system consists of two units. A Pedestrian Detection Unit (PDU) mounted in a roadside cabinet, and a Camera Head Unit mounted on a signal pole adjacent to the waiting area. Video images from the camera are analysed within the PDU using computer vision techniques to provide the required output signals.

To install the detector the software set-up procedure (starting on page 5) should be followed, once the preparation (page 4) has been completed.

Purpose

The electrical connection of the hardware components of the *MoniPed*© system required prior to software set-up.

Additional equipment required for the set-up process is:

VGA monitor (VGA signal lead and power lead)

IBM AT Keyboard (5-pin din connector - NOT PS/2 type!)

Procedure

1. The Camera head is attached to the signal pole at a height of between 3.0 and 4.0 metres, overlooking the area to be monitored.
2. Turn off the mains power supply to the PDU.
3. Connect the power supply to the VGA monitor from the PDU.
4. Connect the VGA signal lead to the VGA connector on the PDU.
5. Connect an IBM AT Keyboard to the keyboard connector on the PDU.
6. Connect the video output from the camera head unit to the video input connector on the PDU.
7. Connect the 12-17v supply to the camera head unit.

The *MoniPed*© system is now ready for software set-up.

Purpose

The set-up of the *MoniPed*© system software for calibration to the characteristics of the operating site.

Procedure

Switch on the mains power to the *MoniPed*© system.

The operational display which shows the current status of the system will appear after a few seconds warm up.

To leave the operational mode and access the menu screen, press 'Q'.

You will then be presented with the following options:

A to Align and Check Camera

M to Enter Calibration Measurements

C to Check Calibration

D to Specify Detection Zone

P* to Set Operating Parameters

O* to Set Operating Mode

V* to View Current parameters

S to Start Detector

L* to View Log File (ADVANCED)

R* to Restore Default Parameters (ADVANCED)

E* to Edit Parameter File (ADVANCED)

Q Quit

* Options NOT for user adjustment - for research purposes ONLY!

The user options (A, M, C, D, S & Q) should be activated in the order in which they appear in the menu to complete the set-up.

The following pages explain the use of each of these options in more detail. This information supplements the "on-screen" summary provided when each option is selected.

Purpose

The alignment procedure is used to ensure the correct positioning of the camera head unit and for the determination of the centre of the resulting field of view, which is required for calibration purposes. This option is also the most convenient means of checking that the camera and connections to the PDU are all correct by confirming that an image of the detection area is visible on the screen.

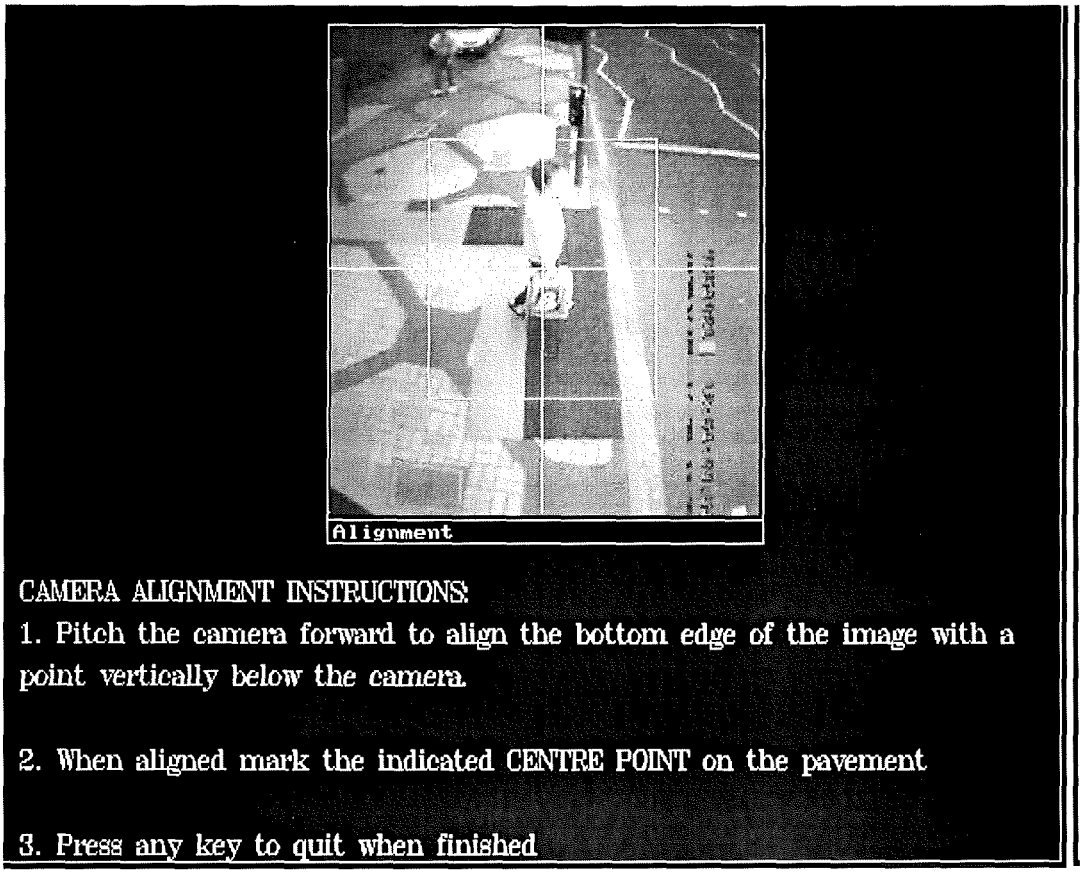


Figure 1. Typical Align/Check Screen

Procedure

When the Align/Check menu option is selected, the user will be presented with a camera image. A typical image can be seen in Fig.1. above.

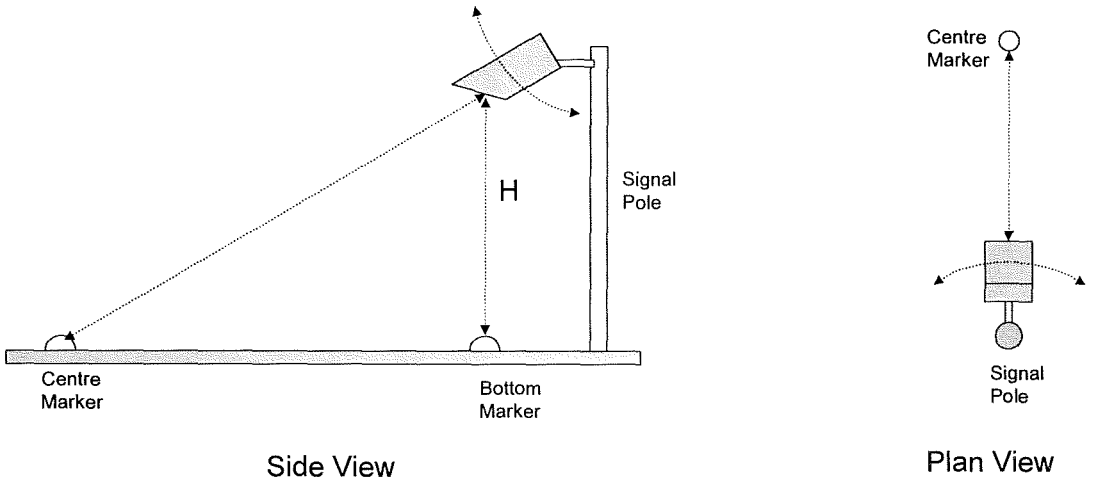


Figure 2. Camera head positioning (mount allows rotation about horizontal (pitch) and vertical (yaw) axes)

