Pedestrian Detection for Underground Mine Vehicles using Thermal Imaging

John Simon Dickens

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, in fulfilment of the requirements for the degree of Masters of Science in Engineering

Johannesburg, 2012

Declaration

I declare that this dissertation is my own unaided work. It is being submitted for the degree of Master of Science in Engineering at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination to any other University.

....
John Dickens

Signed on the .... day of 2012

Abstract

Vehicle accidents are one of the major causes of deaths in South African underground mines. A computer vision-based pedestrian detection and tracking system is presented in this research that will assist locomotive drivers in operating their vehicles safer. The detection and tracking system uses a combination of thermal and three-dimensional (3D) imagery for the detection and tracking of people. The developed system uses a segment-classify-track methodology which eliminates computationally expensive multi-scale classification. A minimum error thresholding algorithm for segmentation is shown to be effective in a wide range of environments with temperature up to 26 $^\circ C$ and in a 1000 m deep mine. The classifier uses a principle component analysis and support vector classifier to achieve a 95% accuracy and 97% specificity in classifying the segmented images. It is shown that each detection is not independent of the previous but the probability of missing two detections in a row is 0.6%, which is considered acceptably low. The tracker uses the Kinect's structured-light 3D sensor for tracking the identified people. It is shown that the useful range of the Kinect is insufficient to provide timeous warning of a collision. The error in the Kinect depth, measurements increases quadratically with depth resulting in very noisy velocity estimates at longer ranges. The use of the Kinect for the tracker demonstrates the principle of the tracker but due to budgetary constraints the replacement of the Kinect with a long range sensor remains future work.

Acknowledgements

I would like to thank my supervisor, Prof. Anton van Wyk for his support and for sharing his knowledge on a broad range of fields. I would like to thank my co-supervisor Jeremy Green for always having an open door and taking the time to discuss my research.

I would like to thank the following mines for allowing me access for testing; Gold Reef City mine, Anglo Platinum's Thembelani mine and First Uranium's Ezulwini mine.

Finally I would like to thank my family for their support throughout all of my studies.

Contents

Chapter 1

Introduction

Transportation machinery is responsible for a large portion of mine deaths in South Africa. Being run over or crushed by transportation equipment is the second largest cause of mine deaths after rock-fall-related incidents. A reliable system for detecting people near mining vehicles is needed to prevent collisions between vehicles and personnel. This research develops a thermal imaging-based pedestrian detection and tracking system for underground mine vehicles.

1.1 Problem Statement

The South African mining industry has committed itself to reducing the vast majority of serious mine accidents by 2013 to be in line with global standards and to continue to strive for zero harm [1, 2]. Apart from the obvious societal impacts of the death of a miner, mining deaths support the perception of mining as very dangerous. In addition to negative industry perception and social impact, mine deaths also cost the industry a significant amount. It is estimated that each death costs R12-million [1] in lost productivity, training costs, insurance and the cost of looking after the deceased's family. So taking the number of deaths for 2011 (112) into account, fatalities cost the mining industry R1.3-billion in 2011. The reduction in mine fatalities is essential for the South African mining industry; however, given that the number of mining fatalities up to November 2011 was 112 [3], the goal of zero harm is not going to be achieved without significant improvements in mine safety systems.

Figure 1.1 shows that after falls of ground, which are rock falls occurring in a mine, vehicles are the second leading cause of mining fatalities. In order to reduce the number of vehicle-related fatalities a system that can detect people near a mine vehicle is needed. Proximity detection systems currently exist but the level of adoption of the technology is low. The low acceptance of current systems is believed to be in part due to fears of nuisance alarms and non-compliance with alarms. Current proximity detection systems are purely proximity-based; they will alarm if someone is within a certain distance of the vehicle whether they are in danger or not, which will most likely result in decreased compliance with alarms.

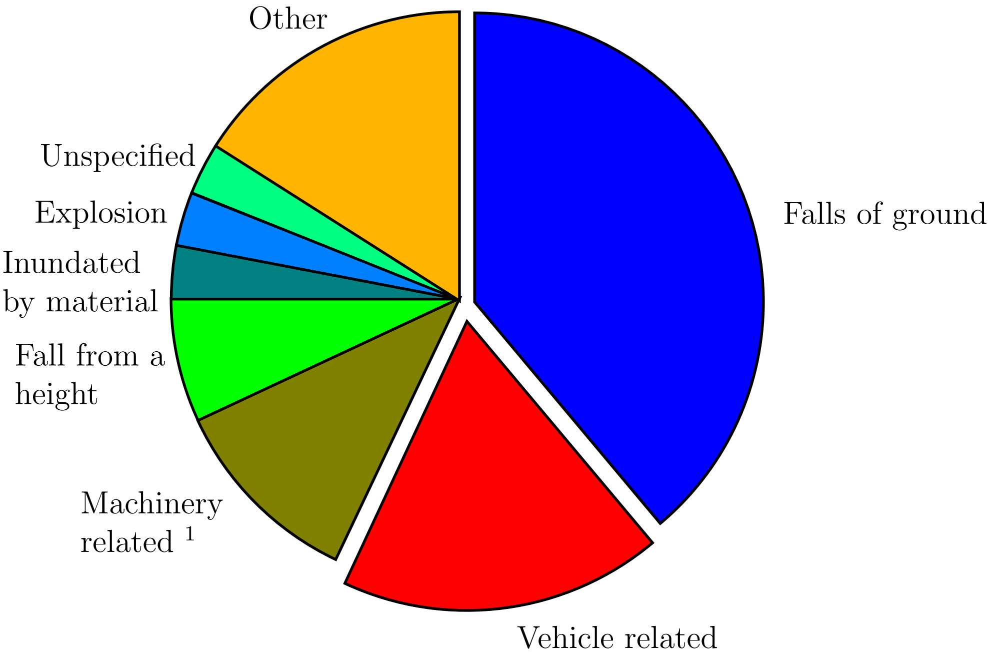


Figure 1.1: A comparison of causes of mine deaths between May 2005 and March 2010 (compiled from DMR fatality reports for both underground and surface mines)

Based on data compiled from the Department of Mineral Resources' (DMR) provisional fatality reports, vehicle-related fatalities account for approximately 18% of mine fatalities. There were 837 fatalities between May 2005 and March 2010, equating to an average of 157 fatalities per year (see *Appendix* A). A pedestrian detection system that can just detect people in front of a vehicle has the potential to significantly reduce the number of vehicle-related fatalities. Such a detection system could reduce frontal collisions with workers, accidents involving two vehicles and accidents where a vehicle hits an object or traps a worker against an object.

[1] Machinery is all mining equipment that is stationary or not self-propelled, such as scrapers, conventional drills, crushers etc.

A number of mine locomotives have the driver sitting behind a fairly small window which restricts the drivers field of view. This may be a contributing factor to some locomotive related fatalities. *Figure* 1.2 shows an example of an underground battery locomotive with an off-centre cabin, it is easy to see from the image that the driver would have a restricted view to the left of the vehicle. The thermal imaging-based pedestrian detection system would be mounted on the from on the locomotive in a position that would allow it to see where the driver could not.

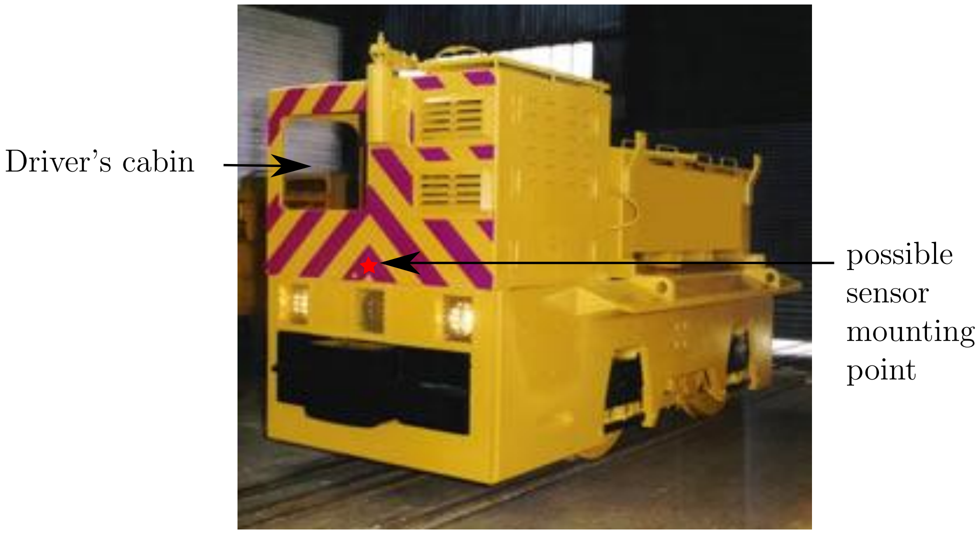


Figure 1.2: An example of an underground locomotive with limited visibility for the driver and a possible mounting point for the thermal imaging-based pedestrian detection system (Adapted from [4])

Underground mines, in particular gold mines, are responsible for the majority of South Africa's mine fatalities and are therefore the nominated environment for the collision avoidance system. A division of fatalities by mining sector is shown in *Figure* 1.3 and it is evident that the majority of fatalities occur in gold mines. The miner detection system will be aimed at gold mines due to the high proportion of deaths in this sector; however the system will be able to operate in other underground mines as well.

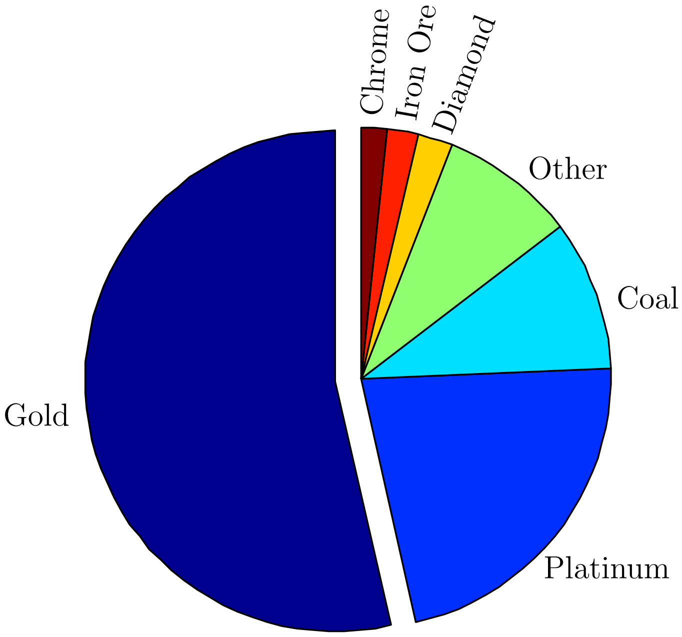


Figure 1.3: A comparison of the percentage of deaths occurring in each mining sector between May 2005 and March 2010

1.2 Background

There are a number of systems that detect the presence of personnel and other vehicles in the vicinity of mining vehicles. These systems have a number of limitations and a thermal imaging-based pedestrian detection system is proposed in this work that is believed to address the limitations of the current systems.

The following sections will discuss the current proximity detection systems and then provide some background on thermal imaging and the use of machine vision for proximity detection. The background will focus on the dark, Global Positioning Systems (GPS) deprived underground environment and more specifically the environment of South African hard-rock (gold and platinum) mines.

1.2.1 Proximity Detection

Existing proximity detection systems use radar, GPS, Radio Frequency Identification (RFID) tags, ultrasonics, lasers, cameras or some combination of these technologies to detect workers near mine vehicles.

Radar-based proximity detection has been used for surface mining equipment as an aid to surface dump truck drivers for detecting people and small vehicles behind the truck. The system is fairly effective, with only occasional false alarms [5]. The close proximity of mine walls in an underground setting provides a challenge for radar implementation, with false alarms being frequent [6].

A GPS-based proximity warning system requires all surface personnel, and vehicles, to carry a GPS receiver and a radio for communication with nearby vehicles. Each vehicle and pedestrian broadcasts its position to vehicles in the area and a display in the vehicle shows the position of nearby vehicles and stationary objects. The system alarms if there is an obstacle within a predetermined range [5, 7]. The reliance of the system on satellite signals precludes it from operating in a GPS-deprived underground environment.

RFID tags are popular for collision avoidance systems and a number of systems operating at various frequencies are commercially available. There are a number of commercial collision avoidance systems, most of which use some form of RFID system.

The Becker NCS Collision Avoidance System (CAS) uses a combination of three proximity detection systems concurrently for improved reliability [8]. They use ultra high frequency (UHF) RFID for long-range detection of up to 100 m, a time-of-flight (TOF) RFID for distances of up to 50 m and an unspecified close proximity electromagnetic detection system that is unaffected by metallic objects (most likely a low frequency magnetic field). *Figure* 1.4 shows a vehicle equipped with four detectors; a vehicle with multiple detectors can tell the operator the approximate direction the detected person is approaching from. The critical zone, in *Figure* 1.4, would sound a high level alarm or apply the vehicle brakes. In a surface mine it is easy to keep vehicles and pedestrians more than 5 m apart. This is also possible in certain underground mines where the mined-out voids are large, for example coal mines. In hard-rock mines it is typical to have pedestrian travelling ways next to haulage rails in one tunnel; in a case such as this any person that a haulage locomotive passes would set off a high level alarm.

The HazardAvert® proximity detection system and the Nautilus International Buddy system use low frequency fields [9, 10]. The HazardAvert® system uses a number of zones surrounding the vehicle. The simplest set-up has a single stop zone around the vehicle (*Figure* 1.5); any workers within the zone cause the machine to stop. Around the stop zone is a warning zone that can alert the operator and slow the vehicle; a typical warning zone range is between 16 m and 23 m. The system can dynamically adjust the size of the zones surrounding the vehicle based on its speed. The HazardAvert® system also allows silent zones to be created which allow, for example, the

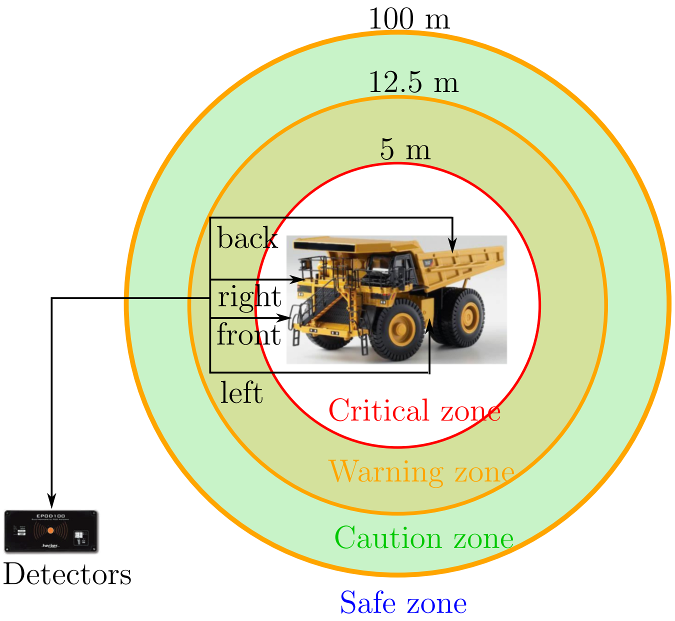


Figure 1.4: The proximity detection zones of the Becker CAS (after [8])

vehicle operator to be within the stop zone without the vehicle stopping.

Minecom's Dynamic Anti Collision System (DACS600) uses RFID tags operating in the 400 MHz frequency range [9, 11]. The DACS600 searches for tags within range and an alarm is set off when a tag is detected. The system has a display that indicates the number of people and vehicles within a single predefined proximity limit.

The Jannatec Advanced Warning System (JAWS) uses radio transceivers fitted to a vehicle to detect the presence of another transceiver carried by a miner, or fitted to another vehicle or fixed hazard [12]. When a hazard is identified within range the system indicates what type of hazard has been detected so appropriate action can be taken, such as giving way to haulage vehicles.

All the RFID systems require some form of transceiver to be carried by personnel, adding to the cost of the systems. The use of tags also raises maintenance issues. A typical gold mine can have 25 000 workers [13], which means that there may be as many receivers that need to be checked and maintained. The actual number of tags may be lower due to the fact that some there are multiple shifts and some of the workers may work in surface plants, however the number of tags is still may thousands. The reliance of RFID-based systems on the tag means that anyone wearing the wrong equipment or whose tag is faulty will be invisible to the system, which is a potentially very dangerous situation. Some tag-based systems only detect the presence

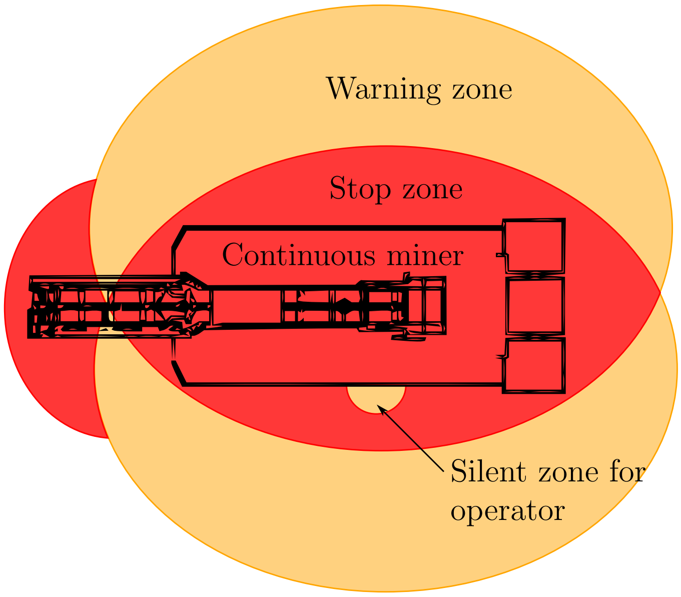


Figure 1.5: The proximity detection zones of the HazardAvert® system

of a tag and not the distance to it and none provide the exact location of the personnel, only how close they are and possibly an approximate direction. It would be advantageous to know where the danger is so that an operator can take appropriate action. It would also be advantageous if the detection system could be integrated into an autonomous system. Automation has the potential to increase safety and productivity [14–16], as well as allow the mining of resources that could otherwise not be mined [17]. An automated system requires knowledge of actual positions of people and obstacles so that it can plan an appropriate response.

1.2.2 Machine Vision

Machine vision provides an accurate way of detecting people and determining exactly where they are in relation to machinery or vehicles [5, 18]. However, visible detection systems can be sensitive to changes in illumination or being obscured by dust or smoke. In a badly lit underground mine, a system using visible-light imaging would require its own light source in order to operate reliably. Using a light source with the camera also creates changing illumination since the illumination intensity increases closer to the light source. Changing illumination is a problem for a number of computer vision algorithms. An additional disadvantage of visible-light imaging is that if a miner points their cap-lamp directly at the camera the intense, focussed beam is sufficient to saturate a visible-light camera. Thermal-infrared is, however, not significantly affected by these problems because the illumination is radiated by people and the long wavelength (7-14 μm) allows it to penetrate dust and smoke [19]. These advantages make thermal imaging-based machine vision an attractive alternative for detecting people near underground mine vehicles.

1.2.3 Thermal Imaging

Every object with a temperature above absolute zero emits infrared (IR) radiation. Thermal imaging systems create images of a scene based on the thermal radiation received from the scene. The radiation received is due to the thermal radiation emitted by objects in the scene, as well as radiation from other sources reflected off objects. Humans generate heat and have an emissivity close to one. An object with an emissivity of close to one will reflect very little radiation and the radiation emitted by the object will mostly be due to its temperature.