To align the camera, the following steps should be performed:

a. Place the bottom marker on the ground directly below the camera head.

b. Manually slacken the camera adjustment fixing screws. Manually adjust the pitch of the camera until the bottom marker is just visible at the centre of the bottom edge of the image.

c. Adjust the camera yaw (rotation about the vertical axis) to obtain the best coverage of the desired area.

d. Reposition the bottom marker to the centre of the bottom of the image (if required).

e. Tighten the adjustment screws. Thus the position of the **camera** is established.

f. Mark the position of the **centre** of the image on the pavement. This is indicated by the centre of the cross hairs.

g. Place the centre marker on this spot on the pavement such that it appears in the centre of the image, overlying the cross hairs.

The user is now able to take the necessary measurements to calibrate the *MoniPed*© system.

Purpose

This section deals with the calibration of the equipment and requires the user to take four measurements, based on the alignment procedure, and enter them into the system.

Procedure

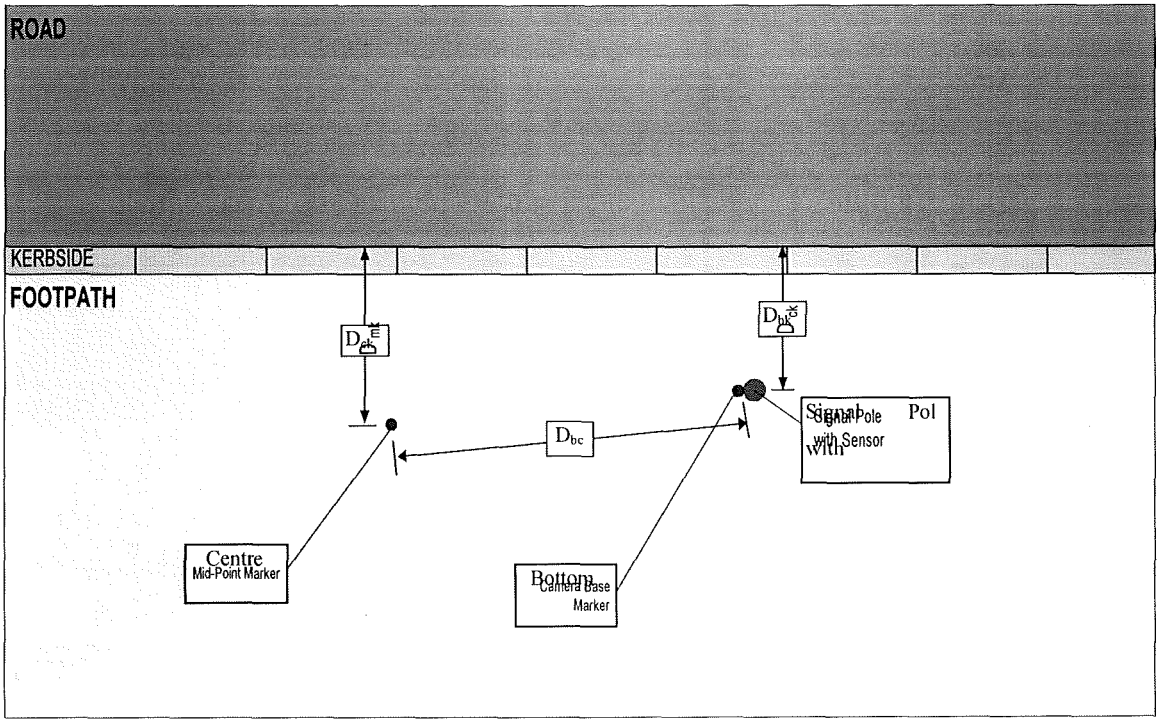


Figure 3. Installation Measurements

Symbol	Description
D_{bc}	Distance between the Bottom marker vertically below the camera's faceplate (near to the base of the signal pole) and the Centre marker.
D_{ck}	Distance between the Centre marker and the outside edge of the kerb.
H	Height from Bottom marker to centre of the camera's faceplate (see Figure.2).
D_{bk}	Distance between the Bottom marker vertically below the camera and the outside edge of kerb.

Table 15: Installation Measurements

Each measurement will be prompted for in turn (all should be in centimetres). When all have been entered the user will be given the opportunity to save (or reject) the changes made. Any mistakes can be corrected at this point, but all the measurements must be re-entered.

If an unacceptable numerical value outside of the permitted ranges (see Table.2) is entered, the value is ignored and the user is prompted to re-enter the value, and this will continue until an acceptable value is input by the user.

WARNING! - ONLY numerical characters must be input.

Measurement	Minimum	Maximum
D_{bc}	200	400
D_{ck}	0	200
H	300	400
D_{bk}	0	100

Table 2: Acceptable Measurement Values

Note: The measurements for the Bracknell test site are attached to this manual as an Appendix.

Purpose

The Check Calibration procedure allows the user to confirm the accuracy of the measurements given in the 'Enter Measurements (M)' section .