The IR spectrum is a large section of the electromagnetic spectrum with a wide variety of uses depending on the part of the IR spectrum being used. The IR spectrum can be sub-divided into five main regions. These are near-infrared, short-wavelength, mid-wavelength, long-wavelength and far-infrared [20]. Near-infrared (0.7 to 1.4 μm) is commonly used for devices like IR remote controls. Near-infrared illumination is also often used for commercial night-vision surveillance since it can be detected using the same imaging sensor used for visible-light. Near-infrared is also used by TOF cameras because it can still be detected with standard complementary metaloxidesemiconductor (CMOS) or charge-coupled device (CCD) sensors but there is less ambient near-infrared than visible light. Short-wavelength IR (1.4 to 3 μm) is used for various process monitoring and inspection tasks such as hot furnace monitoring [21]. Mid-wavelength IR (3 to 8 μm) is used for gas spectroscopy [20]. Long-wavelength infrared (LWIR) (8 to 14 μm) is the region of interest for this research and is used for thermal imaging. The far infrared spectrum extends beyond 14 μm but the definition of where it ends varies between 50 μm and 1000 μm, depending on the application [22, 23].

Thermal-infrared cameras find a number of military and commercial applications, including surveillance, process automation and printed circuit board (PCB) testing [24, 25]. For particles that are significantly smaller than the wavelength of the radiation, scattering is proportional to $1/\lambda^4$, so for particles less than 2 μm in size LWIR is hardly scattered [22, p. 93]. The lack of scattering by smaller particles allows LWIR to penetrate small parti-

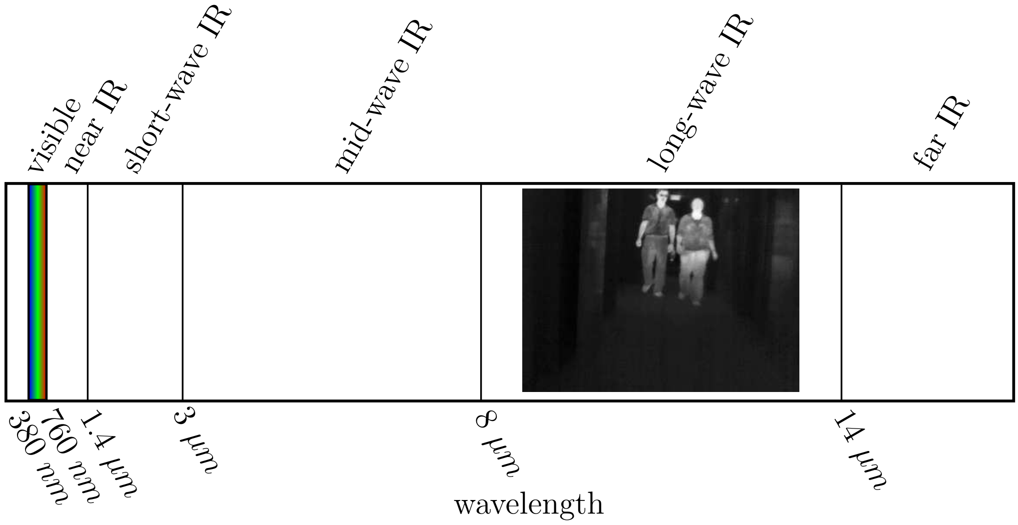


Figure 1.6: A section of the electromagnetic spectrum showing the divisions of the infrared spectrum

cle size dust, smoke and light fog [19]. For large particle sizes such as rain, heavy fog and large particle dust LWIR is scattered similarly to visible-light. The improved dust penetration of LWIR, in comparison to near-IR and visible light imaging, makes it a good choice for the dusty mining environment. LWIR imaging does not require external illumination, which gives it a range advantage over illuminated imaging such as visible or near-infrared systems [19, 26].

The two types of thermal-infrared imaging systems are cooled quantum detectors and uncooled detectors. Owing to the need for cryogenic cooling, quantum detectors are bulky, costly and mostly restricted to military applications [26]. Uncooled detectors are used in the majority of commercial applications. There are two types of uncooled detectors, ferroelectric detectors and microbolometers.

Ferroelectric sensors (a subclass of pyroelectric sensors) make use of the pyroelectric properties of certain materials. Substances that exhibit the pyroelectric effect have an electric polarisation that changes with changing temperature. The changing polarisation will create a changing voltage across the material that can be detected. If a pyroelectric element is exposed to a source of long-wave IR radiation its temperature will rise, which can be detected as a changing voltage across the material. Since pyroelectric sensors can only detect a change in temperature (a change in incoming radiation), the radiation must be chopped [24]. In pyroelectric IR imaging systems the chopping of the radiation is achieved using a rotating chopper. Image capture starts with the chopper closed and the measurement of the potential of each pixel. The chopper rotates to expose the Focal Plane Array (FPA), the open chopper admits IR radiation from the scene. The IR radiation is absorbed by the pixels of the FPA, causing their temperature to change. The change in temperature of each pixel causes a change in the potential across each pyroelectric element. The electronics will measure the new potential across the pyroelectric pixels and subtract the value when the shutter was closed. The change in the potential is proportional to the change in temperature of the detecting element, which is proportional to the amount of incoming radiation.

One of the issues with pyroelectric sensors is that the shutter is not a perfect thermal insulator and this causes the halo effect identified as a problem with thermal imaging [27–29]. If there is a hot object in the scene the chopper heats up and some of the energy gets to the sensor. Since the energy leaking through the chopper is not focused a region larger than the warm object is imprinted on the FPA. Since the FPA is warmed when the shutter is closed the increase in temperature due to the illumination from the scene is reduced; this creates a dark halo around warm objects. Most thermal imaging systems operate at 30 Hz so a typical chopper would rotate at 30 Hz (thermal cameras made in the U.S. are export controlled and a license is required for 30 Hz cameras outside the U.S.).

Microbolometer FPAs do not suffer from haloing and represent the vast majority of modern thermal imaging sensors. While in the past ferroelectric barium strontium titanate (BST) represented the majority of thermal imaging sensors [29], ferroelectric FPAs are now rarely used. For a number of reasons BST sensors have almost universally been replaced with microbolometers. Production of the last commercial camera using a BST sensor ended in 2009 [30].

A microbolometer consists of an infrared-absorbing material and a negative temperature coefficient resistive element thermally insulated in a vacuum [26, 31]. The housing in which the microbolometer is insulated has an infrared-transparent window that allows radiation from the scene to strike the infrared-absorbent material. The temperature change of the microbolometer, due to infrared energy absorption, is determined by measuring the resistance change of the resistive element. The temperature of the object being imaged can be calculated using the Stefan-Boltzmann law, assuming a certain emissivity.

IR image interpretation poses a number of challenges. Thermal cameras have relatively low resolution, contrast and signal to noise ratio [32, 33]. Typically the resolution of thermal cameras is 320×240 but 640×480 resolution cameras are available at significantly higher cost. An original equipment manufacturer (OEM) thermal camera core with a 320 × 240 resolution costs approximately R30 000 while a similar camera with a 640 × 480 resolution costs approximately R80 000. Even high resolution thermal cameras do not come close to the multiple mega-pixel resolutions easily available from visible light cameras. Since objects at similar temperatures have similar thermal radiance, the contrast and therefore the number of identifiable image features is relatively small.

1.2.4 Mining

Various mining methods and associated mine layouts are used to exploit different minerals. The pedestrian detection system is designed for hard-rock mines such as gold and platinum, with a focus on gold mines, so the layout of a typical gold mine will be introduced.

To begin the discussion on underground mines, some basic mining terminology is outlined below [34, 35].

- Ore: a mineral deposit that has sufficient value to be mined profitably
- Waste: material associated with an ore deposit that must be mined to get to the ore and is then discarded
- Stope: the working area in an underground mine from which the ore is extracted
- Reef: a body of rock containing the ore
- Haulage: The horizontal transport of ore, supplies, and waste, or the passageway for such transport
- Cross-cut: a horizontal tunnel going from haulage tunnels to intersect the reef
- Raise: excavations that follow the angle of the reef, from which mining of the reef proceeds
- Dip: a direction parallel to the incline of a reef
- Strike: a direction perpendicular to the incline of the reef
- Hanging-wall: the ‘roof’ of a tunnel or stope
- Foot-wall: the ‘floor’ of a tunnel or stope

- Pillar: a section of rock left to support the hanging-wall of a mined-out void
- Bar-down: To prise down a loose rock from the hanging-wall

A shaft is sunk vertically from the surface and horizontal haulage tunnels are excavated at various levels. The levels are between 60 m and 200 m apart; on each level cross-cuts are excavated horizontally from the haulage tunnels to intersect the reef. When the cross-cut meets the reef a raise is tunnelled that follows the angle of the reef ($Figure$ 1.7).

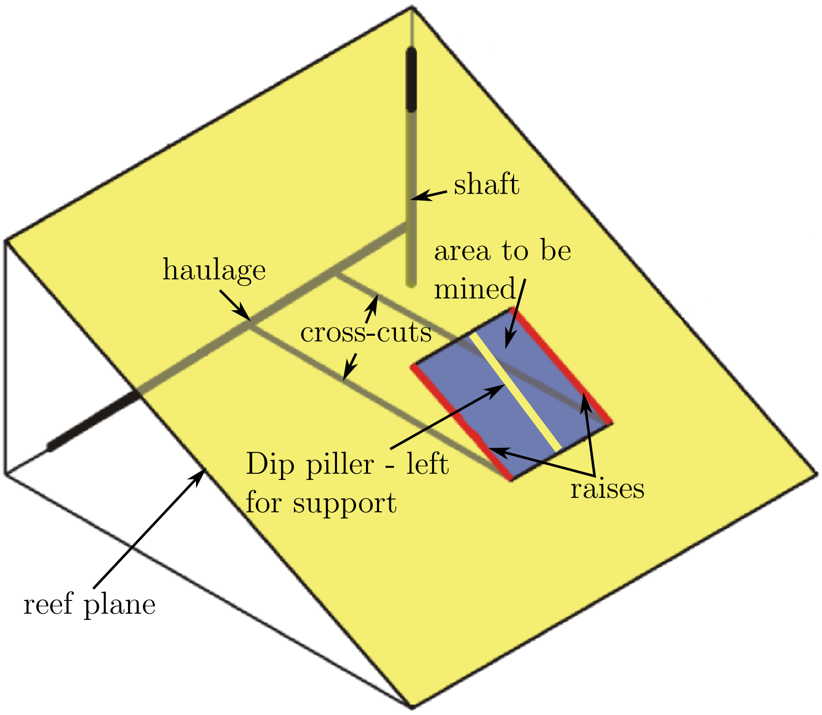


Figure 1.7: A figure showing a typical gold mine layout, adapted from [36]

Mining proceeds from two raises inwards along the strike in stoping panels, as shown in $Figure$ 1.8. Mining usually stops just before the panels meet in the middle, leaving a dip pillar to support the hanging-wall.

Mining in hard-rock mines is predominately cyclical, consisting of a drill-blast-clean cycle. The cycle starts with drilling of blast holes in the stope face. After all the blast holes are drilled and filled with explosives, all the workers exit the mine and the explosives are detonated. After the explosives' fumes dissipate and seismicity decreases back to background levels, the mine workers re-enter and begin the process of entry inspection and support installation. Entry inspection involves checking the hanging-wall of the stope

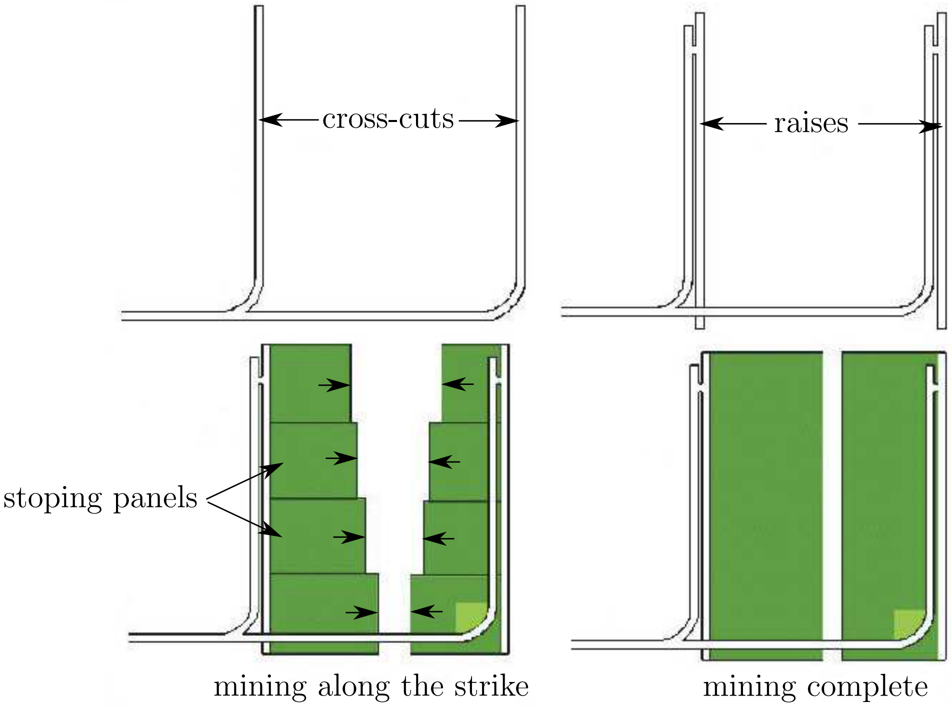


Figure 1.8: The progression of mining operations, after [36]

for loose rocks and either supporting or barring-down those rocks that are found to be unstable. After entry inspection the broken ore from the last blast is removed from the stopes using scrapers and then loaded into haulage equipment. In gold mines the haulage equipment is usually rail-bound and the haulage locomotive transports the ore to a tip point where it falls to the bottom level of the mine and is then hoisted to the surface for processing.

The underground environmental conditions are an important consideration for this research. The thermal conditions in a mine are central to the operation of a thermal imaging-based system. The virgin rock temperatures of deep South African gold mines can be up to approximately 60 $^{\circ}C$; however ventilation and other cooling (eg. evaporative cooling) reduces the temperature within working areas to below 30 $^{\circ}C$ [37]. Work conducted to model the heat flow from advancing stopes shows that the rock surface temperature can be assumed to be equal to the ventilation air wet-bulb temperature (T_{wb}) [38]. Significant work has been performed to design ventilation systems to ensure that the air T_{wb} remains below 28 $^{\circ}C$ (heat stress management programmes are required for $T_{wb} > 27.5$ $^{\circ}C$) [39, 40].

The human body surface temperature will always be higher than the ambient air wet-bulb temperature. The air wet-bulb temperature takes into account evaporative cooling; therefore in order for a person to dissipate metabolic heat their surface temperature (on average) must be higher than the ambient air wet-bulb temperature. Knowing that the tunnel walls have a surface temperature equal to the wet-bulb temperature of the air, and that people will have a temperature higher than the air wet-bulb temperature, we can infer that people will have an average surface temperature that is higher than the surroundings. This fact will be used for detecting people in the thermal images.

1.3 Research Objective

The objective of this research is the design, testing and evaluation of a pedestrian detection system for underground mining vehicles. The collision avoidance system is designed to detect people using a combination of a thermal-infrared camera and a depth sensor. The identified people are tracked over a number of frames and using the position measurements from the depth sensor their current trajectories are estimated. The trajectories of the people relative to the vehicle are used to determine whether a collision is likely to occur and if so when.

This research evaluates whether thermal imaging-based human detection can detect workers in an underground mine with sufficient accuracy to prevent collisions with a vehicle. The system is tested in an indoor environment as well as in an actual underground mine, albeit not mounted on a mine vehicle.

Obviously 100% detection is the goal; however this is very unlikely to be achieved by any system. To decide on the required accuracy the detection system will be compared to detection systems representing the current best practice. The most popular systems for underground pedestrian detection are RFID-based systems. The actual detection rates of RFID systems for pedestrian detection have not been adequately determined. Research on RFID systems for other applications has shown that the detection rates depend on a number of factors and may not be as high as claimed by manufacturers [41].

Work on predicting detection rates of RFID tags on vehicles shows that the detection rates depend significantly on the speed of the vehicle, angle of the receiver and the position of the tag on the vehicle window [42]. A study performed by Clarke et al. [41] on RFID detection rate on shipping containers showed that detection rates depended on the tag orientation and the contents of the package. The work by Clarke et al. showed that even under ideal conditions the highest detection rate was approximately 99%. In the abscence of any information on proximity warning RFID systems, a specificity of 99% or higher will be considered sufficient.

In order to prevent a collision the system will need to provide sufficient time to stop the vehicle or warn the pedestrian. The average reaction time of locomotive drivers has been found to be approximately 2 s [43]; therefore sufficient warning will be assumed to be greater than 2 s.

1.4 Scope and Limitations

This research looks at the design and evaluation of a pedestrian detection system that can detect people in the path of a forward-moving underground mine locomotive. The scope is limited to underground locomotives because it is believed that a machine vision detection system would be most useful on a locomotive. Since a locomotive cannot turn around sharp corners, it cannot turn into an area that is not visible to the camera. The tracks of a locomotive ensure that its trajectory remains smooth, which is necessary to predict a collision.

There are a number of causes of vehicle-related accidents in mines, as shown in *Appendix* A. A large portion of these causes of vehicle-related deaths is unknown. Of the known causes the main cause is a vehicle driving into someone or something.

While it would be advantageous if the detection system could be interfaced with autonomous mine vehicles in the future, the system is intended to assist a human vehicle operator. The research is restricted to the design of the system software and testing using a prototype hardware set-up. Specific hardware considerations, such as the design of an intrinsically safe system for use in a mine containing flammable gases, are excluded.

Only limited range 3D sensors (5 m) are available for this system; the ranges are insufficient to provide adequate warning to the pedestrian or driver but the principles remain the same. The addition of a longer-range sensor will remain as future work. The underground speed limit is generally 16 km/h (4.4 m/s), with faster speed requiring special arrangement. In order to provide the required 2 s warning the system will need to be able to determine the trajectory of a person up to 9 m from the vehicle. The available hardware is constrained to a thermal camera and a Kinect or SR4000 time-of-flight camera.

The available mines to test the system were less than 1000 m deep with ambient air temperatures below 26 $°C$. A number of South African go significantly deeper (deeper than 3000 m, [44]) but for the purpose of this research the scope will be limited to shallower, cooler mines. The system will be limited to operate in environments with air temperatures up to 26 $°C$.

1.5 Assumptions

It is assumed that the temperature of the mine walls will always be measurably different to human body surface temperature. As discussed in *Section* 1.2.4 the average body temperature is greater than the air temperature; however in the presence of high velocity ventilation air this difference may be small, due to the increased cooling effect of the fast moving air. It is assumed that the system will be used in areas with moderate ventilation.

1.6 Overview

Chapter 2 discusses some of the relevant methods for thermal image segmentation, classification and camera-based object tracking. This is followed, in *Chapter* 3, by a discussion of the preliminary system design and the methods evaluated for use in the system. The final system is described in *Chapter* 4 and the results of the system testing are outlined in *Chapter* 5. The important conclusions are summarised in *Chapter* 6 and possible future work is suggested.

Parts of this work was published at three conferences. The three conference papers:

- J. S. Dickens, J. J. Green and M. A. van Wyk, "Human Detection for Underground Autonomous Mine Vehicles Using Thermal Imaging," *26th International Conference of CAD/CAM, Robotics and Factories of the Future*, Kuala Lumpur, Malaysia, July 2011.

- J. S. Dickens, J. J. Green and M. A. van Wyk,"Pedestrian detection for underground mine vehicles using thermal images," *IEEE Africon*, Livingston, Zambia, September 2011.

- J. S. Dickens and J. J. Green, "Segmentation techniques for extracting humans from thermal images," *4th Robotics and Mechatronics Conference of South Africa*, Pretoria, South Africa, November 2011.

Copies of the papers are attached as *Appendix* D.

Chapter 2

Literature Survey

The system will be required to detect, classify and track humans in thermal images. It will need to determine how far away from the vehicle the people are and track them to determine whether they are on a collision course with the vehicle. There are a number of actions that will need to be performed for the collision avoidance system to function.

A commonly used paradigm for object detection and tracking in video is first to extract regions of interest and then classify or validate them [32, 45–50]. This is the approach that has been used for this system for the reasons described in *Section* 2.1. So the first stage is segmentation of the image into candidate regions that are then classified.

The second stage is to classify the extracted segments to remove regions that do not represent humans. Regions that are of interest (humans) are tracked over a number of frames and a trajectory for each object is estimated. The true 3D trajectory of the objects is calculated using a 3D sensor which is calibrated to the thermal camera. Using the calibration, the human identified in the thermal image will correspond to a certain region of the point cloud produced by the 3D sensor.

2.1 Thermal Image Segmentation

Extracting candidate regions for later classification reduces the computation required for the classification step. Without a region of interest the classifier would need to divide the image into regions and determine whether each region contains the object of interest. For objects at various distances from the camera the classifier would need to repeat the process using many different sized blocks.

There are a number of ways of extracting candidate regions for later classification. Segmentation algorithms can basically be divided into two types, static and motion-based. Static algorithms use a single frame and information about the foreground and background objects to segment the image. The information used to segment the image may be intensity, colour, edges etc. This type of segmentation is usually used when there is a specific attribute of the object that can be easily segmented such as tracking a blue ball, or someone wearing a red shirt. Motion-based segmentation makes use of the motion of foreground objects to identify them. Image differencing is the most basic motion-based segmentation method. Two subsequent images are subtracted from each other; the difference will be close to zero everywhere except where objects have moved.

2.1.1 Motion-based Segmentation

Motion-based segmentation for thermal images is used by Dai, Zheng and Li [32] and Fernández-Caballero et al. [25]. Dai, Zheng and Li use a segmentation algorithm that separates the image into a background or still layer and a foreground or moving layer. Once the foreground objects are extracted, they are classified to separate the pedestrians from other foreground objects. The extraction algorithm cannot run in real time since it operates on a number of images. The images are registered and adaptively averaged. The foreground objects are removed from the images based on the estimated background and then the background is re-estimated until convergence.

Fernández-Caballero et al. use an intensity-threshold and motion-based segmentation to detect people in infrared images from a mobile robot. The first segmentation is a static threshold where warmer zones are extracted. The motion-based segmentation uses optical flow or image subtraction. When the robot is stationary, image subtraction is used to segment moving people. When the platform is moving an optical flow-based segmentation is used. A histogram of magnitudes of the optical flow vectors is created. The flow magnitude is thresholded to create an image containing fast-moving objects. While the optical flow method allows the extraction of moving objects with a moving camera, it assumes that the optical flow due to the platform moving is less than due to the moving people.

Motion-based algorithms will only detect moving objects (and usually require a stationary camera) which does not make them useful for detecting people who may be stationary in the path of a moving vehicle.

2.1.2 Static Segmentation

Static segmentation algorithms are used by a number of researchers for segmenting thermal images, some of whose works are described below. Bertozzi et al. [45] use the fact that pedestrians have a high degree of vertical symmetry. They produce a histogram of the symmetry of grey-levels in the image, the symmetry of vertical edges and the density of vertical edges. Regions with high symmetry are determined by thresholding the overall histogram. Since pedestrians are hotter than the background the symmetry is only calculated for hot regions, instead of performing an exhaustive search on the entire image. This also reduces the number of false positives. In small mine tunnels people are unlikely to be walking upright reducing their vertical symmetry. The system is required to detect people in a variety of poses so using symmetry is fairly restrictive.

To account for the changing dynamic range of images, Xu, Liu and Fujimura [50] use a threshold that is a balance between the mean and highest intensity in an image. The threshold (T) chosen by Xu et al. is shown in *Equation* 2.1. If the pixels are uniformly distributed between zero and I_H then the threshold in *Equation* 2.1 will extract the brightest 10% of pixels. Xu et al. perform the thresholding on a histogram-equalised thermal image which provides some basis for the assumption of uniform pixel distribution

$$T = 0.2I_M + 0.8I_H \tag{2.1}$$

Where I_M is the mean intensity of the image and I_H is the maximum intensity in the image.

Thornton, Hoffelder and Morris [49] use a normalised intensity deviation and edge information to determine the regions of interest. The normalised intensity deviation is defined as

$$N\left(x, y\right) = \frac{I\left(x, y\right) - m}{\sigma} \ . \tag{2.2}$$

Where $I\left(x, y\right)$ is the intensity of the pixel (x, y), m is the mean and σ is the standard deviation. The mean and standard deviation are calculated both locally and globally. The edge information is obtained using the gradient magnitude. Using an empirically determined threshold, binary images are created from the local and global deviation images, as well as the edge map.

Nanda and Davis [48] extract regions of interest by thresholding thermal images using a threshold determined from training images. The means and standard deviations of pixels belonging to pedestrians and non-pedestrians are determined. Using Bayes' classification, assuming equal priors for pedestrians and non-pedestrians and a Gaussian distribution, the threshold (T) is given by

$$T = \frac{\sigma_p \sigma_n}{\sigma_p + \sigma_n} \ln \left(\frac{\sigma_p}{\sigma_n} \right) + \frac{\sigma_p \mu_n + \sigma_n \mu_p}{\sigma_p + \sigma_n} \tag{2.3}$$

where
σ_p is standard deviation of pixels belonging to pedestrians
μ_p is mean of pixels belonging to pedestrians
σ_n is standard deviation of pixels belonging to non-pedestrians
μ_n is mean of pixels belonging to non-pedestrians.

Pixels with intensities of greater than the threshold calculated in *Equation* 2.3 are target pixels. Nanda and Davis use the threshold to remove the background pixels but foreground pixels retain their value. Background pixels (pixels with a value of less than the threshold) are equated to zero while foreground pixels are left unchanged.

The segmentation method used by Haritaoglu, Harwood and Davis [47] falls somewhere between a motion-based and static segmentation algorithm. It has features of both methods; Haritaoglu et al. statistically model the background in order to extract foreground objects. Since the background is modelled the camera must remain stationary or the background must be remodelled every time the camera moves. Having a background model allows the segmentation to detect people even if they are stationary and therefore this method is considered a static method. The maximum and minimum intensity, as well as the maximum intensity deviation of each pixel in a scene containing no people, is recorded. The pixel at (x, y) in image I is considered a foreground pixel if

$$|M(x,y) - I(x,y)| > D(x,y) \text{ or } |N(x,y) - I(x,y)| > D(x,y) \tag{2.4}$$

where
$M(x,y)$ is the background maximum of pixel (x,y)
$N(x,y)$ is the background minimum of pixel (x,y)
$D(x,y)$ is the inter-frame difference for pixel (x,y).

After thresholding, a morphological erosion operation is applied to the image to eliminate single pixel noise. A binary connected component operation is applied and small regions are eliminated and then a dilation operation is used to restore the objects to their original size (after the erosion).

A number of methods for segmenting standard grey-scale images exist; one of the most popular is Otsu's method [51]. Otsu's method is based on the assumption that the optimal threshold is the one that maximises the separation between the grey-scale histograms, i.e. it is the threshold that maximises the between-class variance. The derivation of the optimal threshold is shown in *Appendix* B to be

$$k_{opt} = \underset{k}{\operatorname{argmax}}\ \sigma_b^2(k) \tag{2.5}$$

Where k_{opt} is the optimal threshold and $\sigma_b^2(k)$ is between-class variance as a function of the threshold.

A good survey of other grey-scale thresholding methods performed by Sezgin and Sankur [52] identified a number of thresholding techniques which perform well on various grey-scale images. The most effective methods identified by Sezgin and Sankur are those of Kittler and Illingworth [53], Kapur, Sahoo and Wong [54] and Sauvola and Pietikäinen [55].

Let us assume that a picture has a total of N pixels that fall into a total of L grey-levels. The number of pixels that fall into each grey-level (i) of the image histogram is denoted by n_i. The normalised grey-scale histogram can be considered an estimate of the probability distribution of pixel intensities, i.e.

$$p_i = n_i/N \tag{2.6}$$

Where N is the total number of pixels in the image and p_i is the probability that a pixel belongs to the i^{th} grey-level.

The cumulative probability function for the k^{th} grey-level is defined as

$$P(k) = \sum_{i=1}^{k} p_i\ . \tag{2.7}$$

Kittler and Illingworth [53] view the probability density function as an estimate of a mixture population of grey-levels from the foreground and background. For each possible threshold they calculated the value of a criterion function which produces an indirect estimate of the overlap of two Gaussian distributions. The criterion function provides a less computationally complex method of estimating the overlap of the distributions than fitting distributions to the probability density functions. The derivation of the criterion function can be found in [53], giving the final function

$$\begin{aligned} J(k) \;&=\; 1 + 2\left[P(k)\ln\left(\sigma_1(k)\right) + (1 - P(k))\ln\left(\sigma_2(k)\right)\right] \\ &\quad -2\left[P(k)\ln\left(P(k)\right) + (1 - P(k))\ln\left(1 - P(k)\right)\right] \end{aligned} \tag{2.8}$$

Where $\sigma_1(k)$ is the standard deviation of the background up to grey-level k and $\sigma_2(k)$ is the standard deviation of the foreground, from k to L.

The minimum of the criterion function corresponds indirectly to the minimum overlap of the two distributions and therefore the optimal threshold is the one that corresponds to the minimum of the criterion function. As the level of the threshold (k) is varied, the models of the two populations change. The better the fit between the models and the data, the smaller the overlap between the density functions and therefore the smaller the classification error. Using *Equation* 2.8 the optimal threshold can be determined using

$$T_{opt} = \underset{k}{\operatorname{argmin}} \; J(k) \; . \tag{2.9}$$

The entropic thresholding method described by Kapur et al. [54] exploits the entropy of the foreground and background of the image that is thresholded. Maximising the entropy of the thresholded image maximises the information between the foreground and background distributions in the image [52, 54]. It makes sense that the best threshold would be the one where the segmenting into foreground and background retains the most information. For a threshold at grey-level k the entropy of the background up to grey-level k is

$$H_b = -\sum_{i=1}^{k} \frac{p_i}{P(k)} \ln \frac{p_i}{P(k)} \tag{2.10}$$

and the entropy of the foreground is

$$H_f = -\sum_{i=k+1}^{L} \frac{p_i}{(1-P(k))} \ln \frac{p_i}{(1-P(k))} \; . \tag{2.11}$$

Defining the sum of the two entropies as $\psi(k)$ we get

$$\psi(k) = -\sum_{i=1}^{k} \frac{p_i}{P(k)} \ln \frac{p_i}{P(k)} - \sum_{i=k+1}^{L} \frac{p_i}{(1-P(k))} \ln \frac{p_i}{(1-P(k))} \; . \tag{2.12}$$

Maximising $\psi(k)$ gives the maximum total information between the two distributions. So the optimal threshold is

$$T_{opt} = \underset{k}{\operatorname{argmax}} \; \psi(k) \; . \tag{2.13}$$

Sauvola and Pietikäinen [55] propose a thresholding method for binarising text documents. Their method uses an adaptive threshold; instead of having one threshold (T) which is applied to the entire image, the threshold is a matrix the same size as the image ($T(x, y)$). Sauvola and Pietikäinen adapt their threshold-based on the mean and standard deviation of the pixels in a window around each pixel. The threshold is calculated according to the formula

$$T(x,y) = m(x,y) \cdot \left[1 + k\left(\frac{s(x,y)}{R} - 1\right)\right] \tag{2.14}$$

where

$m(x, y)$ is the mean of the window centred on pixel xy
$s(x, y)$ is the standard deviation of the window centred on pixel xy
R is the range of the standard deviation
k is a user-defined constant.

The purpose of the adaptive threshold is to correct for non-uniform illumination and stains on the paper. In badly illuminated regions the mean pixel value in the region will be lower and the threshold will be lowered. In regions with a high standard deviation the threshold will be increased (closer to the mean pixel value), while in uniform regions the threshold will approach $(1 - k)$ times the mean. So in noisy regions the threshold is raised to reduce the thresholding of noisy pixels. At first it would appear that this method should work well for thermal images; changes in the background temperature will change the mean, which will change the threshold. So the threshold adapts itself to a changing background temperature. However, the fact that the people that we are trying to segment are larger than the window used for calculating the mean adversely affects the method. As the window gets towards the centre of a warm person, the mean increases and likewise the threshold and therefore it is only the edges of people that are segmented successfully.

2.2 Classification

Once candidate regions have been segmented, the regions must be verified as human or not. There are a large number of classification methods that could be used for pedestrian detection. Some of the popular methods, especially those used for thermal imaging, are discussed below.

2.2.1 Template Classifiers

Template classification involves the classification of candidates using a template that represents the structure of the object to be classified.

Bertozzi et al. [45] first filter candidates by removing regions that do not meet certain criteria, such as aspect ratio and size constraints and restrictions on a histogram of vertical edges. An example of the filtering constraint is that the histogram of vertical edges of a region cannot be empty in the centre. A region with strong edge and grey-level symmetry but no edges in the centre is likely to be a smooth vertical object such as a pole. Once a candidate has passed all the filtering it is validated through a match with a simple morphological model. The morphological model encodes the shape and average temperature characteristics of the pedestrian. While Bertozzi et al. only use a single model, for a standing pedestrian; multiple models could be used to match humans in a variety of poses.

Nanda and Davis [48] use a probabilistic template to classify pedestrians in thermal infrared videos. A template is created from training images of humans in various orientations but having the same height. Each $m \times n$ (128×48) pixel training image is thresholded and a probabilistic template is created based on the frequency with which each pixel is extracted as a foreground pixel. The probability of a pixel belonging to a pedestrian is calculated using the frequency that it appears as foreground. The 'probability' ($p_c(x, y)$) that the $m \times n$ region around the pixel (x, y) contains a pedestrian is defined to be

$$p_c(x, y) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(t_{xy}(i, j) - 127\right) \cdot \left(p(i, j) - 0.5\right) \qquad (2.15)$$

Where t_{xy} is a $m \times n$ window around pixel (x, y) in the thresholded image and p is the probability that the pixel is from the foreground (i.e. a pedestrian) of the templates. The 127 is due to the thresholded image being an 8 *bit* image with values of 0 and 255. The value $p_c(x, y)$ is not a probability, despite its definition as such by Nanda and Davis, but is a correlation value proportional to the likelihood that there is a pedestrian in the $m \times n$ region around the pixel.

The whole image is tested using *Equation* 2.15 for three different scales to identify pedestrians at various distances from the camera. The 'probability' map is thresholded to find the pedestrians in the image.

Another template-based classification method for identifying pedestrians in thermal imaging is used by Olmeda et al. [56]. Olmeda et al. start with a similar method to Bertozzi et al. They start by identifying warm regions with high vertical symmetry; these regions are thresholded and these candidates are validated using a template similar to Nanda and Davis'. Instead of using a single template, Olmeda et al. use four templates representing people in various stages of a stride, from closed legs to open legs.

2.2.2 Statistical Classifiers

Fehlman and Hinders [27] use three different classifiers to determine the best features to use to classify non-heat-generating objects in unstructured environments. Fehlman and Hinders use a large number of features for classification, such as ambient temperature, emissivity, entropy, energy and homogeneity. One of the classifiers used by Fehlman and Hinders is the Parzen classifier, which is similar to a Bayesian classifier except that it uses a Parzen density estimate to estimate the likelihood function instead of a K nearest neighbour approach. The Parzen density estimate approximates the conditional probability of getting a given feature vector ($\mathbf{D}$) given that the image is of class j (O_j) [27], i.e.

$$P(\mathbf{D}|O_j) = \frac{1}{N_j h^d} \sum_{q=1}^{N_j} H\left(\frac{\mathbf{D} - \mathbf{D_{qj}}}{h}\right) \ . \tag{2.16}$$

H is the Parzen window function

$$H\left(u\right) = \begin{cases} 1 & |u_p| \leq \frac{1}{2} \; p = 1, ..., d \\ 0 & otherwise \end{cases} \tag{2.17}$$

and

$|u_p|$ is the magnitude of the p^{th} dimension of u
D_{qj} is the q^{th} training feature of class j
N_j is the number of feature vectors belonging to class j
d is the dimensionality of the feature space
h is the length of one side of a d dimensional hypercube.

The Parzen classifier uses Bayes' theorem and the Parzen density estimation in *Equation* 2.16 to determine the probability that the image belongs to a certain class given the observed feature vector. The posterior probability given by the Parzen classifier is

$$P(O_j|\mathbf{D}) = \frac{P(\mathbf{D}|O_j)P(O_j)}{P(\mathbf{D})} \tag{2.18}$$

$$= \left[\frac{1}{N_j h^d} \sum_{q=1}^{N_j} H\left(\frac{\mathbf{D} - \mathbf{D_{qj}}}{h}\right)\right] \frac{P(O_j)}{P(\mathbf{D})} \tag{2.19}$$

where $P(O_j)$ is the prior probability of getting an object of class j and $P(\mathbf{D})$ is called the evidence and normalises the posterior probabilities so that they sum to one.

2.2.3 Support Vector Classifiers

Support vector classifiers are classifiers that determine a classification function that separates two linearly separable classes with a plane that has the greatest margin. The margin is the distance between the closest points on either side of the decision plane. Maximising the margin intuitively should maximise the generalisation of the classifier, if the decision boundary is maximally separated from both classes. New data points can be perturbed about the training data and will still remain on the correct side of the classification boundary.