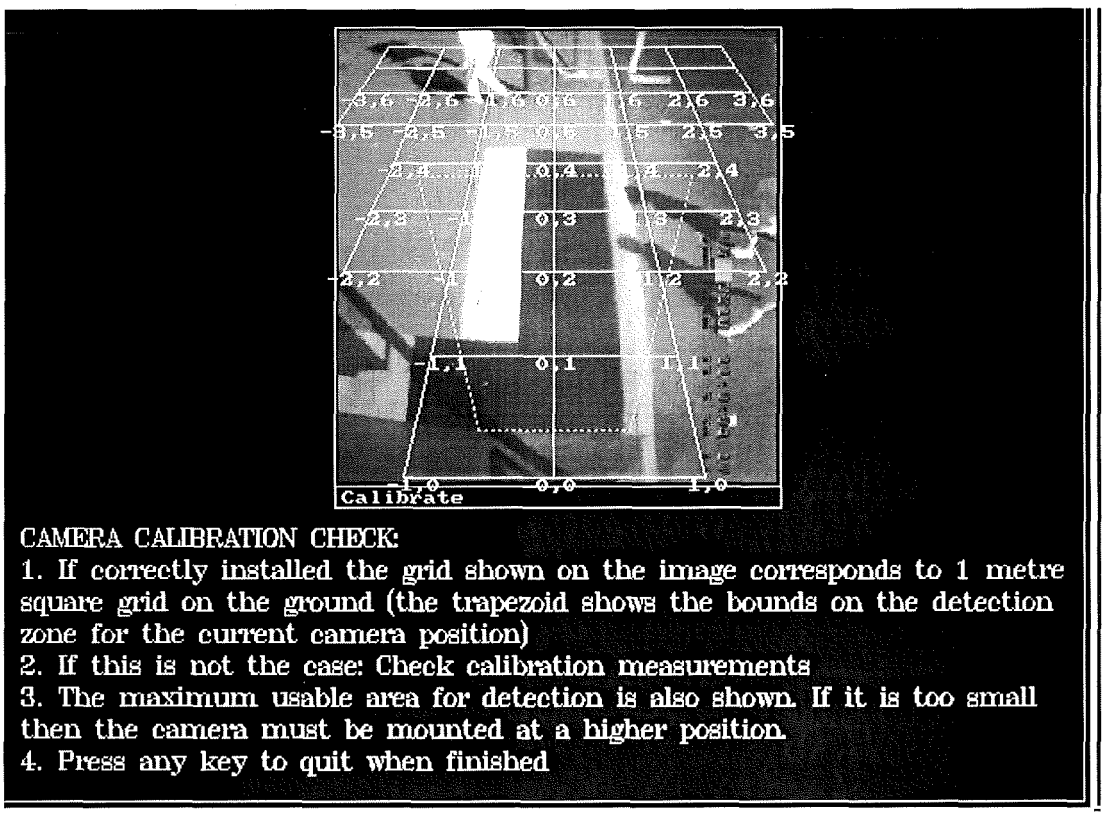


Figure 4. Typical Check Calibration Screen

Procedure

The user will be presented with a view of the scene. A typical image can be seen in Fig. 4. above. A grid is placed over the image. Each square in the grid represents 1 square metre on the ground.

This grid should match the ground fairly accurately, although this will not be perfect due to the presence of slight curvature of the image caused by the camera lens. If it is noticeably out of alignment then return to the alignment and measurement steps and ensure that the measurements are correct.

Explanation of the numerical values on the grid

The vertices of the metre squares are labelled with x,y co-ordinates, relative to an origin at the bottom marker. The y axis is parallel to the kerbside. The x,y co-ordinates indicate the position, in metres, of corresponding points on the pavement.

Purpose

This option allows the user to specify a detection zone to cover the area of pavement to be monitored.

Definitions

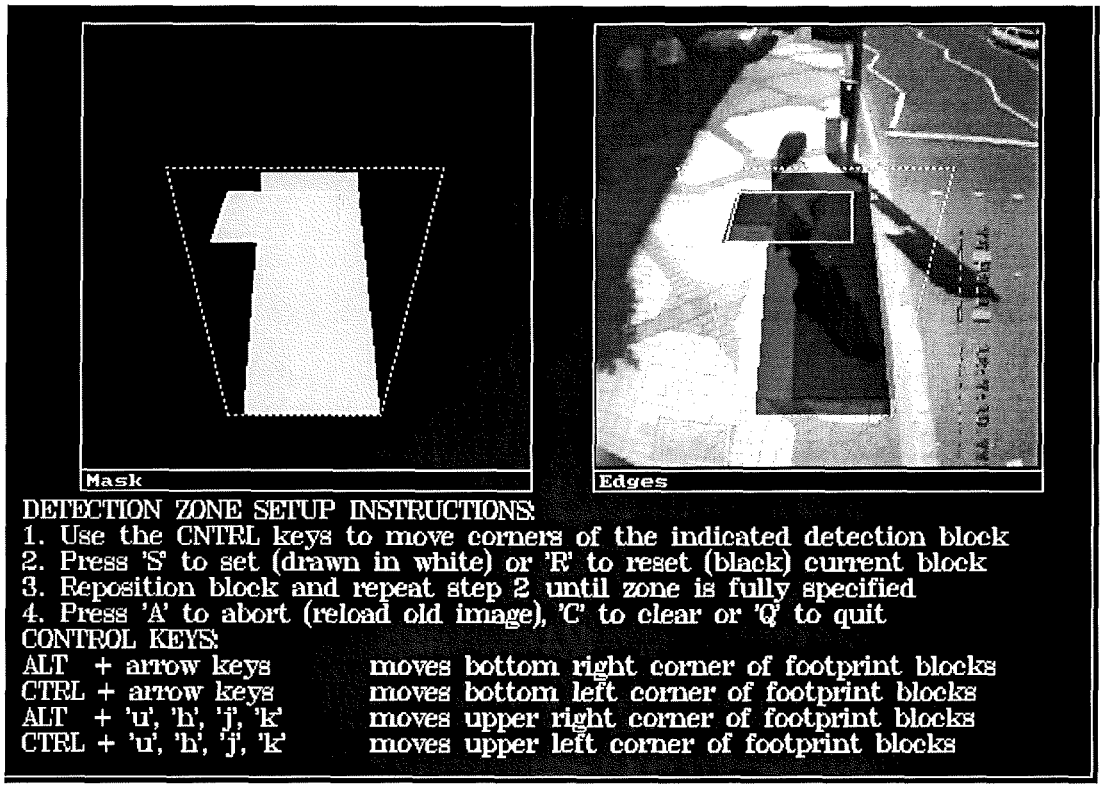


Figure 5. Sample view of Detection Zone specification

Two images are presented - the left-hand image shows the current active detection zone (in white). The right-hand image also indicates the current detection zone (dark shaded area) overlaid onto an image of the scene. In addition both images show the

detection zone limit (the quadrangle with a dashed border), and the painting region (the quadrangle with a solid border).

Detection zone: The white area in the left-hand image. This specifies the pavement area within which pedestrians are to be detected if their feet fall within it.

Detection zone limit: This is the "dashed area" (originally shown on the calibration screen (C)) which indicates the maximum limits of the usable detection area for the current camera head unit position. Although the borders of the detection zone may be specified outside the limit, they will automatically be clipped to the limit boundary when the algorithm is started ("S"), and any pedestrian activity in these areas will not be detected.

Painting region: Shown with solid border lines and initially a square lying at the centre of the image. Its position and shape can be manipulated by adjusting its corner positions according to the on-screen instructions.