Let us assume we have m training points; each point, $\mathbf{x_i}$, is a point in d dimensional space. Each point belongs to one of two classes and has a label $y_i = \pm 1$. We will therefore have a training dataset in the form

$$\{\mathbf{x_i}, y_i\} \quad \text{where} \quad \mathbf{x_i} \in \Re^d, \quad y_i \in \{-1, 1\}\,, \quad i = 1 \ldots m\ . \tag{2.20}$$

Assuming the two classes are linearly separable; there will be a hyperplane, $\mathbf{w} \cdot \mathbf{x_i} - b = 0$, that will separate the two classes. The classes are separated such that the class with $y_i = 1$ will fall in the space $\mathbf{w} \cdot \mathbf{x_i} - b > 0$, and the class $y_i = -1$ will fall in the space $\mathbf{w} \cdot \mathbf{x_i} - b < 0$. Let the smallest value of $|\mathbf{w} \cdot \mathbf{x_i} - b|$ be δ, i.e. the point closest to the hyperplane is a distance δ from it. The training data can now be described by

$$\mathbf{w} \cdot \mathbf{x_i} - b \geq \delta \quad \text{for} \quad y_i = 1\ , \tag{2.21}$$

$$\mathbf{w} \cdot \mathbf{x_i} - b \leq -\delta \quad \text{for} \quad y_i = -1\ . \tag{2.22}$$

For simplicity we rescale these equations to get

$$\mathbf{w_s} \cdot \mathbf{x_i} - b_s \geq 1 \quad \text{for} \quad y_i = 1\ , \tag{2.23}$$

$$\mathbf{w_s} \cdot \mathbf{x_i} - b_s \leq -1 \quad \text{for} \quad y_i = -1\ . \tag{2.24}$$

It can be seen that the distance between the two planes, the margin, is equal to $2/\left\|\mathbf{w_s}\right\|$. The problem is then to find the values of $\mathbf{w_s}$ and b_s that maximise the margin while still satisfying *Equations* 2.23 and 2.24. Maximising the margin is equivalent to minimising $\left\|\mathbf{w_s}\right\|^2/2$ and *Equations* 2.23 and 2.24 can be combined into a single constraint, so we need to find

$$\min_{\mathbf{w_s}, b_s} \frac{1}{2}\left\|\mathbf{w_s}\right\|^2 \quad \text{s. t.} \quad y_i\left(\mathbf{w_s} \cdot \mathbf{x_i} - b_s\right) - 1 \geq 0 \quad \forall i\ . \tag{2.25}$$

Using a Lagrange multiplier α_i for each inequality constraint yields the Lagrangian equation

$$L_p = \frac{1}{2}\left\|\mathbf{w_s}\right\|^2 - \sum_{i=1}^{m} \alpha_i y_i \left(\mathbf{w_s} \cdot \mathbf{x_i} - b_s\right) + \sum_{i=1}^{m} \alpha_i\ . \tag{2.26}$$

By the Wolfe dual [57], minimising *Equation* 2.26 is equivalent to maximising

$$\begin{aligned} L_d = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \mathbf{x_i} \cdot \mathbf{x_j} \quad \text{s. t.} \\ \sum_{i=1}^{m} y_i \alpha_i = 0, \quad \alpha_i \geq 0 \quad \forall\; i \;. \end{aligned} \tag{2.27}$$

Representing the optimisation as in *Equation* 2.27 allows it to be solved using robust quadratic programming methods. It also allows the easy extension to the use of non-linear discriminant functions. To create a non-linear classifier the dot product in *Equation* 2.27 is replaced by a non-linear kernel function to yield

$$\begin{aligned} L_d = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j K\left(\mathbf{x_i}, \mathbf{x_j}\right) \quad \text{s. t.} \\ \sum_{i=1}^{m} y_i \alpha_i = 0, \quad \alpha_i \geq 0 \quad \forall\; i \end{aligned} \tag{2.28}$$

where K is the kernel function, for example the radial basis function kernel

$$K\left(\mathbf{u}, \mathbf{v}\right) = e^{-\frac{\|\mathbf{u}-\mathbf{v}\|^2}{2\sigma}} \;. \tag{2.29}$$

Support vector classifiers have been successfully used for a number of applications in computer vision and many other fields. Some of the applications involving thermal image classification are discussed below.

Xu et al. [50] use a support vector machine for pedestrian classification. They evaluate the performance of a single classifier for various pedestrian subtypes (including cyclists) or multiple classifiers in cascade. They also evaluate the difference between classifying grey-scale images and classifying binary images. Xu et al. found that the single classifier performed better than multiple classifiers; however, this comes with a speed penalty. It was also shown that due to the sensitivity of the binary image to the threshold, better classification was achieved with the grey-scale image.

In a recent paper, Navarro-Serment et al. [58] use two Support Vector Machines (SVMs) for classifying whether objects in a 3D point cloud are pedestrians or not. They use the unique elements from the covariance and normalised moment of inertia matrices as some of the features for the classifiers. A Principal Component Analysis (PCA) is performed; the pedestrians are assumed to be upright so the principal component is expected to be vertically aligned with the person's body. The second-largest component would be horizontal across the width of the person. The principal components are used to create planes that divide the object into an upper half, and two lower halves which correspond to the trunk and legs of the pedestrian respectively. The covariance matrix (2D) for each planar region is calculated and provides three features for each region. The remaining features are from histograms for each of the two principal planes (the width-height and depth-height planes or the first-second and first-third principal components).

2.2.4 Neural Network Classifiers

Neural networks have been used for a number of visual classification problems such as handwritten digit recognition [59], face detection [60, 61] and pedestrian detection [62].

Artificial Neural Networks (ANNs) are designed to mimic biological neurons. An ANN consists of layers of interconnected artificial neurons. Each neuron takes multiple input values, multiplies each by a weight and sums them. The sum of the weighted inputs is the input to a non-linear activation function, as shown in *Figure* 2.1. The output of a neuron with n inputs can be expressed mathematically as

$$O = f\left(\sum_{i=1}^{n} w_i x_i\right) \ . \tag{2.30}$$

A commonly used activation function is the sigmoid (*Equation* 2.31), which is smooth and easily differentiable

$$f\left(x\right) = \frac{1}{1 + \mathrm{e}^{-x}} \ . \tag{2.31}$$

Another choice is the hyperbolic tangent function, which has a shape similar to the sigmoid except that it has an output range of $O \in (-1, 1)$ while a sigmoid has a range of $O \in (0, 1)$

$$\tanh\left(x\right) = \frac{1 - \mathrm{e}^{-2x}}{1 + \mathrm{e}^{-2x}} \ . \tag{2.32}$$

There are a wide variety of ANN structures but the most popular is the Multi Layer Perceptron (MLP). The MLP consists of an input layer of neurons, one or more hidden layers and an output layer, as shown in *Figure* 2.2. The input layer in *Figure* 2.2 is passive; it simply distributes the inputs

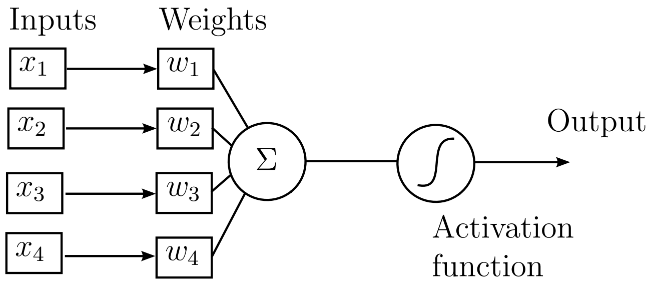


Figure 2.1: A single artificial neuron

to all of the hidden neurons. Each neuron in the hidden layer takes all the inputs and produces an output that is fed into all of the neurons in the next layer. In this case the layer after the hidden layer is the output layer; however, it is possible to have multiple layers of hidden neurons. In *Figure* 2.2 the network only has a single output; however, it is possible to have multiple outputs from a single network for multiple classifications.

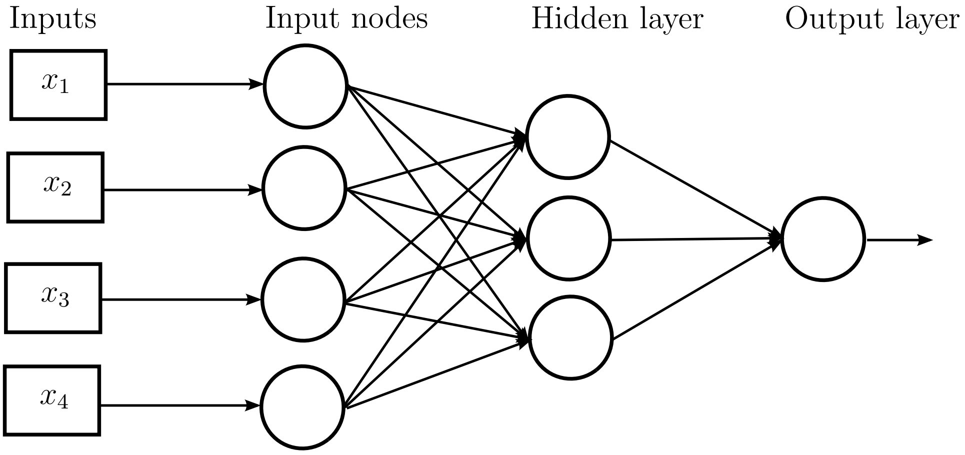


Figure 2.2: An artificial neural network with a single hidden layer

There are a number of parameters that can be tuned when designing an ANN classifier: these include the type of preprocessing of the data, the number of hidden nodes, the number of layers and the activation function.

Zhao and Thorpe [62] use a neural network to classify pedestrians. They produce a gradient magnitude image from segmented stereo images. The segmented images are scaled to (30×65) pixel images and the image gradient magnitude is calculated. The gradient magnitude has the advantage of being robust to illumination changes. The whole gradient image is then used, unprocessed, as the input to a neural network classifier.

Rowley, Baluja and Kanade [60] produce an image pyramid by subsampling an image multiple times. Windows of 20×20 pixel are extracted from each image in the pyramid. Each window is processed to correct the lighting and then histogram-equalised. The hidden neurons used by Rowley et al. are not fully connected, like the ones shown in $Figure$ 2.2, but each one has a receptive field of pixels that form its input. The receptive fields are: four which look at 10×10 pixel subregions, 16 which look at 5×5 pixel subregions and six which look at 20×5 pixel overlapping horizontal stripes.

Le Cun et al. [59] use a neural network to classify handwritten digits. They skeletonise 32×32 pixel images of handwritten digits and then extract a number of features. The features extracted represent lines, line ends and arcs. After some processing, 18 3×5 feature maps are produced; these give 270 inputs to the neural network chip used by Le Cun et al. The network used by Le Cun et al. has 40 hidden neurons, fully connected to the input, and 10 output nodes, one corresponding to each digit.

2.3 Object Tracking

Once pedestrians have been detected in each image, they need to be tracked through an image sequence and their trajectory determined. There are a number of tracking methods that can be used to track people in video sequences. When pedestrians are well separated in the image, geometric proximity is usually sufficient for determining the correspondence between objects and therefore determining their paths [32].

Dai et al. [32] use a combination of geometric proximity and shape similarity for tracking. Dai et al. scale all pedestrians to the same scale and then projects them onto eigenvectors. The similarity is defined as the L_2 norm calculated in the space spanned by the eigenvectors. A graph-matching framework is used where the graph G has two sets of nodes for the detected pedestrians, one for each of the two frames that are being used for detection. Dai et al. assume that the same number of pedestrians (Q) are detected in both frames; the graph therefore has $2Q$ nodes denoting the detected pedestrians, where $U = \{u_1, ..., U_Q\}$ are the nodes from image k and $V = \{v_1, ..., v_Q\}$ from image $k + 1$. For any $u \in U$ and $v \in V$ there is an edge between them that has a weight

$$w\left(u, v\right) = \alpha d_{sa} + \left(1 - \alpha\right) d_{eu} \tag{2.33}$$

where

d_{sa} is the shape similarity between two pedestrians

d_{eu} is the Euclidean distance between two pedestrians

α is the overlapping ratio of u and v.

The overlapping ratio, α, is the ratio of the area of overlap between pedestrians and the pedestrian window size. From *Equation* 2.33 it is evident that for well-separated pedestrians only the Euclidean distance is used while when they overlap the appearance term becomes significant.

Haritaoglu et al.[47] use a second order motion model to predict the location of a tracked region in subsequent frames. The position of foreground object is compared to the predicted position of the object and used as an initial estimate of the object's position. Haritaoglu et al. use the median coordinate of the object and state that it is a more robust estimate of a person's position and is not affected by large movement of the extremities. The initial estimated position is used to narrow the search space for a silhouette correlation which is used to compute the actual position of the person. To deal with the case of two people meeting and then reappearing separately Haritaoglu et al.'s system creates a textural template during tracking for each object. The template is a weighted sum of foreground pixel intensities weighted according to the frequency with which they are detected as foreground pixels.

Navarro-Serment et al. [63] use object overlap and a Kalman filter for tracking objects. They define a bounding box around each object. The positions of the objects are fed into a Kalman filter that estimates the object velocity. Using the velocity estimate the position of the bounding box in the current frame is determined and the object that falls within this box is assumed to be the tracked object. The robustness of the measurements is increased by validating the velocity estimates. Velocity estimates are required to stay consistent for a minimum amount of time and have a standard deviation below a certain threshold.

2.4 Distance Determination

There are a number of ways that can be used to determine the distance to objects that are being tracked by the thermal camera. Depth can be determined from an image using stereo cameras, structure from motion, depth from focus or defocus or by fusing the thermal image with depth information from a separate sensor. The options for depth determination are discussed below.

2.4.1 Stereo Imaging

Stereo imaging uses the disparity between two images captured simultaneously by two cameras. In order to determine the disparity between two images the image requires sufficient texture to determine correspondences. Using two thermal cameras in order to get the depth is currently not a viable option due to the high cost of thermal cameras. Even if the cost was not a problem, the lack of texture in thermal images would make accurate depth determination difficult.

2.4.2 Depth from Focus and Defocus

Depth from focus is a technique where the focus of a camera is actively searched in order to maximise sharpness. The camera focus is adjusted until an object of interest reaches maximum sharpness; at this point the distance to the object is simply determined using the focal length of the lens and the distance to the image plane. Depth from focus is an active search method that involves having to change the focus of the camera to get maximum sharpness. Having to adjust the camera focus for each object of interest makes depth from focus too slow for use on a moving platform [64].

Depth from defocus uses a number of images taken with different camera geometries to determine depth. When a point in the scene is not in focus, its image is spread over some blur circle. If the radius of the blur circle can be estimated then the distance to the image point can be calculated. At least two different images of the same scene are usually used to determine the defocus [65]. This can be done by taking an image with multiple cameras simultaneously, or by taking multiple images with a single camera. Pentland [66] uses two optically co-aligned cameras with different apertures to calculate distance. Pentland uses a pinhole camera with a camera having a finite aperture. The radius of the blur circle is estimated and then used to calculate the distance to the point using

$$d = \frac{fv}{v - f - \sigma F} \tag{2.34}$$

where
f is the focal length of the lens
v is the distance from the lens to the image plane
F is the f-number of the lens
σ is the radius of the blur circle.

Using a depth from defocus method with two cameras would also require two cameras, with the associated cost. Using co-aligned cameras has the advantage that the pixels are the same in both images and therefore the correspondence problem faced with stereo cameras is eliminated. Using two cameras with different apertures, however, creates difficulties with differing illumination. Using a single camera and taking multiple images has similar disadvantages to depth from focus. It is slow due to the need to acquire multiple images for a single depth calculation.

2.4.3 Structure from Motion

To determine the depth of points in a scene, two images of a matched point are needed. Then given knowledge of the translation and rotation between the camera poses, the depth can be estimated [67]. Stereo vision acquires two images simultaneously from two cameras a known distance apart. Structure from motion takes two images at different times from a moving camera to get two views of the scene. If the displacement of the camera between two images is known then the distance to the objects in the scene is fairly simple to calculate. If the displacement of the camera is not known then structure from motion involves determining the movement of the camera and then using that to calculate the depth of objects in the image. Without additional information, structure from motion can only determine the depth of objects up to a global scaling factor [67, 68]. Further information is required for absolute depth, such as the size of the object of interest, additional knowledge of the camera motion, or the distance to one point on the object [68, 69].

2.4.4 Active Distance Sensors

In addition to the above-mentioned passive distance sensors there are a number of active sensors that can be fused with the thermal camera to provide position information. Active sensors are the most popular sensors for obstacle detection and avoidance owing to the simplicity of reading out position information [64]. Active ranging sensors therefore provide a computationally simple method of determining the 3D information about a scene. The two main types of active range sensors are TOF-based sensors and triangulation-based sensors.

Time-of-flight Distance Sensors

There are a number of ranging sensors that all measure distance using the round-trip time of a signal reflected off the object being measured. These systems include laser scanners, ultrasonic rangers, radar, TOF cameras and range-gated cameras.

Ultrasonic sensors use the speed of sound to determine the distance to an object that reflects a transmitted sound wave. Using the round-trip time t, the distance to the object being measured is

$$d = \frac{ct}{2} \, . \tag{2.35}$$

The speed of sound c in air is given by

$$c = \sqrt{\gamma R T} \tag{2.36}$$

where

γ is the adiabatic ratio of specific heats of air
R is the specific gas constant
T is the temperature in Kelvin.

There are a number of characteristics of ultrasonic distance sensors that limit their use to very simple ranging tasks. One of the major problems is that the sound propagates in a fairly wide cone-shape with an opening angle of between 20° and 40° . The wide beam width produces distance measurements that are the distance to a region of constant depth instead of a point depth measurement. Narrower beams can be produced; however, they require a physically large transducer; for example, a transducer with a 2° beam-width which is 335 mm in diameter [22, pp. 176-179]. Another limitation is the speed of sound, which limits the cycle time of ultrasonic sensors. Using *Equation* 2.36 and assuming a temperature of $20°C$ then the speed of sound in air can be shown to be 343 m/s; so for an object 5 m from the sensor the round-trip time is about 29 ms. The round-trip time will limit the maximum measurement frequency to about 30 Hz, which is acceptable for a single sensor but is too slow to allow for any form of scanning to produce a depth image.

It has been shown that the attenuation of millimetre-wave radar signals by mine dust and thick fog is negligible [70]. This makes radar a promising sensor for distance sensing underground. Radar systems usually consist of a single beam which is scanned to produce a two-dimensional (2D) or 3D range image. Millimetre-wave radar has a small enough aperture and narrow enough beam width to allow it to be scanned using mirrors. The wavelength of millimetre-wave radar requires mirrors that are large in size compared to those of laser scanners. The large size of the scanning system makes the use of radar problematic in a confined underground environment. Single beam radar systems similar to those being added to high-end cars provide a compact alternative; however, the close proximity of tunnel walls causes a high number of false positives [6].

Range gated cameras are another TOF-based distance sensor, even though they are frequently not used as a ranging sensor. Range gated cameras are usually used for enhanced visibility through smoke, dust and bad weather conditions [71–73]. Gated cameras operate by transmitting a very short laser pulse which is reflected off the scene and captured by a camera with a very short exposure time. By varying the time delay between when the laser pulse is transmitted and when the camera begins its exposure a section of the scene at a certain distance from the camera can be imaged. Since the camera only receives the light from objects at the desired distance the backscatter from obscurants is significantly reduced. Light scattered off particles between the camera and the target is not imaged because it will arrive at the camera before the gate is open. Owing to the very short exposure time the amount of light received by the camera is very low, so the camera is usually a photo-cathode image intensifier [74]. The exposure of the camera is controlled by switching the high voltage supply to the image intensifier on and off. In order to produce 3D images the delay between the laser pulse and the camera exposure is incrementally increased. This is equivalent to taking slices of the scene at increasing distances from the camera. The range to each pixel is measured by determining at which range setting a particular pixel is lit up. The rise time of intensifier tubes is typically of the order of 10 ns [74] which limits the application of gated cameras to very long distances (in the order of kilometres) with low resolution (metres).

Laser scanners and TOF cameras operate on the same basic principle. Both have of an emitter that emits a pulse of light and a receiver that measures the round-trip time of the light. For typical measurement distances of a few metres the round-trip time is in the order of picoseconds and therefore the electronics required to directly measure the elapsed time are expensive. A lower-cost method of determining the distance is to measure the phase shift between the transmitted and reflected light. TOF cameras measure the phase shift for all pixels simultaneously. Laser scanners have a single receiver that is mechanically scanned and can use pulse travel time or phase shift measurement methods. Laser scanners sweep a single beam in one or two planes using rotating mirrors.

Commercial TOF cameras use a modulated near-infrared light source and measure the phase shift between the transmitted and received light [75, 76]. The maximum unambiguous distance (D_{unamb}) to a target would be

$$D_{unamb} = \frac{c}{2f} = \frac{\lambda}{2} \tag{2.37}$$

where f and λ are the modulation frequency and wavelength respectively.