Procedure

The detection zone is edited by means of the painting region. Using the keys described on the screen, move each corner of the painting region to the desired positions. Edit the detection zone shown in the left-hand image by the addition (set key "S") and removal (set key "R") of the bounded region to/from the detection zone.

It is possible to build up a larger, more complex detection area, by adding more than one detection zone to the image. Press "A" to abort the changes and reload the original settings. Press "C" to clear the current detection zone.

Note: As the detection zone is edited by the addition and removal of regions specified using the painting region the changes are immediately shown in the left-hand image as modifications of the white area. In the current version the changes are not shown in the right-hand image (due to memory limitations). However, the right-hand image

update can be achieved by quitting back to the menu (press "Q"), and then re-entering (press "D").

Press Q when finished editing the detection zone. This action returns you to the menu screen.

Purpose

This option shows the user the operating screen of the *MoniPed*© system - the system is now running in detection mode.

Procedure

No further user input is required when this option is running. You may pause and restart by pressing the space bar or press Q to return to the menu system.

Purpose

To allow the user to quit from the menu screen to the computer operating system.

Procedure

To quit the pedestrian detection program completely, press "Q" from the menu. This will return you to the computer operating system.

To restart the system, either press the reset button, or turn off the power to the unit for a short period and then turn power on again. The system will then boot straight into operating mode.

APPENDIX to MoniPed Manual

Installation Measurements for Bracknell Test Site

Sensor 1: EAST Side, Facing South

Camera Height	317	cm
Distance from Bottom marker to Centre marker	311	cm
Distance from Bottom marker to Kerbside	96	cm
Distance from Centre marker to Kerbside	95	cm

Sensor 2: West Side, Facing North

Camera Height	325	cm
Distance from Bottom marker to Centre marker	230	cm
Distance from Bottom marker to Kerbside	83	cm
Distance from Centre marker to Kerbside	76	cm

Reference Information

Black (pressure sensitive) Tiles	80cm square
Red (Tactile) Tiles	40cm square

12 Appendix C: Frame-Grabber Design

Binnie T.D. and Reading I.A.D., "Image Capture Board for the PC", International Journal of Electrical Engineering Education, vol.32, no. 3, pp 235-255, July 1995.

PUBLISHED PAPER(S)
NOT INCLUDED WITH THESIS

13 Appendix D: Information on Test Sites

13.1 Site 1: West End of Princes Street, Edinburgh

13.1.1 Reference Code: WEPS

13.1.2 Site Description

This site is a crossing sited on a busy junction at the western end of Princes street in central Edinburgh. The camera is mounted parallel to the direction of Princes street which runs approximately due east to west. Early morning sunshine therefore is directed straight towards the camera. During the day the sun projects shadows of vehicles from the busy road area over the pedestrian area creating fast moving patterns of intensity change. Bus traffic passes frequently down the near side lane of the road throwing shadows that frequently fill the entire observed kerbside waiting area. In the evening the sun is almost directly behind the camera causing very bright reflection off objects into the lens and causes very high contrast between pedestrians and their background.

The site is fairly open and Edinburgh, a windy city, combined with the presence of a Macdonald's nearby often leads to litter moving through the observed area.

A further valuable feature of this site is a nearby mast mounted camera intended for traffic monitoring and use by the police. It also gives an overview of the crossing permitting parallel acquisition of plan view video, which may be useful for later evaluations, on a transportation basis, of crossing performance. A plan view of the site layout is shown in below.

The co-operation of the City of Edinburgh Council for granting permission to use this site and for assisting with the required underground wiring is gratefully acknowledged.

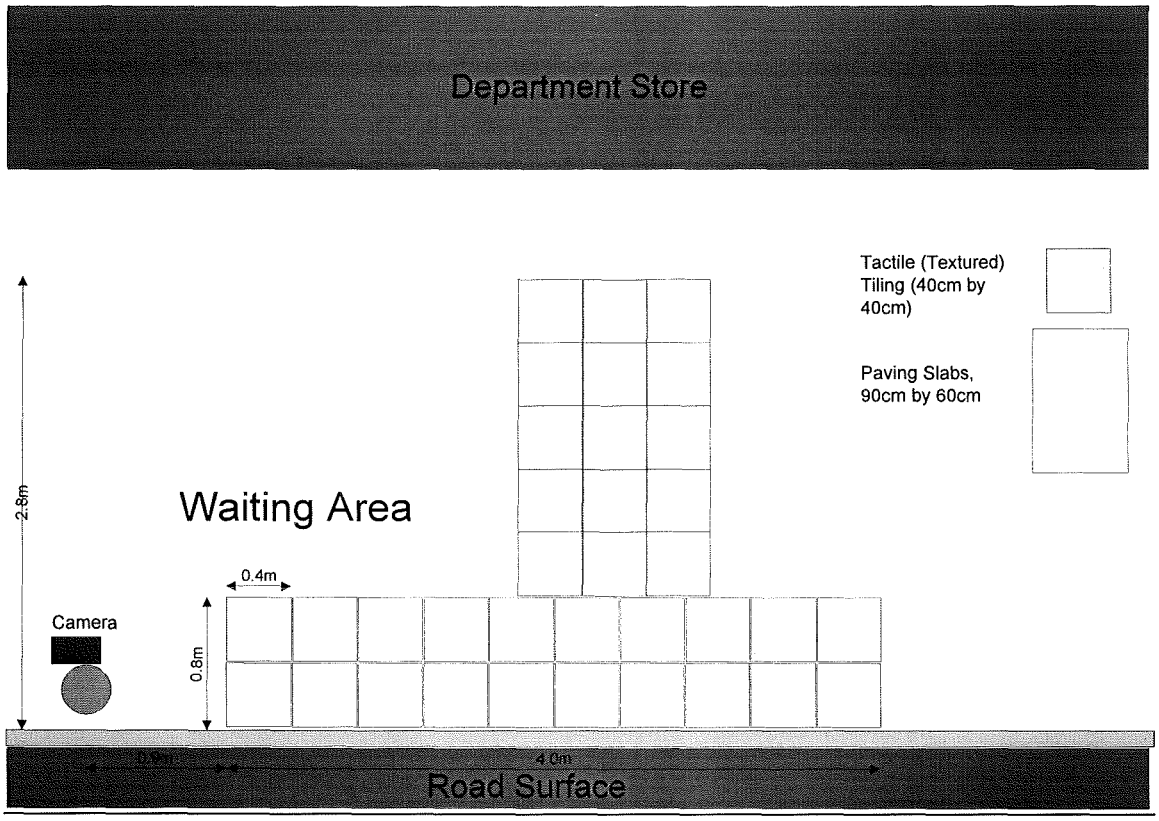
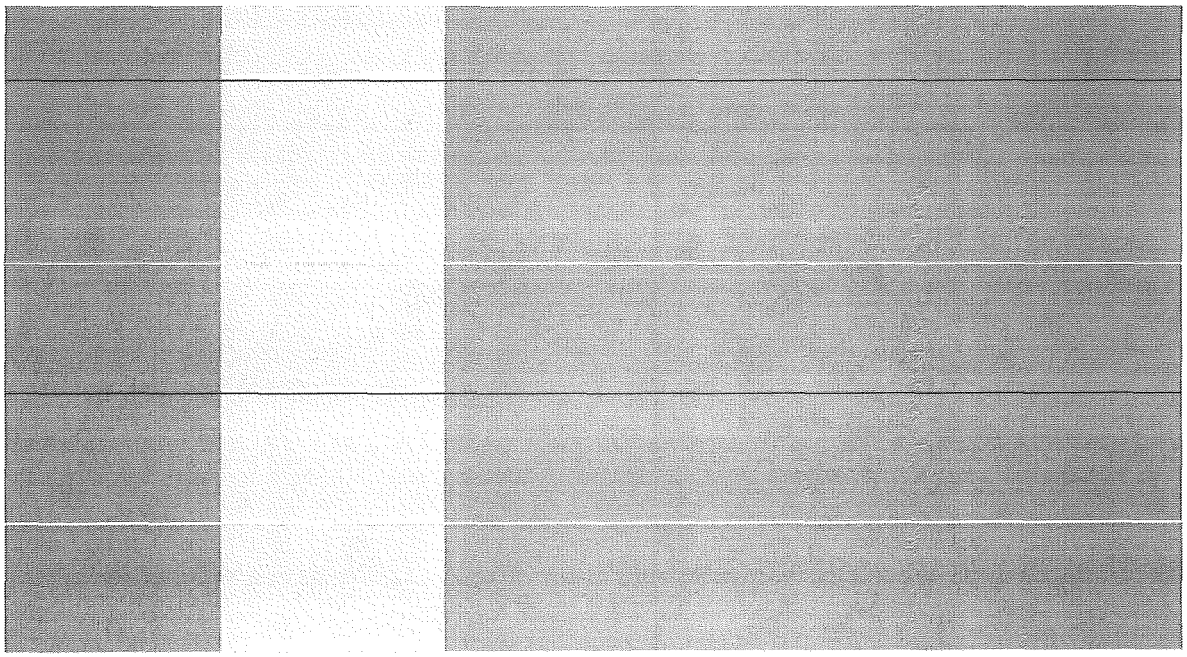