Any distance less than D_{unamb} is calculated by measuring the ratio of the phase shift (ϕ) to a full cycle and multiplying it by the maximum distance

$$\begin{aligned} d &= \frac{\phi}{2\pi} D_{unamb} \\ &= \frac{\lambda}{4\pi} \phi \ . \end{aligned} \tag{2.38}$$

One of the problems with TOF cameras is the phase shift ambiguity. A phase shift of slightly over 2π would be measured as a shift of just greater than zero and according to *Equation* 2.38 the calculated distance would be small. There are a number of methods to correct for the phase ambiguity; however, each has its own challenges. One method is to ignore any measurement with a low intensity (because it is likely to be from a distant object); this method is very simple but it results in the loss of information and can still produce errors with distant, highly reflective objects. There are also methods that can unwrap the ambiguous measurements [77]; however, these methods are computationally expensive and/or require multiple frames captured with different modulation frequencies.

Geometric Distance Sensors

The simplest geometric distance sensors are triangulation-based sensors. The principle of a triangulation-based distance sensor is simple. A collimated beam of light is directed towards a target; the reflected light is focussed by a lens on a position sensitive device (PSD) or linear camera. From the geometry shown in *Figure* 2.3, it is easy to see that

$$d = f \frac{b}{x} \ . \tag{2.39}$$

Single spot triangulation-based distance sensors provide a low-cost, high bandwidth alternative to ultrasonics.

Triangulation-based sensors that use a 2D camera instead of the linear camera or PSD are known as structured light sensors. Structured light sensors project a known pattern onto a surface and record the deformation of the pattern using a camera a certain distance from the projector. The most common pattern for a structured light scanner is a single laser stripe that is swept across the object [78]. *Figure* 2.4 shows the principle used to calculate the distance by triangulation. It can be shown using similarity of triangles that the x and z coordinates of the target are

$$x = \frac{bu}{f \cot \alpha - u} \tag{2.40}$$

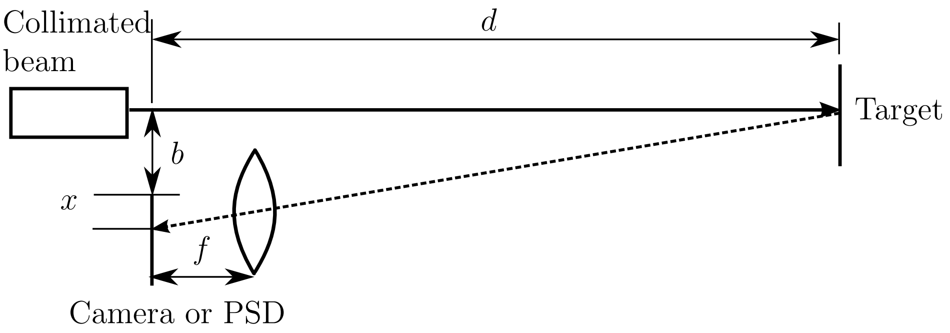


Figure 2.3: Geometry of a triangulation-based distance sensor

and

$$z = \frac{bf}{f\cot\alpha - u} \ . \tag{2.41}$$

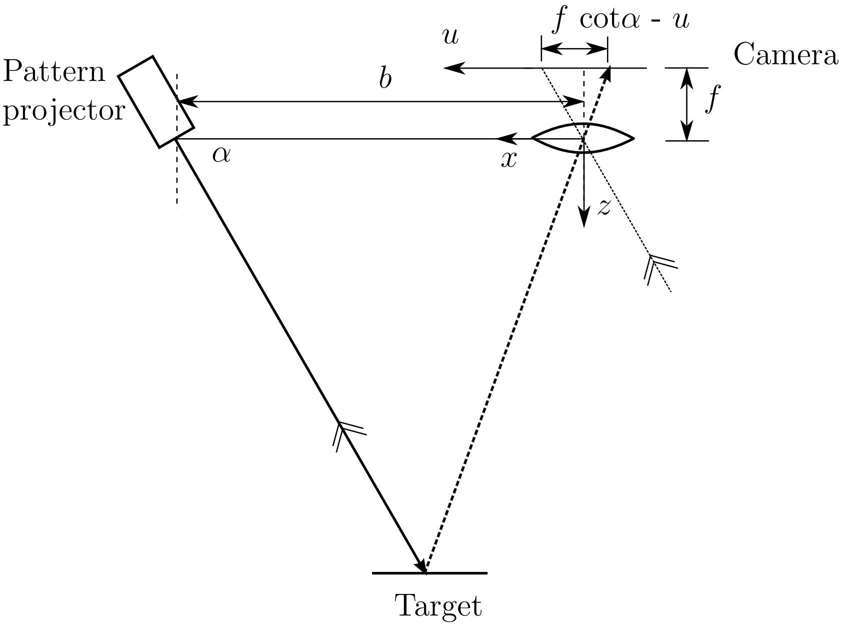


Figure 2.4: Schematic showing the principle of structured light triangulation (adapted from [64])

The disadvantage of using a single line and scanning it across the scene is that it is slow, having to acquire an image each time the laser moves. Another major disadvantage of using a single line is that it can only be used for static objects due to the scanning time. One approach to circumvent these disadvantages is to project a 2D pattern onto the entire scene and reconstruct the depth values from a single image. There are a number of patterns, such as parallel stripes, a grid of lines or blocks and a matrix of dots. Using a 2D pattern requires a method of uniquely identifying which reflected point corresponds to which projected point. In order to solve the correspondence problem a coded pattern is projected onto the scene. There are many different coding schemes such as spatial coding, colour coding and grey-level coding [79]. A binary spatial coding scheme will be described because it is the method used by the Kinect sensor [80] that has been used for this work. The pattern used by the Kinect is an uncorrelated pattern of dots which is such that its auto correlation for any shift less than the maximum shift that would occur over the range of depths is negligible. Therefore the correlation of the received pattern with the projected pattern will only return a result when the patterns correspond. Correlating the projected and received patterns allows the shift in the pattern to be determined for each position and thus the depth.

2.5 Camera Calibration

Camera calibration basically involves the recovery of the parameters that map 3D points in the world reference frame to 2D points in the camera's reference frame. A pinhole camera model is shown in *Figure* 2.5. From the

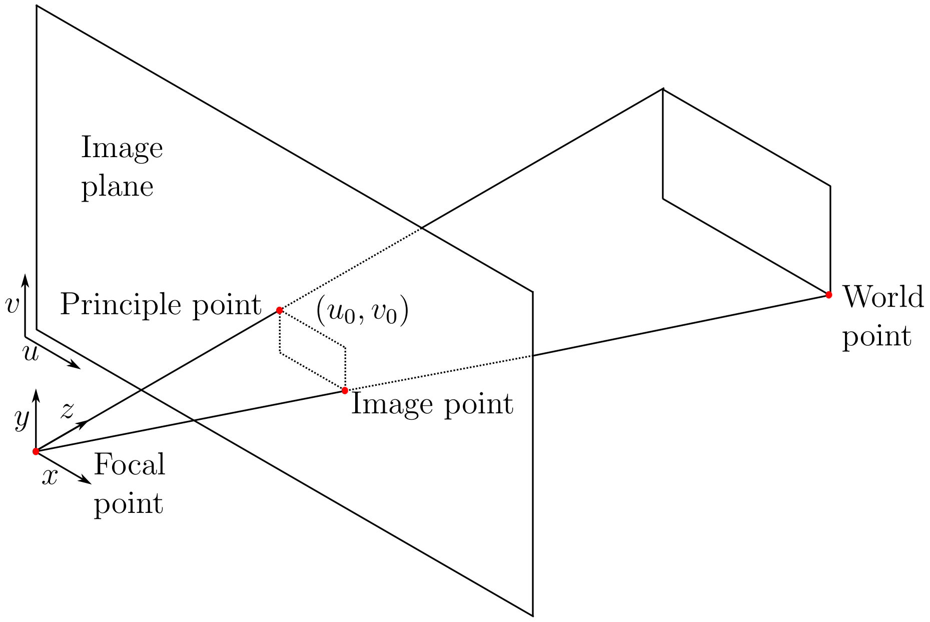


Figure 2.5: A pinhole camera model

figure it can be seen that

$$\frac{u}{f} = \frac{x}{z}$$
$$\therefore \; u = \frac{fx}{z} \tag{2.42}$$

and

$$v = \frac{fy}{z} \; . \tag{2.43}$$

We can combine *Equations* 2.42 and 2.43 to get

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \tag{2.44}$$

in homogeneous coordinates.

To get the actual 2D image points from the 3D homogeneous coordinates we divide through by w, which gives the point expected from *Equations* 2.42 and 2.43. The points in an image are not measured relative to the principle point but instead relative to a corner. The translation between the corner and principle point of an image is (t_u, t_v); therefore, *Equations* 2.42 and 2.43 become

$$u = \frac{fx}{z} + t_u \tag{2.45}$$
$$v = \frac{fy}{z} + t_y \; . \tag{2.46}$$

Therefore we can rewrite *Equation* 2.44 as

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} f & 0 & t_u \\ 0 & f & t_v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \; . \tag{2.47}$$

Equation 2.47 projects a 3D world point into the camera's reference frame. The projected point is still, however, in the same distance units as the world point (e.g. m). The camera image provides image point positions in pixels and not as a true distance; therefore, we need to know the camera resolution in pixels per metre. If the pixel resolutions in the u and v directions are m_u and m_v respectively then the projection matrix becomes

$$\begin{aligned}\begin{pmatrix} u \\ v \\ w \end{pmatrix} &= \begin{pmatrix} m_u f & 0 & m_u t_u \\ 0 & m_v f & m_v t_v \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \\ &= \begin{pmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \\ &= K \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \end{aligned} \tag{2.48}$$

The matrix K defines the intrinsic properties of the camera, such as its focal length and principle point. In cases where the image axes are not perpendicular an additional parameter, the skew parameter (s), is added to the intrinsic parameter matrix (K)

$$K = \begin{pmatrix} \alpha_x & s & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2.49}$$

In the case where the camera is not at the origin, facing in the z-direction of the world reference frame, then a reference frame transform is required to align the reference frames. A rotation is applied to align the z-axis of the world reference to the principle axis of the camera and then the frame is translated to move the origin to coincide with the focal point of the camera. This can be represented by a generic coordinate transform as shown in *Equation* 2.50.

$$\begin{pmatrix} x_t \\ y_t \\ z_t \\ 1 \end{pmatrix} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2.50}$$

The 3D rotation (direction cosine) matrix is defined as R and the 3×1 translation matrix is the matrix T. If we define a $n \times m$ zero matrix as 0_{nm}, then *Equation* 2.50 can be written as

$$\begin{pmatrix} x_t \\ y_t \\ z_t \\ 1 \end{pmatrix} = \begin{pmatrix} R & T \\ 0_{13} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}. \tag{2.51}$$

The final calibration matrix going from four-dimensional (4D) homogeneous coordinates to 3D homogeneous coordinates of the camera is

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} K & 0_{31} \end{pmatrix} \begin{pmatrix} R & T \\ 0_{13} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix} \tag{2.52}$$

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} \alpha_x & s & u_0 & 0 \\ 0 & \alpha_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}. \tag{2.53}$$

Chapter 3

Preliminary Design

This chapter outlines the design of the pedestrian detection system created for this study. A block diagram indicating the major subsystems of the detection system is shown in *Figure* 3.1. There are a large number of segmentation and classification methods used for thermal image processing. The choices of segmentation and classification methods appear to be largely ad hoc; therefore, a preliminary system was designed to test various algorithms and determine which are most appropriate for detecting people underground, with thermal imaging.

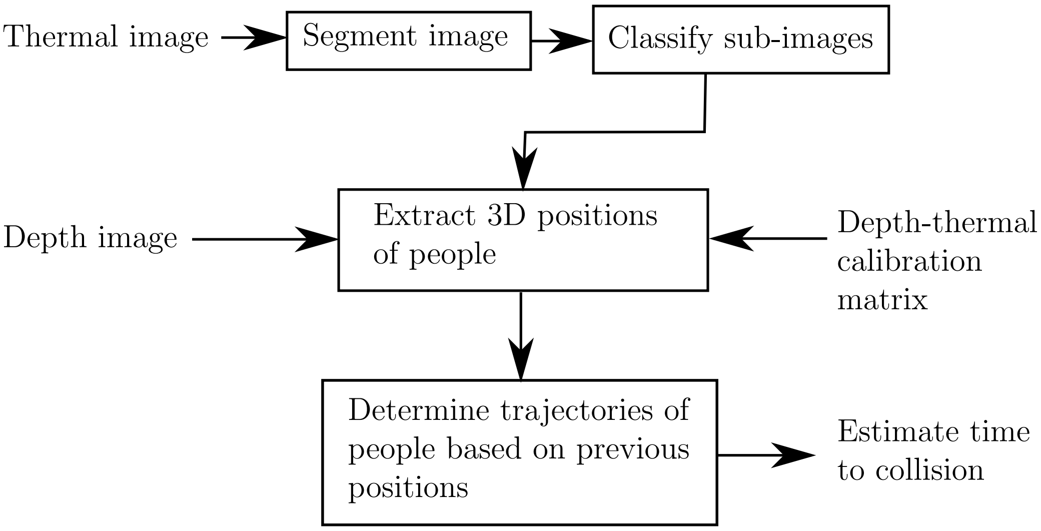

Thermal image
Segment image
Classify sub-images
Depth image
Extract 3D positions of people
Depth-thermal calibration matrix
Determine trajectories of people based on previous positions
Estimate time to collision

Figure 3.1: A block diagram of the pedestrian detection system

3.1 Segmentation

As discussed in *Section* 1.2.4, we know that people will be hot spots in the thermal images. The process of segmentation involves the determination of a threshold that optimally extracts people from the background.

Motion-based segmentation methods were not considered for the preliminary evaluation because the camera is mounted on a moving vehicle so the background is constantly in motion. Additionally since the vehicle is moving a person who is stationary may still be in danger; however, a motion-based segmentation method would not extract stationary people.

Using a thresholding method that determines a single threshold is not desirable because the threshold would need to be modified each time the vehicle moved into a new area with a slightly different background temperature. For this reason the thresholding methods that were evaluated for segmenting the thermal images were those which adapted the threshold based on each image.

The static segmentation methods of Xu et al. [50], and Thornton et al. [49] are designed for thermal images that vary their threshold based on each image. There are also a number of thresholding methods, shown to be effective on grey-scale images, that were modified and evaluated on thermal images. The grey-scale segmentation methods use statistical information about the images to segment them, so whether the underlying signal is a grey-scale intensity or a temperature signal does not matter.

3.1.1 Qualitative Comparison

Xu's Thresholding

Xu et al. [50] use a threshold that is based on the mean and standard deviation of a histogram-equalised image. They determine a threshold using *Equation* 2.1, which extracts the brightest 10% of the pixels in an image. Segmenting the warmest 10% of the pixels works fairly well for the images where the people are far from the camera because they are the warmest objects in the image. As the people come closer the difference in temperature between exposed skin and clothing becomes evident. With people close to the camera the warmest 10% of the pixels are only parts of each person. This is evident in *Figure* 3.2, image *h* shows a fragmented segmentation result because the warmest 10% of pixels represent the exposed skin of the people in the image.

Thornton's Normalised Intensity Thresholding

Thornton's method [49] can be shown to be equivalent to thresholding the image at a number of standard deviations above the mean. Thornton's threshold is based on a normalised intensity deviation image (n_{ij}). If the segmented image is $n_{ij} > T$ then using *Equation* 2.2 the thresholding is equivalent to

$$x_{ij} > m + T\sigma \ . \tag{3.1}$$

From *Equation* 3.1 it is evident that there is an implicit assumption about the relative number of pixels belonging to people versus those of the background. If it is assumed that the number of pixels belonging to foreground objects is small then the mean and standard deviation of the image can be assumed to be approximately equal to those of the background. If this is the case then the threshold selects pixels that have a probability below some threshold of belonging to the background.

The thresholding results shown in *Figure* 3.3 indicate that Thornton's method is effective on the first three images; however, investigating the underlying assumption reveals the underlying disadvantage of the method. *Figure* 3.3, *h* shows the thresholding result where the number of background pixels is not very much larger than the number of foreground pixels. As the proportion of foreground pixels increases the image mean increases because the foreground is warmer than the background. In addition the standard deviation of the image also increases. Both of these effects combine to increase the threshold, causing the foreground to be severely fragmented.

Thornton's method creates a local and global threshold and then fuses the two. The local threshold is calculated using a mean and standard deviation which is calculated in some region around each pixel. The local threshold only works well when the foreground objects are smaller than the window, otherwise the centre of the person is segmented as background because the mean is now the mean of the foreground. It may seem that using a large window would improve the results; however, there is no window size that will always be larger than the person because as a person moves closer to the camera their size in the image increases. Sauvola's method (*Section* 3.1.1) also uses a local window and suffers from similar problems to Thornton's local threshold.

Sauvola's Locally Adaptive Thresholding

The locally adaptive thresholding method does not improve the results when compared to a global threshold but actually performs worse. The method extracts the edges of people and a fair amount of noise from the background, as seen in *Figure* 3.4. In the work of Sauvola and Pietikäinen [55], the authors segmented printed characters which were smaller than the size of the window used to calculate the local mean and standard deviation. Unlike text, the foreground objects in the thermal images are large. In the images in *Figure* 3.4 the people (foreground objects) are larger than the window so the mean increases near the centre of the object where the window encloses the whole object. The increasing mean towards the centre of the foreground objects causes a commensurate increase in the threshold and hence the tendency of this method to extract edges only. Thornton's method also suffers from a similar problem when using a local threshold.

Otsu's Method

Otsu's method [51] produces acceptable results for images where the numbers of foreground and background pixels are approximately equal [52]. This is not the case in the thermal images that are typically segmented by the system. Having approximately the same number of foreground and background pixels would mean that people would take up half of the image. This would only happen when there were people very near the camera. When there is a significantly larger number of pixels in one class than the other, then Otsu's method tends to split the larger mode in half [53]. This is exactly what is seen in *Figure* 3.5: the threshold is dividing the background distribution and extracting the higher intensity background pixels as foreground pixels.

Kapur's Entropic Thresholding

The entropy-based threshold performs well on all of the images except image g (see *Figure* 3.6). The poor performance on image g is most likely due to the high variance in the foreground.

Kittler's Minimum Error Thresholding

The results of Kittler and Illingworth's minimum error thresholding [53] algorithm, shown in *Figure* 3.7, indicate that the minimum error thresholding technique performs well on all of the input images, even image g which was a problem for the previous algorithms. It is evident that the minimum error thresholding method is the most robust of the methods tested. It performs well on all the images and is the only method that effectively segments image g.

Looking at all the results it is evident that none of the methods completely extract the people in image e. While this is unfortunate it is not possible for a global thresholding method to perform better. Parts of the people (their hard-hats, gum-boots and cap-lamp batteries) are at the same temperature as the background and therefore cannot be segmented from the background unless the threshold is locally adapted.

3.1.2 Quantitative Comparison

The thresholding methods are evaluated qualitatively in the preceding sections and it appears that Kittler's minimum error thresholding performs the best on a selection of images. To verify this result quantitatively, the performance of the thresholding methods is evaluated using 40 ground-truth images. The performance of each method is tested using the misclassification error (M_E), region non-uniformity (R_N), the relative foreground area error (R_F) and modified Hausdorff distance (M_H) metrics [52].