Figure 52: *Plan view of the Princes street test site.*

13.1.3 Calibration Information

Parameter	Value (cm)	Description
D_{bc}	288	Distance between the Bottom marker vertically below the camera's faceplate (near to the base of the signal pole) and the Centre marker.
D_{ck}	100	Distance between the Centre marker and the outside edge of the kerb.
H	305	Height from Bottom marker to centre of the camera's faceplate.
D_{bk}	103	Distance between the Bottom marker vertically below the camera and the outside edge of kerb.



13.2 Site 2: Bracknell Central East

13.2.1 Reference Code: BCe

13.2.2 Site Description

This site was set-up and monitored by consultants Atkins Wootton Jeffies as a trial site for testing of the volumetric puffin. The author (and colleagues) installed the detection system described in this work with their assistance.

The road through the crossing is oriented in an approximately north to south direction. Both the East and West sides of this crossing were monitored as part of the trials.

Tall buildings surround the site. A shopping centre to the south casts shadows on the left-hand edge of the image. An office building on the west has a large windowed area that can causes reflected time-variant light patterns to be projected onto the pavement area when the sun is in the south. The high buildings also cause a funnelling of the wind in a south to north direction. This, in combination with its situation next to a shopping centre (it is adjacent to a fast food store), leads to the presence of significant amounts of litter being moved around by the wind. In autumn a similar effect with moving leaves was observed. The air flow is further complicated by a second stream of air coming from the entrance to the shopping centre (a tunnel) to the east causing swirling eddies where they meet at the crossing's waiting area.

The East side camera was mounted facing south. During the mid-morning period the observed area is covered by a slow moving shadow region due to the shopping centre buildings to the south.

13.2.3 Calibration Information

Parameter	Value (cm)	Description
D_{bc}	311	Distance between the Bottom marker vertically below the camera's faceplate (near to the base of the signal pole) and the Centre marker.
D_{ck}	95	Distance between the Centre marker and the outside edge of the kerb.
H	317	Height from Bottom marker to centre of the camera's faceplate.
D_{bk}	96	Distance between the Bottom marker vertically below the camera and the outside edge of kerb.

13.2.4 Reference Information

Black (pressure sensitive) Tiles 80cm square

Red (Tactile) Tiles 40cm square

13.3 Site 3: Bracknell Central West ()

13.3.1 Reference Code: BCw

13.3.2 Site Description

For general site description see section on Bracknell Central East above.

The West side camera was mounted facing north. The usage of the waiting area is such that most of the pedestrian flow passes through on the side furthest from the camera.

13.3.3 Calibration Information

Parameter	Value (cm)	Description
D_{bc}	205	Distance between the Bottom marker vertically below the camera's faceplate (near to the base of the signal pole) and the Centre marker.
D_{ck}	80	Distance between the Centre marker and the outside edge of the kerb.
H	325	Height from Bottom marker to centre of the camera's faceplate.
D_{bk}	83	Distance between the Bottom marker vertically below the camera and the outside edge of kerb.

13.3.4 Calibration Information (After realignment on 24.6.96)

Parameter	Value (cm)	Description
D_{bc}	230	Distance between the Bottom marker vertically below the camera's faceplate (near to the base of the signal pole) and the Centre marker.
D_{ck}	76	Distance between the Centre marker and the outside edge of the kerb.
H	325	Height from Bottom marker to centre of the camera's faceplate.
D_{bk}	83	Distance between the Bottom marker vertically below the camera and the outside edge of kerb.

13.3.5 Reference Information

Black (pressure sensitive) Tiles	80cm square
Red (Tactile) Tiles	40cm square

13.4 Laboratory Calibration Sequences.

A set of short sequences were captured under indoor, controlled conditions to allow the accuracy of the calibration process to be assessed. The camera was fixed to the wall of the laboratory, and a calibration grid of 2x4 metres in size is laid out on the laboratory floor using masking tape. The bottom marker corresponded with the central point on the bottom line of the grid. The calibration parameters relevant to these captured sequences are as follows:

Camera Height	295 cm
Distance from Bottom marker to Central 2 metre mark on grid	200 cm
Distance from Bottom marker to 2 metre mark on grid to the right	223 cm
Distance from Bottom marker to 2 metre mark on grid to the left	223 cm
Distance from Wall to Central 2 metre mark on grid	158 cm

The parameters of each test sequence are encoded into the filename assigned to it. Each filename consists of three parts:

1) The first letter of the filename denotes whether the camera was centred, to the left or to the right of the grid marked out on the laboratory floor. This is either L for Left, C for Centre or R for Right.
2) The number (followed by an 'm') denotes at what distance from the camera the marker pole is placed, i.e. 1m for one metre, 2m for two metres, etc.
3) The rest of the filename denotes whether the marker pole is to the left, right or centre of the grid, i.e. Left for Left, Cent for Centre or Right for Right.