The M_E metric reflects the proportion of pixels that are wrongly classified: foreground as background and vice versa. The M_E varies from zero, for a perfectly segmented image, to one and can be expressed as

$$M_E = 1 - \frac{card\,|B_G \cap B_T| + \text{card}\,|F_G \cap F_T|}{\text{card}\,|B_G| + \text{card}\,|F_G|} \tag{3.2}$$

where
B_G and F_G are the background and foreground pixels of the ground-truth image, respectively
B_T and F_T are the background and foreground pixels of the thresholded image, respectively
card $|\cdot|$ denotes the cardinality of the set.

The R_N measures the difference in variance between an image and its foreground

$$R_N = \frac{\text{card}\,|F_T|}{\text{card}\,|F_T| + \text{card}\,|B_T|}\frac{\sigma_f^2}{\sigma^2} \tag{3.3}$$

where
σ_f^2 is the variance of the foreground
σ^2 is the variance of the whole image.

A well-segmented image should have a non-uniformity measure that is low, approaching zero. The worst case would be a R_N of one, corresponding to indistinguishable foreground and background.

The R_F is based on the ultimate measurement accuracy defined by Zhang [81], using the foreground area as the object feature. Using Zhang's definition of the ultimate measurement accuracy (U_A) with the foreground area as the feature of interest gives

$$U_A = \frac{|A_G - A_T|}{A_G} \tag{3.4}$$

where
A_G is the foreground area of the ground-truth image
A_T is the foreground area of the thresholded image
$|\cdot|$ denotes the absolute value.

With a slight change we get the R_F as defined by Sezgin and Sankur [52], which is now bounded within the interval $[0, 1]$. The R_F is defined as

$$R_F = \begin{cases} \frac{A_G - A_T}{A_G} & if\ A_T < A_G \\ \frac{A_T - A_G}{A_T} & if\ A_T \geq A_G\ . \end{cases} \tag{3.5}$$

An ideally segmented image will have the same foreground area as the ground truth, and therefore a R_F of zero.

The modified Hausdorff distance (MHD) measures the difference in the shape of the foreground of the segmented and ground-truth images. The Hausdorff distance measures how far two subsets of a metric space are from each other, or expressed formally

$$D_H(X, Y) = \max\left\{d\left(X, Y\right),\ d\left(Y, X\right)\right\} \tag{3.6}$$

where

$$d\left(X, Y\right) = \sup_{x \in X}\ \inf_{y \in Y}\ \|x - y\| \tag{3.7}$$

and
$\|\cdot\|$ is the underlying norm on the points in X and Y.

The function $d\left(X, Y\right)$ is known as the directed Hausdorff distance (DHD) and if the sets X and Y are finite, as is the case for sets of image pixels, the DHD is equal to

$$d\left(X, Y\right) = \max_{x \in X}\ \min_{y \in Y}\ \|x - y\|\ . \tag{3.8}$$

The DHD defined in *Equation* 3.8 is, however, sensitive to outliers due to the max. Dubuisson and Jain [82] tested various modifications to the Hausdorff distance and showed that replacing the max in the DHD with a mean performs best for shape matching. The M_H metric used for the evaluation of the segmentation methods is, therefore, defined as

$$M_H(X, Y) = \max\left\{d\left(X, Y\right),\ d\left(Y, X\right)\right\} \tag{3.9}$$

where

$$d\left(X,Y\right) = \frac{1}{\mathrm{card}\left|X\right|} \sum_{x \in X} \min_{y \in Y} \; \|x - y\| \; . \tag{3.10}$$

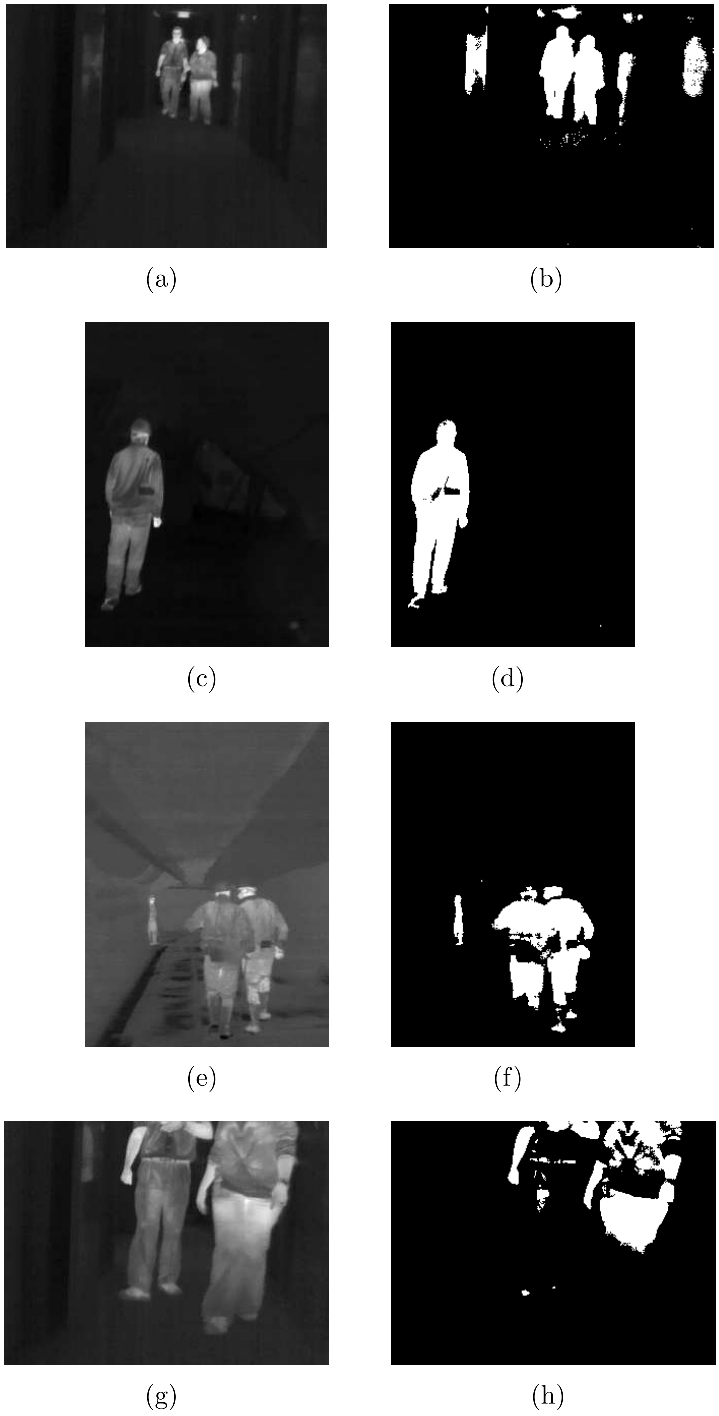
Unlike the previous metrics the MHD is not bounded within the interval $[0,1]$; therefore, the M_H metric is normalised by dividing each MHD by the maximum obtained for all the tests.

The average performance of the segmentation algorithms, in *Figure* 3.8, shows that Kittler and Illingworth's minimum error thresholding out-performs the other algorithms in all but one of the tests. Minimum error thresholding performs best overall using the four performance metrics on all 40 ground-truth images.

3.1.3 Summary

Both the qualitative and quantitative comparisons of the thresholding techniques support the conclusion that minimum error thresholding is the best performing thresholding technique for segmenting people in thermal images; consequently, a modified version of minimum error thresholding is used for the segmentation system.

All the tests performed in this section segmented images that contain people but the segmentation system needs to handle images that may or may not contain people. A modification to Kittler and Illingworth's method allows it to detect when image that does not contain a person and not segment it. The modification will be discussed in *Section* 4.1.



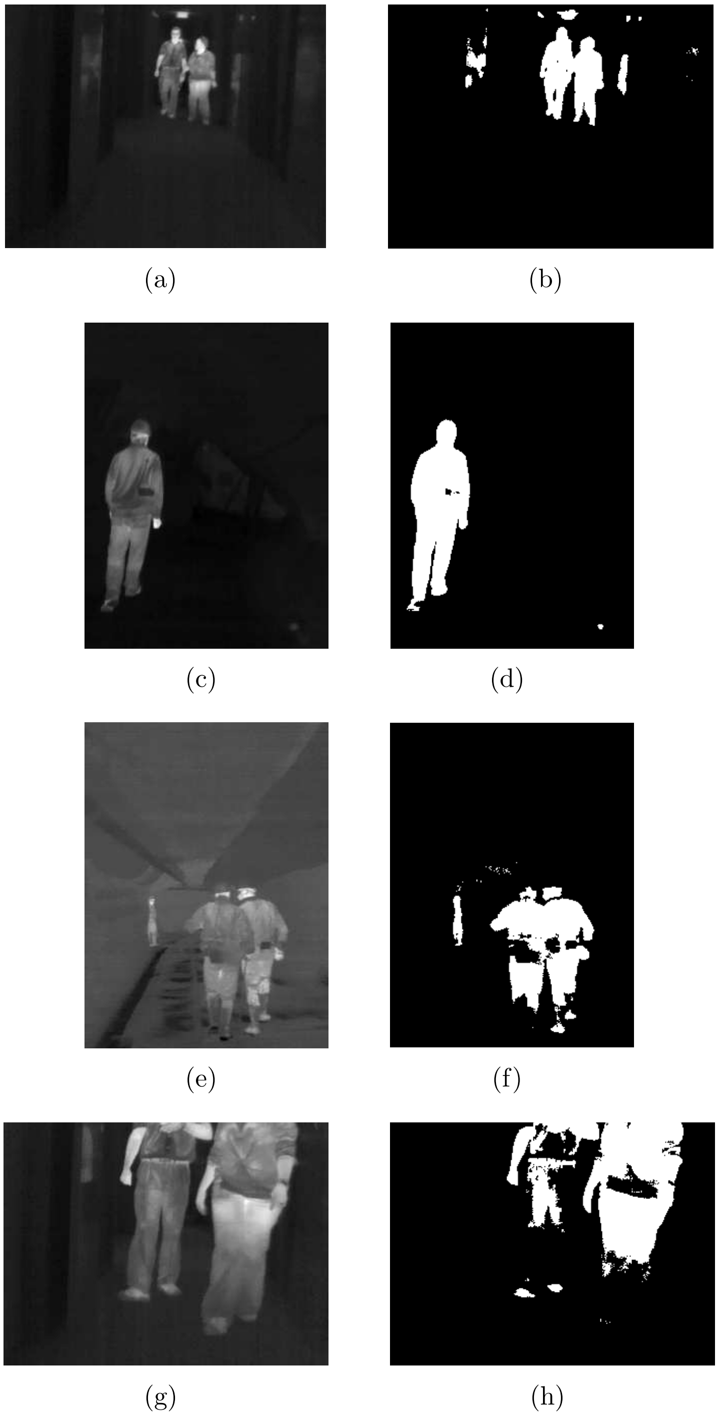
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.2: Thresholding results using Xu et al.'s method [50] - input images in the left column and thresholding results on the right



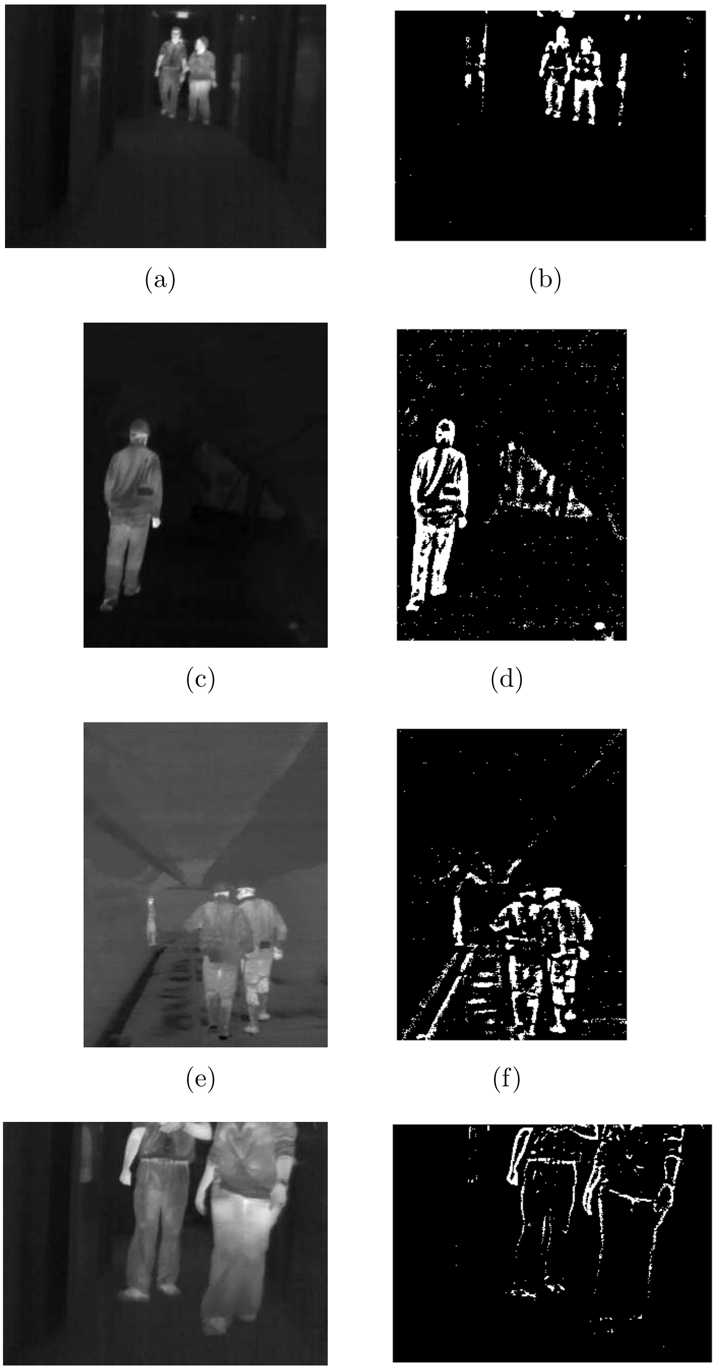
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.3: Thresholding results using Thornton's method [49] with a global threshold - input images in the left column and thresholding results on the right



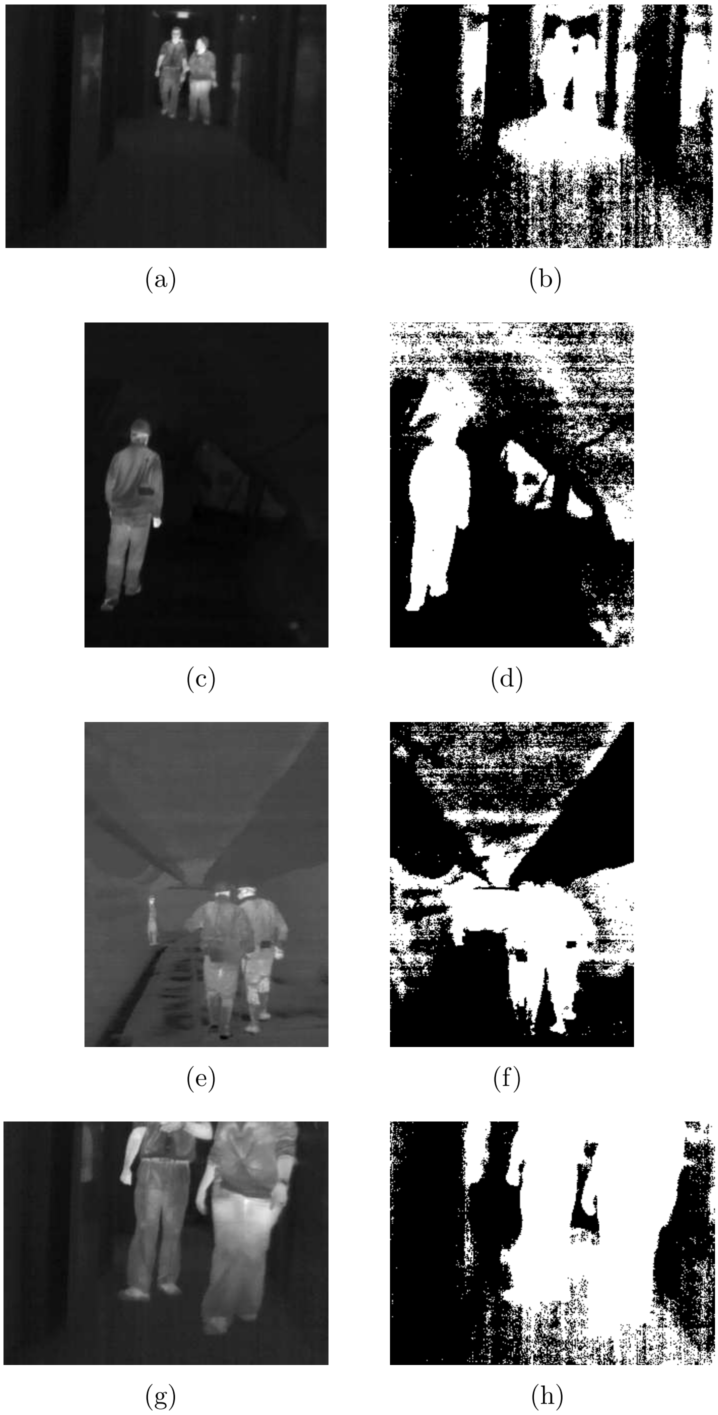
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.4: Thresholding results using Sauvola and Pietikäinen's method [55] - input images in the left column and thresholding results on the right



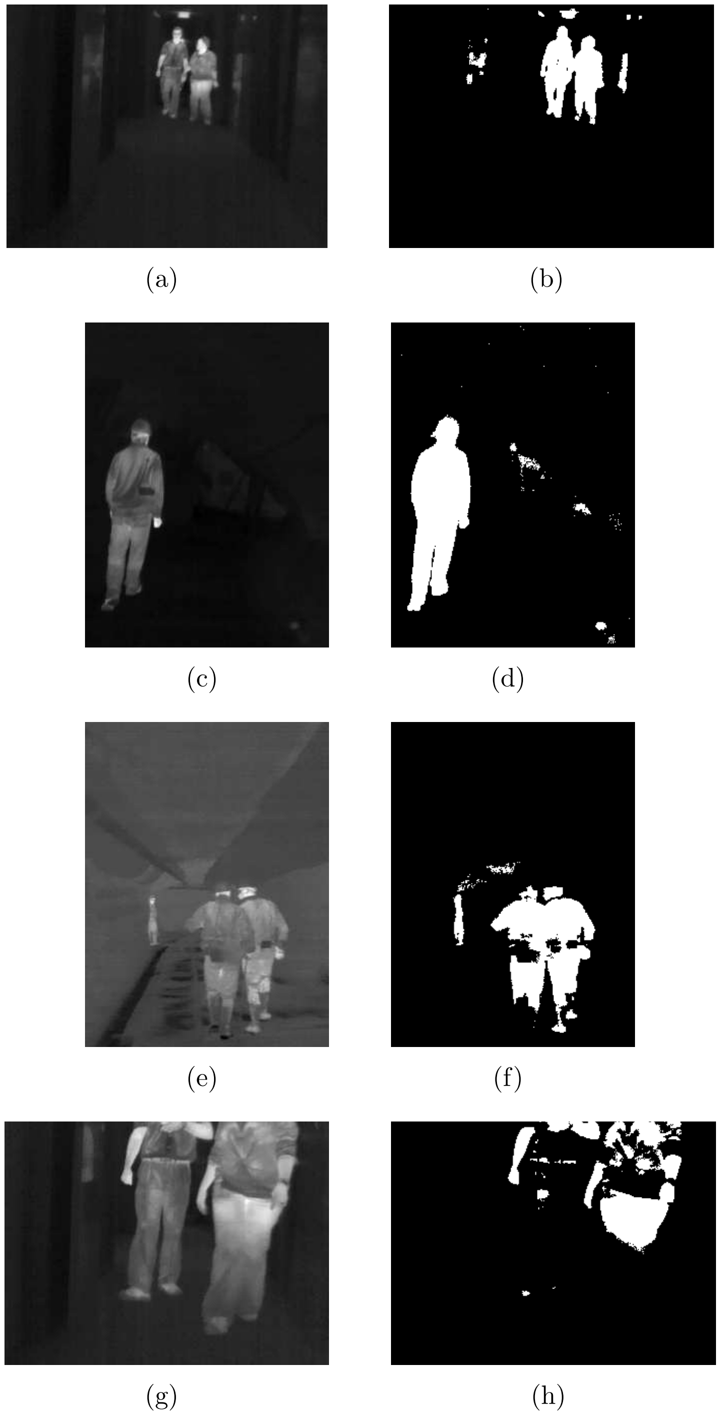
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.5: Thresholding results using Otsu's method [51] - input images in the left column and thresholding results on the right