An example of a filename is R3mRight, which means that the camera is pointing to the right of the grid, and the marker pole is three metres from the camera on the right of the grid

14 Appendix E: Three Dimensional Modelling

14.1 Introduction

This section develops the necessary mathematics to support a top-down three-dimensional modelling of pedestrians and of the scene. For the pedestrian model to be of use it was necessary to relate its position in the image plane to a real world position by projecting it into the scene where it could be matched against the image data. Furthermore in order to be able to evaluate a hypothesis about a pedestrian being present at a particular pixel position in the image it was also necessary to be able to perform the inverse of this projection process. Inverse projection seeks to relate a particular image pixel position to a real-world position in the scene.

In the following, firstly the nature of the model and scene are described and appropriate co-ordinate systems defined. Then by assuming the availability of a small set of calibration measurements from the installation process and a flat pavement area the projection equations and their inverses are derived.

14.2 Definition of Co-ordinate Systems

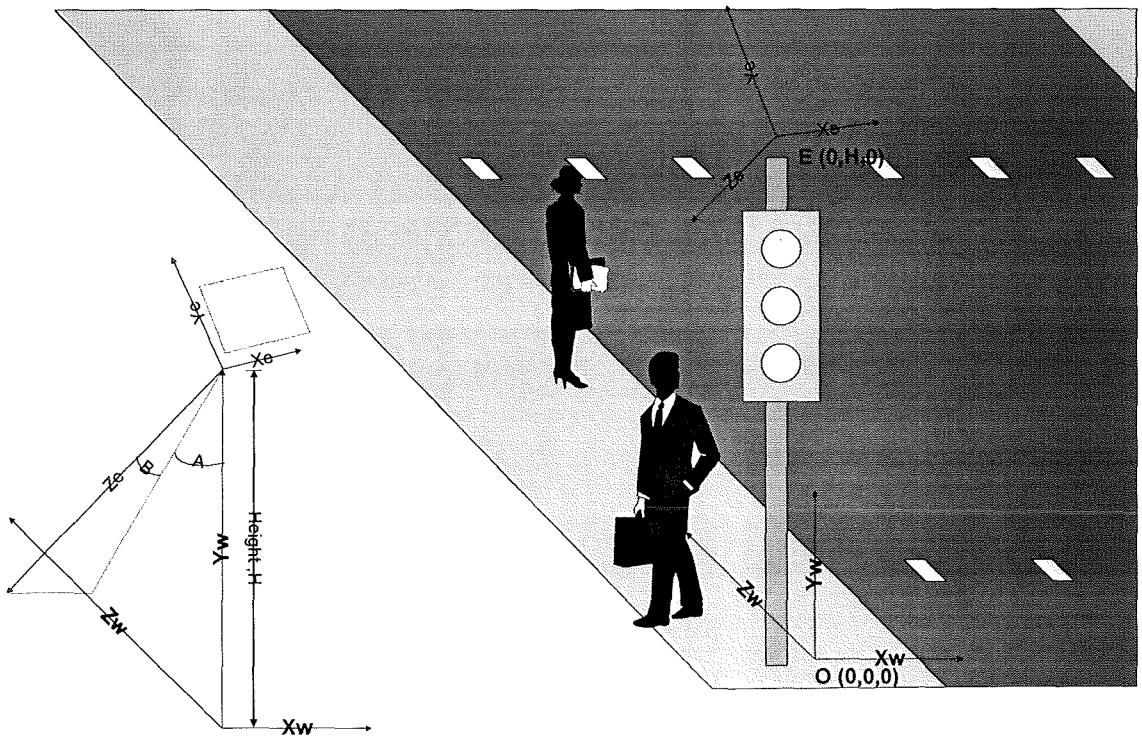


Figure 53: *Definition of world and eye co-ordinate systems*

The diagram shows the relationship of the world co-ordinate system (defined with respect to an origin at the base of the signal pole, with the z-axis along the surface of the pavement) to the eye (i.e. camera) co-ordinate system defined from an origin at the camera position with the z-axis looking down at the centre of the detection zone. Both world and eye co-ordinate systems are left-handed.

14.2.1 Pedestrian Model

A pedestrian was modelled by a 3D bounding box. Initial dimensions used for search were based on knowledge of typical pedestrian size and expected rotation. The reference point used as the origin of a pedestrian's co-ordinate system was the point on the ground plane which lay vertically beneath their head.

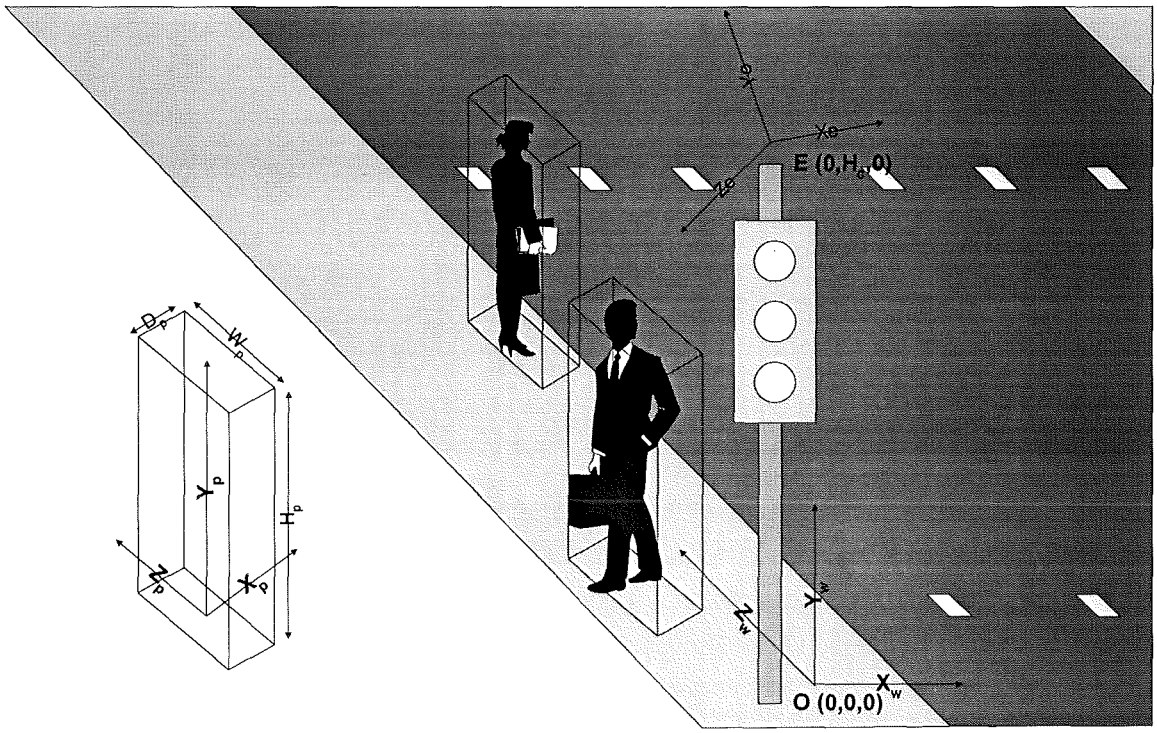


Figure 54: *Three Dimensional Pedestrian Model*

14.2.2 View Point Transformation

The following develops the equations necessary to transform from object positions described in world co-ordinates $\underline{x_w}$ to the position they will appear in the display co-ordinate system $\underline{x_d}$ of the camera image. This is achieved by transforming through the intermediate stages of the eye co-ordinate system $\underline{x_e}$ and the perspective projection onto the image sensor $\underline{x_{pp}}$.

14.2.2.1 Transformation from world co-ordinates to eye (i.e. camera) co-ordinates.

The first step is to obtain the co-ordinates of vertices defined in the world co-ordinate system when viewed from the position of the camera (the eye co-ordinate system). This involves translation up the height of the pole, H (performed by matrix T), followed by rotation about the Y axis by angle B (matrix R_Y) and then a second rotation about the X axis by angle A (matrix R_X).

$$\underline{x_e} = R_x . R_y . T . \underline{x_w}$$

where $\underline{x_w} = \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix}$ and $\underline{x_e} = \begin{bmatrix} x_e \\ y_e \\ z_e \\ 1 \end{bmatrix}$ are the world and eye position vectors in homogeneous co-ordinates.

Substituting the required rotation and translation matrices this expands to become:

$$\underline{x_e} = \begin{bmatrix} x_e \\ y_e \\ z_e \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & SinA & CosA & 0 \\ 0 & -CosA & SinA & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} CosB & 0 & -SinB & 0 \\ 0 & 1 & 0 & 0 \\ SinB & 0 & CosB & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -H \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \underline{x_w}$$

Resulting in:

$$\underline{x_e} = \begin{bmatrix} CosB & 0 & -SinB & 0 \\ CosASinB & SinA & CosACosB & -HSinA \\ SinASinB & -CosA & SinACosB & HCosA \\ 0 & 0 & 0 & 1 \end{bmatrix} . \underline{x_w}$$

14.2.2.2 Perspective Projection to transform eye co-ordinates into image plane co-ordinates and transformation to pixel co-ordinates

The vertices defined in eye co-ordinates are now transformed using the perspective projection, P to take account of their distances from the camera, into points in the image plane. This is achieved using the back projection model (Schalkoff, 1989).

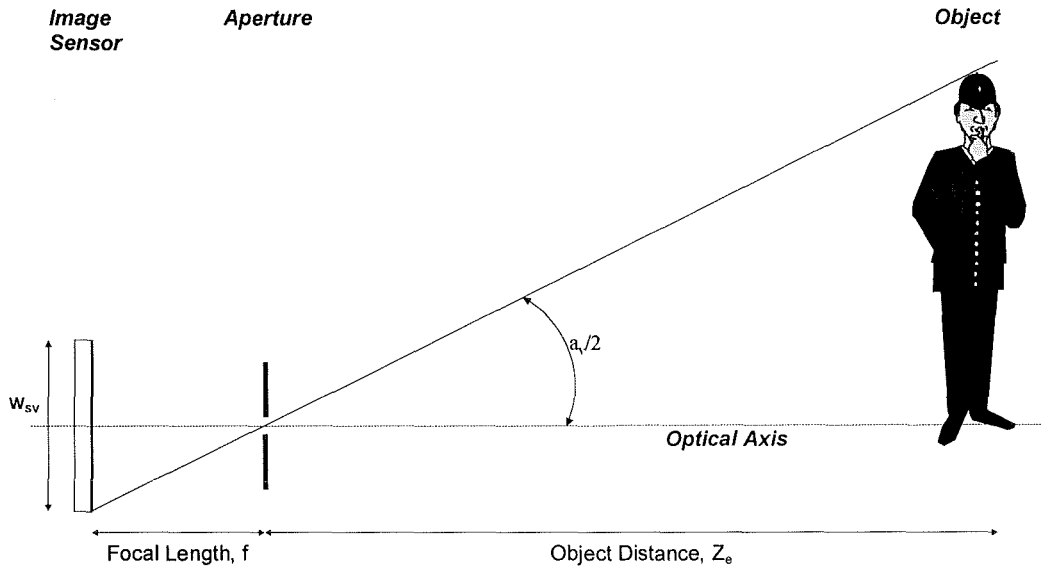


Figure 55: *Perspective transformation by Back Projection*

It is evident from the figure above that for a point at the limit of the sensor's field-of-view:

$$\frac{w_{sv}}{2.f} = Tan\left(\frac{a_v}{2}\right)$$

Note the above makes use of the far field (or large magnification) assumption that $z_e >> f$ such that it is immaterial whether z_e is defined with respect to the lens or to the image sensor position. Using this assumption means that as the sensor width is known in units of pixels the scaling conversion from units of metres to units of pixels in the image plane can be achieved without requiring knowledge of the actual focal length and sensor pixel dimensions in metres. The parameter w_{sv} can therefore be expressed in pixels as the sensor's resolution, r.

Rearranging the above equation the focal length can therefore be calculated (in pixels) in terms of the field-of-view a_v and the resolution, r parameters of the camera - both of which are known.

$$f = \frac{r}{2.Tan\left(\frac{a_v}{2}\right)} \text{ pixels}$$

Once f is known then by similar triangles the position of a point in the image plane due to a world point known in eye co-ordinates can be calculated from:

$$y_{pp} = -\frac{f.y_e}{z_e}$$

The negation can be ignored if it assumed that the camera is mounted (upside down) such that an upright image is obtained. The above applies with respect to both the x and y co-ordinate transformations. However using a sensor with non-rectangular pixels or operating differing subsampling factors in the x and y axes means that differing scale factors may need to be applied for each plane. This is accommodated by expressing this as a difference in the focal lengths f_x and f_y for the two planes although, of course, they both have the same physical focal length.