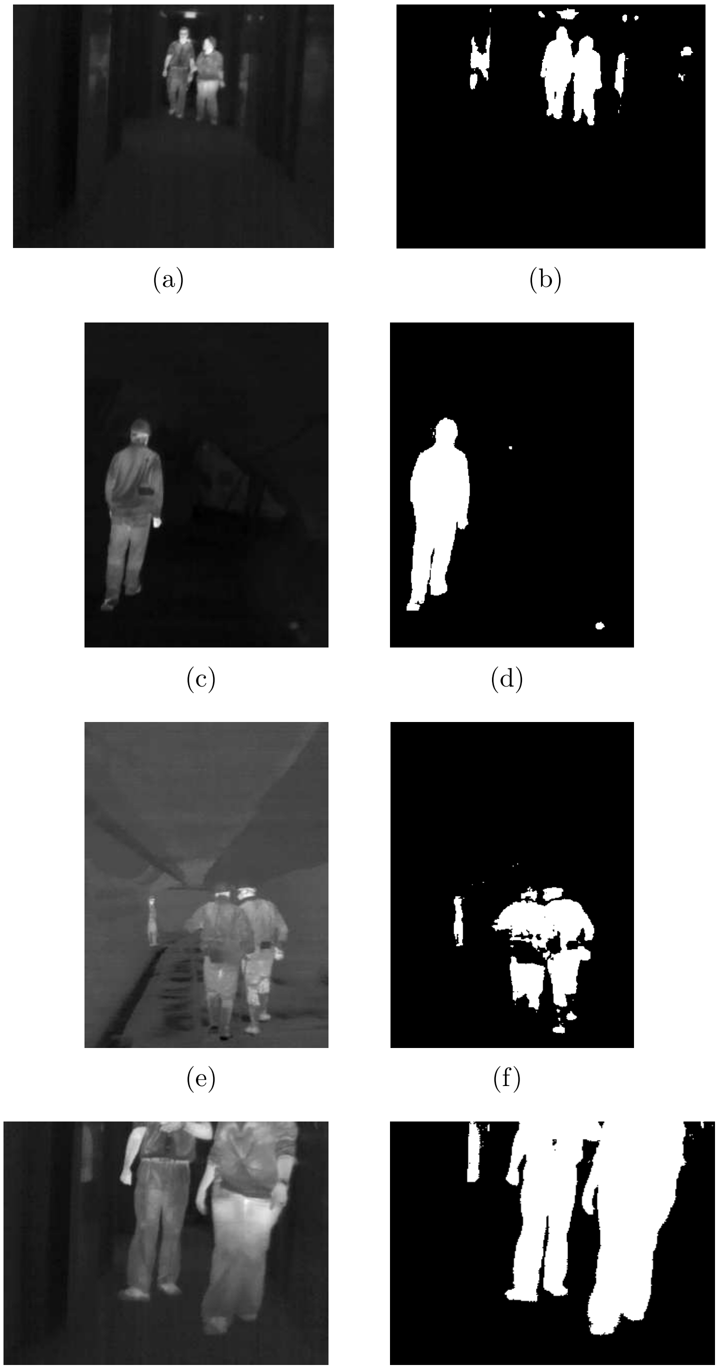
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.6: Thresholding results using Kapur et al.'s entropy-based threshold [54] - input images in the left column and thresholding results on the right



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 3.7: Thresholding results using Kittler and Illingworth's minimum error thresholding [53] - Input images in the left column and thresholding results on the right

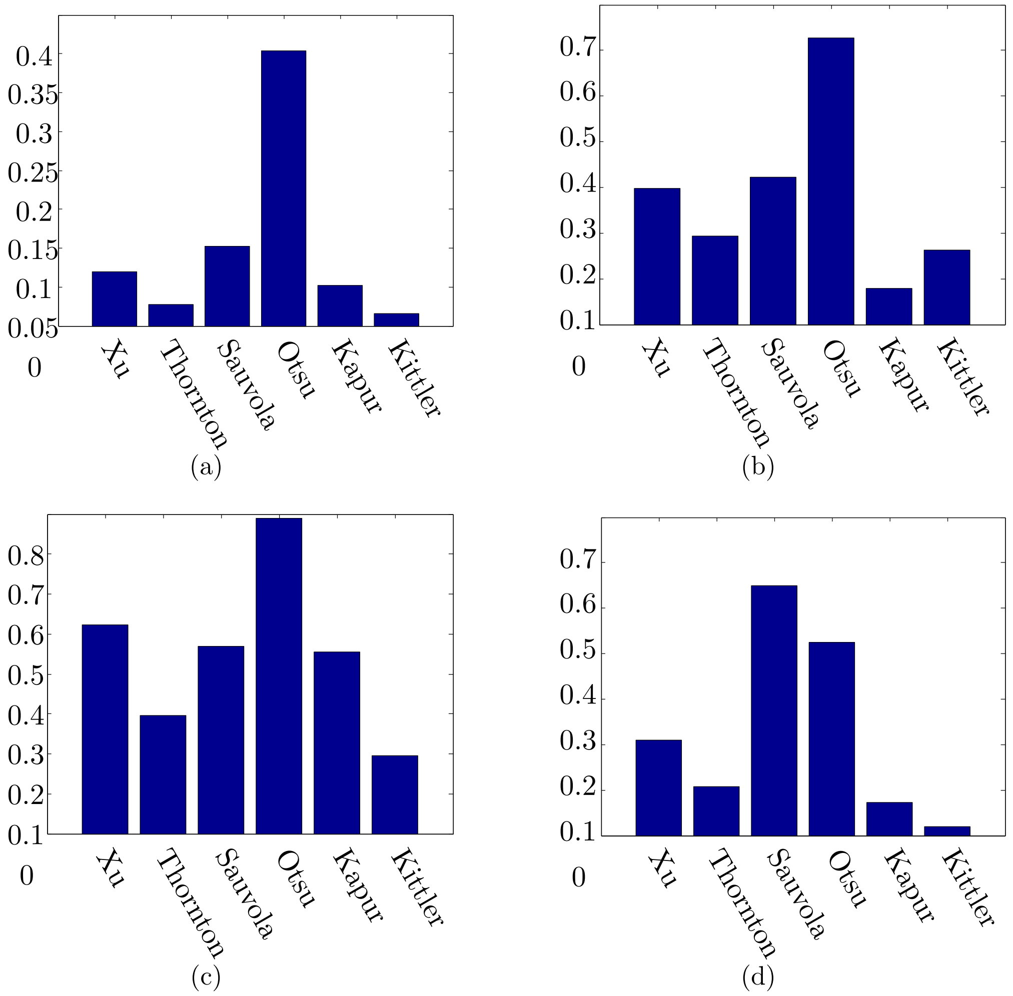


Figure 3.8: Average performance scores for the segmentation algorithms. a) misclassification error, b) region non-uniformity, c) relative foreground area error and d) the normalised modified Hausdorff distance

3.2 Classification

After segmentation the regions of interest undergo a validation stage to classify the regions that are actually pedestrians. Many image classification techniques exist; however, due to the wide range of image-processing applications, an objective comparison is difficult. To provide a more rigorous basis for the choice of classifier than a personal preference, some popular classification algorithms were compared.

3.2.1 Template Classifier

The first classifier that was tested was a template-based classifier, chosen because of its simplicity. The template method uses the total absolute difference between the template and the candidate image as the similarity measure. The template-based classification methods of Nanda and Davis [48], and Olmeda et al. [56] use templates that are based on the frequency with which each pixel is extracted as a foreground pixel. It was decided not to use a method based on the segmented images because there is a large amount of information that is lost, for example the fact that a person's face is warmer than the rest of their body, and arms and legs are cooler. The images of humans in the training data are rescaled to form an $M \times N$ image (30×12). A template is created by calculating the mean of the scaled images. The candidate regions are rescaled to the same dimensions as the template and the two are compared using an absolute difference, D, i.e.

$$D = \sum_{i=1}^{M} \sum_{j=1}^{N} |T_{ij} - I_{ij}| \tag{3.11}$$

where

T is the template image
I is the image to be classified
$|\cdot|$ denotes the absolute value.

If the difference, D, is less than a threshold, then the candidate image is classified as human.

3.2.2 Parzen Classifier

The second classifier tested was a Parzen classifier, using a small number of image features. The features used with the Parzen classifier are the mean, standard deviation, aspect ratio, the entropy and fill ratio (the ratio of foreground pixels to the total) of the images. It was decided to keep the number of features fairly small to reduce the computational requirements of the classifier. Normally a decision is made based purely on the posterior probability: an image is classified as human if the probability that it is human is greater than the probability that it is not. The consequences of false positives and false negatives are very different: a false positive will add an object that will need to be tracked, while a false negative may result in the system missing a potential collision. In order to adjust the sensitivity and false positive rates, an offset in the range of -1 to 1 exclusive is added. A negative offset will increase the probability that an image is classified as a human, i.e. it will result in an increased number of true positives but also increase the number of false positives. A positive offset has the opposite effect, biasing the classifier towards returning fewer false positives, as well as fewer true positives.

3.2.3 Principle Component Analysis

Most classification methods require the extraction of features from an image that are then classified. The Parzen classifier discussed above uses a number of high level features but finding sufficient, suitable features can be difficult and calculating them can be computationally expensive. In order to create an effective classifier the input images need to be mapped to a suitable feature space. The work by Turk and Pentland [83] shows that a face image can be classified and indeed uniquely identified by using principle component analysis (PCA). For the neural network and support vector classifiers, PCA is used to reduce the dimensionality of the feature space. Once the mapping has been determined using the training data, the feature can be determined from the test image using a single matrix subtraction and a single matrix multiplication. PCA was chosen above other feature extraction methods due to its computational simplicity since the detection system is required to run in real-time.

Figure 3.9 shows a plot of the magnitude of the eigenvalues calculated as part of the PCA. It is evident from the figure that the magnitude of the eigenvalues has decayed to an insignificant value by approximately the 80^{th} component. Since the magnitude of the eigenvalue represents the significance of the corresponding eigenvector to the data [84], the majority of the significant information in the data is represented by the first 80 components. For this reason the neural network and support vector classifiers were tested using 80 features.

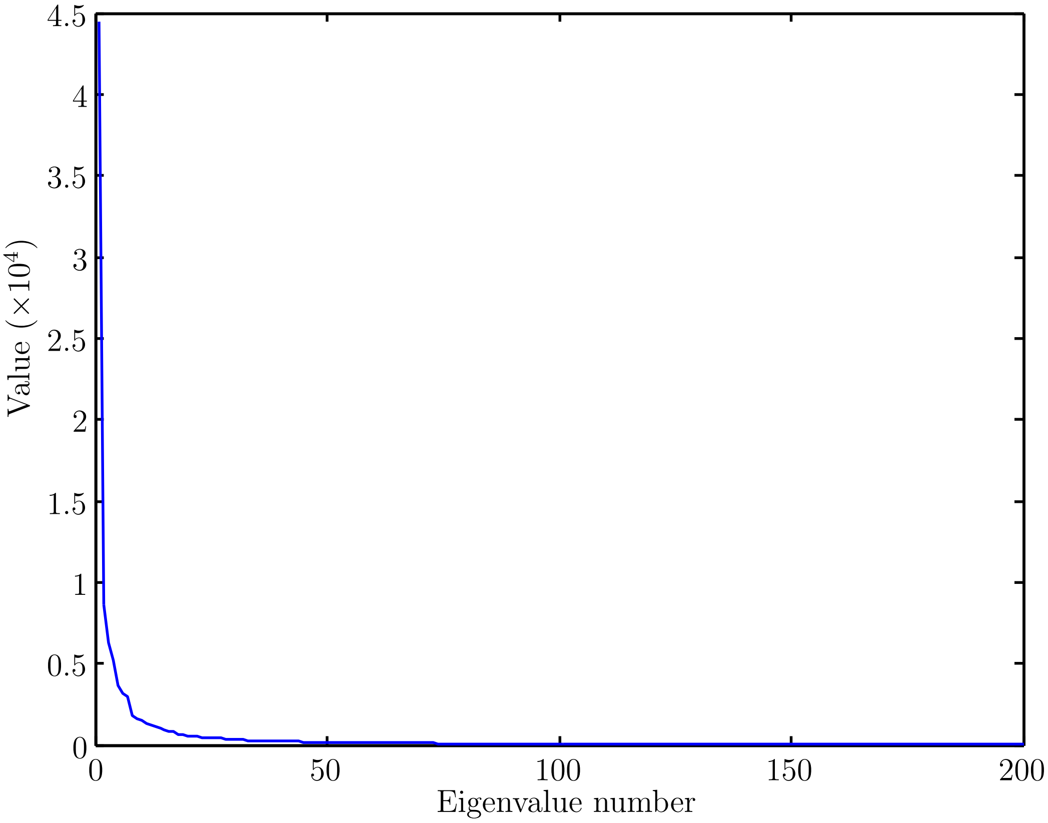


Figure 3.9: Plot of the magnitudes of the first 200 eigenvalues of the covariance matrix of the test data

3.2.4 Neural Network Classifier

The third classifier was a single hidden layer neural network. The input images are rescaled to produce a 48 × 20 pixel image which is then reduced to a 80-feature vector using the mapping obtained from the PCA.

The neural network that was tested is a single hidden layer perceptron with 12 hidden nodes and sigmoidal activation functions. The network has a single output node which outputs values in the range of (0,1) and is trained using simple back-propagation training. The mapping to the reduced dimension feature space that was determined from the training data is applied to the rescaled test image. The training data is also used to determine the mean and standard deviation of the feature vector. These values are used to normalise the training and test features before they are entered into the neural network.

3.2.5 Support Vector Classifier

The final classifier that was tested was a support vector (SV) classifier with the same data preparation as for the neural network. The input images are scaled and then their dimensionality is reduced and the feature vector is normalised. The support vector classifier implementation used is the popular support vector machine library LIBSVM [85] using the MATLAB interface. The linear, radial basis function and the hyperbolic tangent kernel functions were tested and it was found that the radial basis function kernel performed the best. The performance was optimised for the cost (the cost of a data point being on the wrong side of the decision hyperplane) and then the class weights were modified to produce a receiver operating characteristic (ROC) curve. The class weights scale the cost for each class, which means that the cost of misclassifying a class would be the cost multiplied by the weight for that class. Modifying the cost of making a false positive misclassification versus a false negative classification creates a ROC curve for various values of the class weight.

3.2.6 Comparison

The four initial classifiers were tested using a dataset consisting of 2800 images. The dataset consists of 677 isolated people, 372 partial people, 95 groups and 1656 non-human objects. Of the images 1988 were captured in a corridor and 812 in an underground mine. The classifiers were tested by randomly dividing the dataset into two sets of frames, a training set used to train the classifier and a test set used to evaluate its performance. As discussed in *Chapter* 5, simply dividing the dataset randomly produces a biased classification performance but all of the classifiers should be equally biased.

The ROC curves for the four classifiers are shown in *Figure* 3.10. The ROC curves provide two pieces of information about the classifiers. Firstly it is easy to see in *Figure* 3.10 that the SV classifier performs best, being on the upper left of the other classifiers at all points. The second thing that the curves are useful for is determining what value of the weight, threshold or probability offset will give the desired performance.

It is obvious that the classifier that should be chosen for the system is the SV classifier. The SV classifier achieves greater than 99% true positives at approximately 8% false positives on the initial dataset.

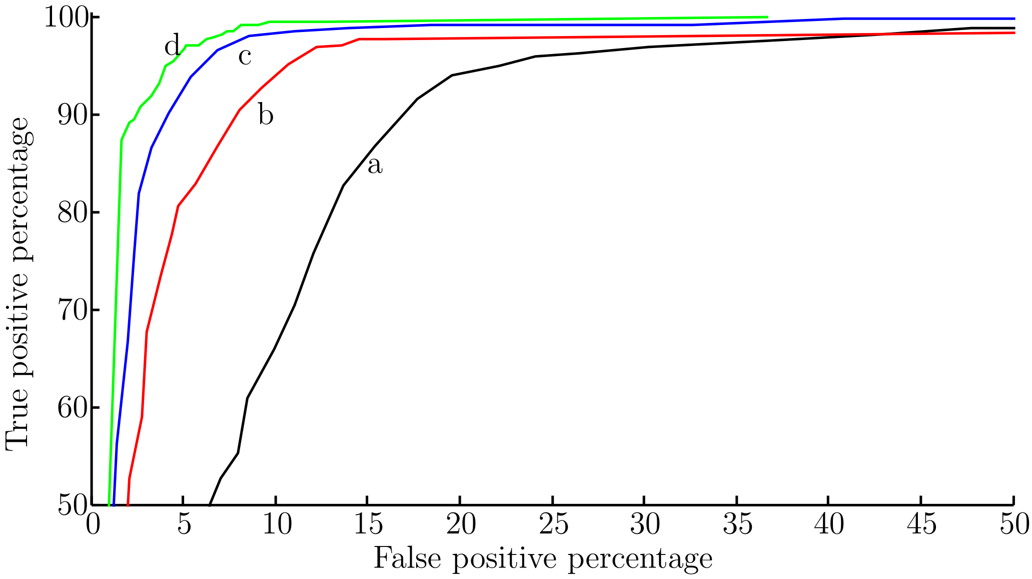


Figure 3.10: The ROCs of the four tested classifiers: (a) template classifier, (b) Parzen classifier, (c) neural network classifier and (d) SV classifier

3.3 Position Sensing

The initial prototype sensor contained a FLIR A300 thermal camera and two depth sensors, a SwissRanger SR4000 from Mesa Imaging and an Xbox Kinect, as shown in *Figure* 3.11.