To perform the perspective transformation it is therefore only necessary to know the sensor resolution and field of view for each of the x and y image dimensions. These results can be expressed in matrix form as a perspective transformation matrix, P so that:

$$\underline{x_{pp}} = \begin{bmatrix} f_x & 0 & 0 & 0 \\ 0 & f_y & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \underline{x_e} \text{ where } \underline{x_{pp}} = \begin{bmatrix} x_{pp} \\ y_{pp} \\ z_{pp} \\ 1 \end{bmatrix}$$

The vector $\underline{x_{pp}}$ being two-dimensional is now expressed in superhomogenous co-ordinates. This result is then incorporated into the transformation process to find the relationship between image co-ordinates and world co-ordinates, as shown below:

$$\underline{x_{pp}} = P.R_y.R_x.T.\underline{x_w}$$

14.2.2.3 Transformation from image plane co-ordinates to graphical display co-ordinates

A further transformation D is required as the display co-ordinates are defined with the co-ordinate origin at the top left-hand corner of the screen with the y co-ordinate increasing down the screen. The image co-ordinates used above are however based around an origin at the centre of the image on the camera's optical axis. It is therefore necessary to first reflect about the y-axis and then to offset both x and y axes by half

the image resolution. The co-ordinates are now expressed as 2D homogeneous co-ordinates giving.

$$\underline{x_d} = \begin{bmatrix} 1 & 0 & r_x / 2 \\ 0 & -1 & r_y / 2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{pp} \\ y_{pp} \\ 1 \end{bmatrix} \text{ where } \underline{x_d} = \begin{bmatrix} x_d \\ y_d \\ 1 \end{bmatrix}$$

The final transformation is therefore:

$$\underline{x_d} = D.P.R_x.R_y.T.\underline{x_w}$$

14.2.2.4 Inverse View Point Transformation

When testing a hypothesis that a pedestrian is present at particular pixel position in the image it is necessary to perform the inverse of the above transformation process to find the world co-ordinates of the hypothesised pedestrian. This information is then used to project the model into the scene. The objective in this section is therefore to find out the position of a given a point in the image in world co-ordinates. This process is however under-constrained as each point on the image plane may correspond to any point on a line in 3D space. To achieve the inversion an additional constraint known as the ground plane assumption is therefore invoked i.e. it is assumed that all objects under observation lie on a flat horizontal ground plane. In this case this amounts to an assumption that the y co-ordinate in the world system is zero i.e. $y_w = 0$. The inversion process therefore involves finding the intersection of the 3D line in space defined by an image co-ordinate with this ground plane.

Firstly the inverse of the transformation between the eye and world co-ordinate systems is found. Rather than calculate the inverse of the combined transformation, of each stage of the transformation, it is simpler to form the inverse of each stage of the transformation. This is achieved by performing the relevant rotations through the reverse angle (and similarly for the translation) and combining the results by matrix multiplication as shown below.

$$\underline{x_w} = \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & H \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} CosB & 0 & SinB & 0 \\ 0 & 1 & 0 & 0 \\ -SinB & 0 & CosB & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & SinA & -CosA & 0 \\ 0 & CosA & SinA & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \underline{x_e}$$

$$\underline{x_w} = \begin{bmatrix} CosB & CosASinB & SinASinB & 0 \\ 0 & SinA & -CosA & H \\ -SinB & CosACosB & SinACosB & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} . \underline{x_e}$$

$$\underline{x_w} = T^{-1} . R_x{}^{-1} . R_y{}^{-1} . \underline{x_e}$$

expanding gives:

$$x_w = x_e . CosB + y_e . CosA . SinB + z_e . SinA . SinB$$

$$y_w = y_e . SinA - z_e . CosA . + H$$

$$z_w = -x_e . SinB + y_e . CosA . CosB + z_e . SinA . CosB$$

using the ground plane constraint of $y_w = 0$ and reversing the perspective transform means that x_e, y_e can then be expressed in terms of z_e . z_e can in turn be found by setting the second equation to zero.

$$x_e = \frac{z_e . x_{pp}}{f_x}, \quad y_e = \frac{z_e . y_{pp}}{f_y}$$

$$z_e = \frac{-H}{\left[\frac{y_{pp} . SinA}{f_y} - CosA \right]}$$

Substituting back gives the required values for x_w and z_w:

$$x_w = z_e . \left[\frac{x_{pp} . CosB}{f_x} + \frac{y_{pp} . CosA . SinB}{f_y} + SinA . SinB \right]$$

$$z_w = z_e . \left[\frac{-x_{pp} . SinB}{f_x} + \frac{y_{pp} . CosA . CosB}{f_y} + SinA . CosB \right]$$

The perspective projected co-ordinates (x_{pp}, y_{pp}) used above can be found from the image display co-ordinates by inverting the display transform via:

$$x_{pp} = x_d - \frac{r_x}{2} \qquad \text{and} \qquad y_{pp} = -y_d + \frac{r_y}{2}$$

The calculation is left in the expanded form shown i.e. without full substitution of variables) in the interest of clarity.

14.2.2.5 Calculation of Transformations

It was observed that the translation and rotation parameters only changed with a change of viewpoint, which should be determined or changed only during installation. Computation was therefore reduced by calculating the transformation matix coefficients just once, during installation, after which they were stored (as scaled integers) ready for use during execution of the detection algorithm.

Further, it is a property of the above perspective transformations that all straight lines map to straight lines in the image plane. Therefore in forward projecting a straight line it was only necessary to calculate the position of the vertices at the line's end points. The projection of all points between the end points could then be constructed by simply drawing a line between their transformed vertices in the image plane.

14.2.2.6 Simplified Pedestrian Model

The bounding box model described was further simplified to aid in the implementation of the evidence-integrating search. The pedestrian model was required to incorporate variation in scale and viewpoint whilst searching the image for potential pedestrians as part of the peripheral vision process. However there were difficulties in using the entire box model such as deciding which surfaces of its surfaces were visible for a given image position and then scanning over each of these surfaces. As this scanning of each surface has to be repeated for a possible pedestrian on each image pixel covering the detection zone it was the slowest part of the peripheral algorithm and any additional computation had a large effect on execution speed.

Figure 56: *An image pair overlaid with a calibration grid and projections of the pedestrian models. They illustrate the simplification of the original rectangular pedestrian model (left) to a diagonal slice (right) taken from the pair of vertices at the bottom-front to those at the rear-top.*

It was therefore decided to use a simplified model as part of the peripheral process consisting of a single diagonal slice through the model from the bottom at the front up to the top at the rear. The justification for this choice was that, given the relative positions of the camera and scene, in the distance it would approach a side-on view whilst when viewed from above it would tend towards a plan view thus giving useful results over the entire field-of-view. Being a simple 2D-slice calculation of the scanning parameters was consequently greatly reduced.