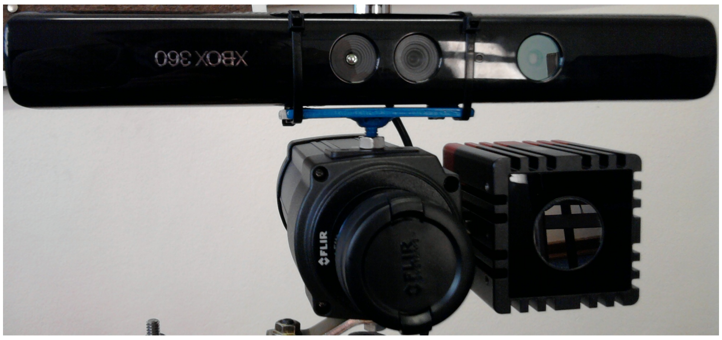


Figure 3.11: The prototype sensor showing the Kinect at the top, the SR4000 on the bottom right and the FLIR A300 in the centre

The SR4000 is a TOF camera that uses modulated near-infrared light to determine the distance to each point in its image. TOF cameras transmit modulated infrared light and measure the light reflected off the object. The camera measures the phase shift between the transmitted and reflected light using a specialised imaging sensor. The phase shift of the returning light allows the distance to be calculated using *Equation* 2.38. One of the advantages of the TOF camera is that it produces depth measurements with a bounded error (for the SR4000 the accuracy is ±10 mm). However, since each pixel is effectively independent of the others, the depth image tends to be spatially noisy. An advantage of a TOF camera is that it produces a complete depth image; each pixel will have an associated depth value. In an environment where there is the possibility of aliased depth measurements it is no longer true that the camera will produce a full depth image since a number of the readings can be invalid and need to be discarded (algorithms do exist to unwrap aliased measurements but they generally require multiple frames and are computationally very intensive).

Some work was performed to unwrap TOF images using a combination of an adaptive integration method with a reflectance estimation method. The adaptive integration method proposed by Strand and Taxt [86] estimates the phase shift that minimises discontinuities in the depth image. This method unfortunately attempts to 'unwrap' portions of the image that have large discontinuities but are not the result of phase wrapping. The intensity of light that returns from an object falls off proportionally to the square of the distance to the object; this can be used to determine whether a portion of a TOF image has experienced a phase wrapping. If phase wrapping occurred an object that is in fact far from the camera would be measured as being close to the camera; however, the intensity of the reflected light would be lower than expected from an object at that distance. This conflict between measured distance and amplitude can be used to unwrap a TOF image. For this research it was originally intended to integrate the two methods but this was not completed because it became evident that the Kinect was a better sensor with which to continue development.

The major disadvantage of TOF cameras is the fact that each phase measurement requires four separate measurements, which means that the effective integration time for each depth image is long. The long integration time produces blurred images with even slight movement, making it inappropriate for a moving vehicle. The blurring can be reduced by shortening the integration time of each exposure; however, this reduces the already limited 5 m range of the SR4000 further. Another disadvantage of the TOF camera is its susceptibility to airborne obscurants. This effect is discussed in more detail in *Appendix* C. The presence of reflective particles in the air causes a change in the net phase shift of the received light and hence a change in the measured distance.

The Kinect is a 3D imager with a built-in colour camera which is used as a controller for the Xbox 360 gaming console. The Kinect is built around

PrimeSense's PS1080 system on a chip [87]. The Kinect uses a triangulation technique to determine distance. A near-infrared light projector projects a pattern that is measured by a near-infrared camera. The distance to an object that reflects the light can be calculated using the angle that the light hits the camera and the distance between the camera and projector. One of the disadvantages of the structured light depth measurement method used by the Kinect is the fact that the accuracy decreases with distance. Since the camera is separated from the projector by some distance the Kinect also suffers from occlusion; parts of the scene illuminated by the pattern projector are not visible to the camera and near objects shadow further objects. The Kinect is a very low-cost sensor with a similar range to the SR4000; over a short range (up to about 2.5 *m*) the Kinect has a very high accuracy and good precision. The accuracy of a properly calibrated Kinect over this range is ±1 *mm* and the precision is approximately 15 *mm* (60th percentile) [88].

Owing to the low cost and good performance of the Kinect, and the results of some initial tests on the TOF camera, it was decided to use the Kinect as the depth sensor for the system. The range of both cameras is too short for the final system since the system will not provide sufficient time to warn the vehicle driver. Owing to hardware cost constraints the inclusion of a longer range sensor will remain future work.

3.4 Sensor Heads

Three sensor heads were used in this research for data capture. The first sensor consisted of three cameras, a FLIR A300 thermal camera, the Kinect and the SR4000 TOF camera. The three cameras shown in *Figure* 3.12 *a* were bolted together such that the SR4000 and FLIR A300 were back-to-back and they were perpendicular to the Kinect.

The initial tests using the SR4000 showed it to be unsuitable for use on a moving system so the SR4000 was removed. The A300 was rotated for the second sensor (*Figure* 3.12 *b*) so that it was parallel to the Kinect, ensuring the best overlap between the 3D and thermal images.

The first two sensors use custom capture software for capturing the 3D and thermal video to a separate notebook computer [89]. The software captures both image streams at maximum speed and the synchronised frames are extracted from the saved data. The high data output rate from the Kinect could not be saved to the hard-drive of the capture laptop fast enough to ensure that no frames where missed. For this reason the capture software buffers the video in RAM and then saves it when the buffer is full.

The third sensor was designed as a stand-alone sensor for capturing ther-

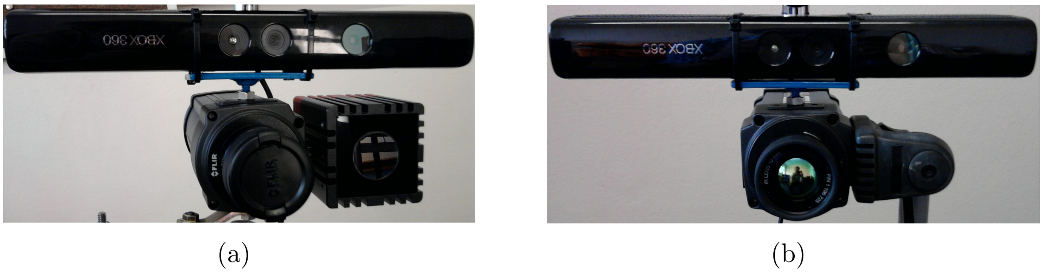


Figure 3.12: The first two iterations of the sensor head: a) with the SR4000 TOF camera and b) without the TOF camera

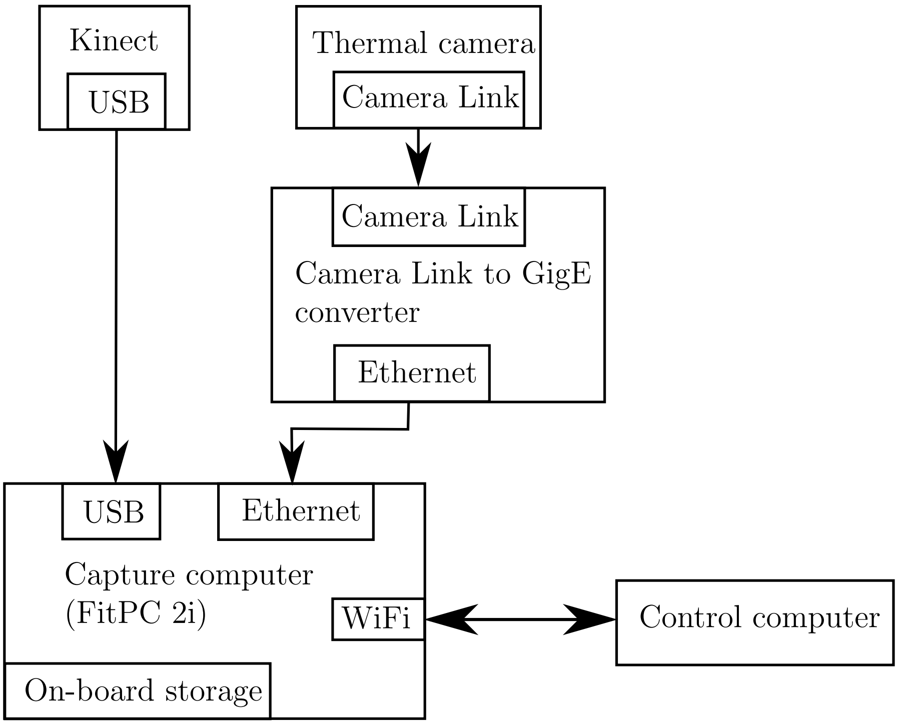


Figure 3.13: A block diagram of the third sensor version

mally textured 3D models of mine stopes, to assist in the assessment of hanging-wall stability. The stand-alone sensor includes co-aligned thermal and Kinect cameras and was used due to problems with the second sensor. For still unexplained reasons, the A300 of the sensor stopped streaming images and had to be returned for repairs. The repairs involved reloading the camera's firmware by the manufacturer. On the camera's return it was discovered that a frame-rate of 1.5 Hz was the highest achievable. Time constraints did not allow this problem to be resolved and it was decided instead to use the stand-alone sensor for the final tests. The third sensor version is a self-contained sensor system, a block diagram of which is shown in *Figure* 3.13.

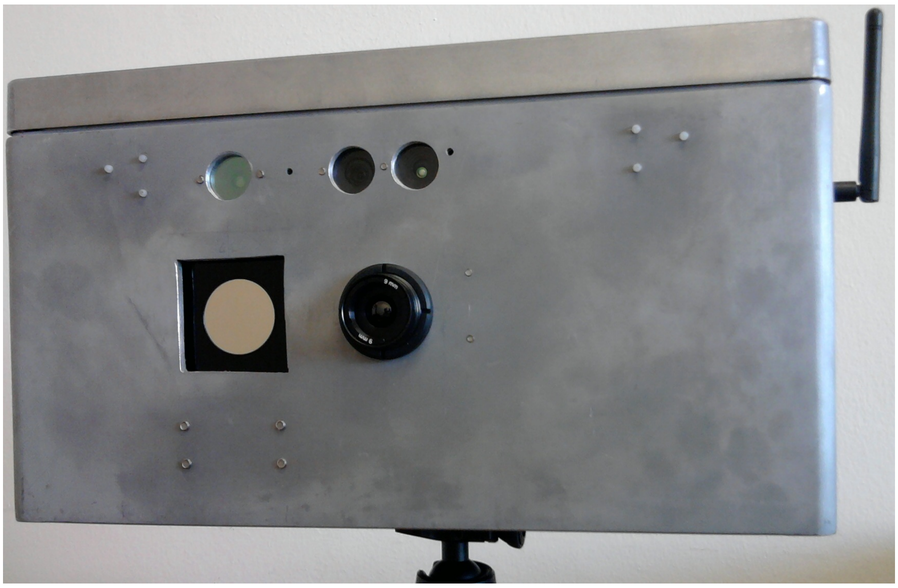


Figure 3.14: The third version of the sensor head

The new sensor is shown in *Figure* 3.14, with the three Kinect lenses visible at the top of the enclosure and the thermal camera, a FLIR TAU320 camera core, in the centre. The SR4000 on the left is currently unused.

The Kinect and the thermal camera both stream images to the capture computer, which stores them on-board for later retrieval and processing. The sensor is self-contained, running all the capture software on-board while being controlled by a remote machine over WiFi.

The capture computer runs software using the same buffer-then-save method of capturing images as the previous sensors. The computer has half the RAM of the notebook used to capture the thermal and depth data from the first two sensors. In order to reduce the size of the required video buffer the capture software was modified to add only synchronised frames to the buffer. Since both cameras are streaming and do not allow triggering (with current hardware) perfect synchronisation between the thermal and depth images is not possible. In order to capture the best synchronised frames the capture software records the most recent depth image when each thermal image is received. This ensures that the time difference between the frames is at most equal to the period of the Kinect capture (33 *ms*). It is possible to check the time stamp of the depth frames received before and after the thermal frame. This ensures that the time difference between the frames is at most half of the Kinect period. This method is, however, more complex and since we only know when a frame is received by the computer, not when it is captured, there is no guarantee that the synchronisation will be any better.

The ultra-low power Fit-PC does not have the processing power to effectively run the capture software through its MATLAB interface with a sufficiently low latency. This caused the third sensor to miss some Kinect frames, resulting in some time differences being greater than 33 ms.

Chapter 4

Final System

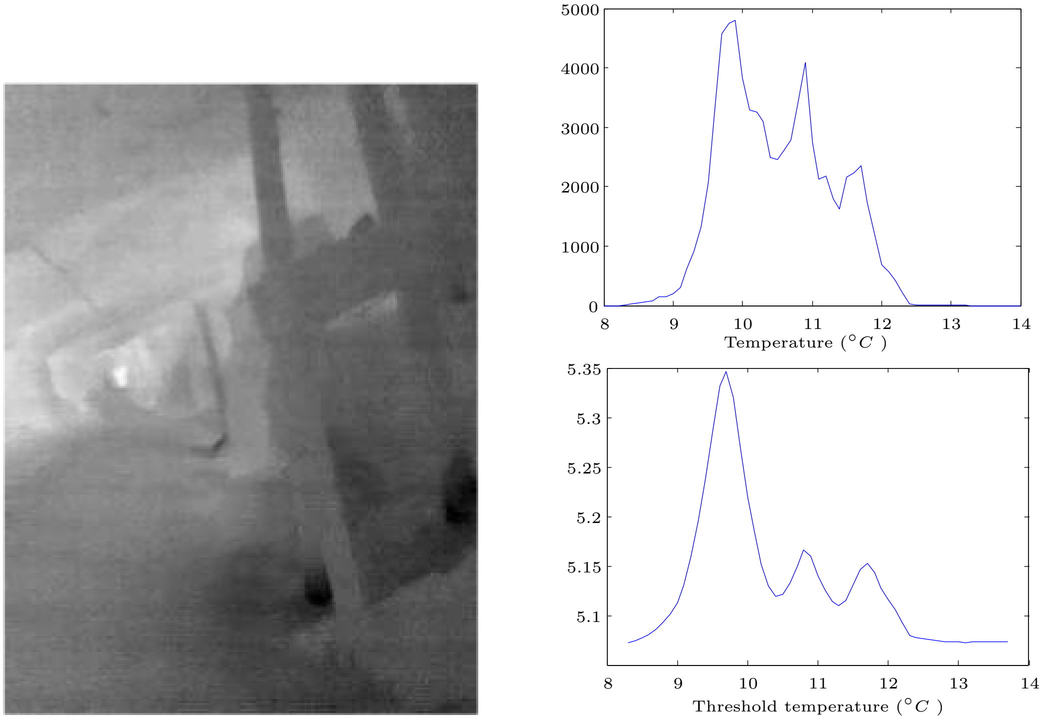
This chapter describes the final methods chosen for testing the pedestrian detection and tracking system.

4.1 Segmentation

The pedestrian detection and tracking system uses a modified version of Kittler and Illingworth's segmentation method for the segmentation of the thermal images. The modification to original algorithm allows the algorithm to handle images that do not contain people and return an image with no foreground objects, instead of trying to optimally segment the background.

The criterion function gives an estimate of the binarisation error if the image histogram consists of two normally distributed populations. The difference in value between the maximum and minimum of the function gives an indication of the difference in error between the best and worst thresholds. For an image that consists of fairly uniformly distributed intensities the range of the criterion function values will be small, while for an image with two well-defined distributions the criterion will have a deep minimum corresponding to the optimal threshold.

Figure 4.1 shows an image that does not contain any people, and has a unimodal histogram. The criterion function shown in *Figure* 4.1 illustrates that for an image with a unimodal distribution the criterion function does not have a true minimum. There are images that consist of a well-separated bimodal distribution but still do not contain any people; an example is shown in *Figure* 4.2. The minimum error criterion function in *Figure* 4.2 does have a well-defined, deep minimum. We know that a criterion function that does not have a deep minimum is unimodal and most likely does not have a person in it; the converse, however, is not necessarily true.

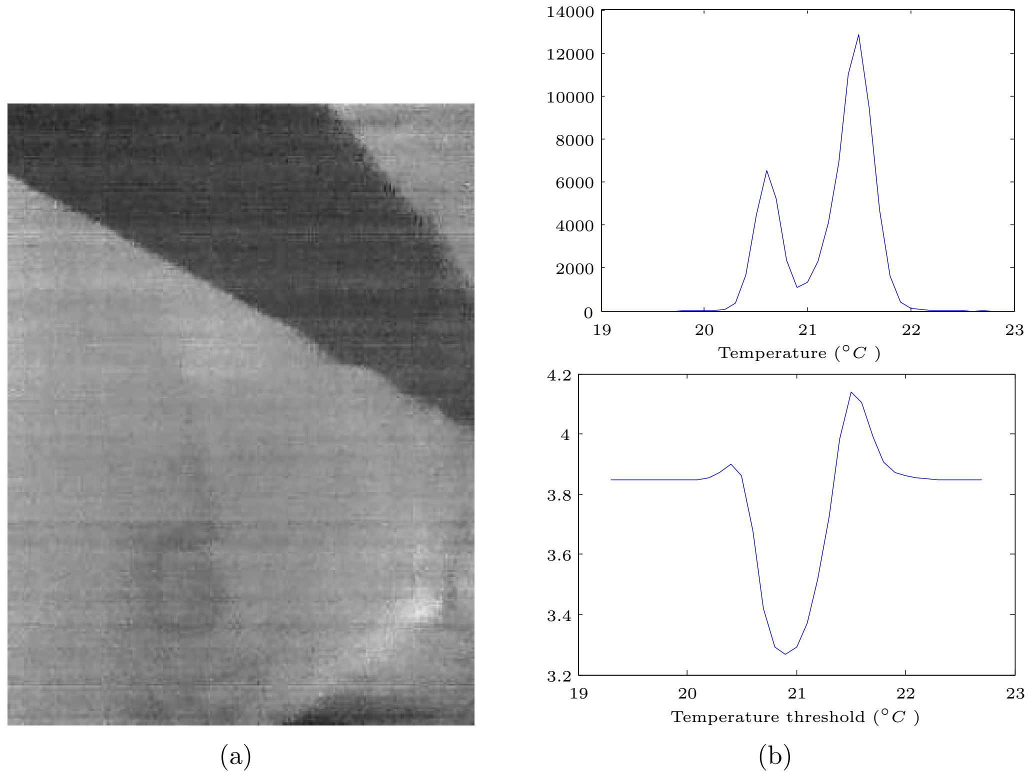


(a) (b)

Figure 4.1: a) An image with a unimodal distribution, b) The image histogram and criterion function

Since it is possible to have an image that does not contain any people but still has two well-separated distributions some additional information is required to determine whether the image should be segmented or whether it is empty. From the data it was found that the difference in the mean temperature of the background and people is at least 1.5 $^\circ C$. In the process of determining the minimum error criterion function, estimates of the foreground and background means are calculated. To determine whether an image may contain a person and should be segmented, a combination of the depth of the criterion function's minimum and the difference in the means is used.

After segmenting the image, a connected component analysis is performed and each individual region is numbered. Small regions of less than 160 pixels are ignored and the others are classified to verify which are actually human.



(a) (b)

Figure 4.2: a) An image with a bimodal distribution but no people, b) The image histogram and criterion function

4.2 Classification

As discussed in *Chapter* 3 the initial tests indicated that a SV classifier is the best classifier for identifying people in thermal images. It was initially expected that a cascade of classifiers, each trained to classify a specific pose, would provide better performance than a single classifier. This was tested and the indication was that there was common misclassification of the poses; groups of people were classified as single or partial people etc. This indicated that the features of the various poses were fairly similar and it was decided to try a single classifier for all poses in the training data. It was found that a single classifier performed better that the cascade classifier.

The final classifier was trained using 17 video sequences which were segmented and manually classified to produce a dataset of 3300 training images. The training images consist of: 834 images of isolated people, 412 of partially occluded people, 178 of groups of people and 1876 images of non-human objects.

Initially three classifiers in a cascade were used, one for each sub-class of pedestrian; single standing people, groups of people and partially occluded people.It was believed that using three classifiers would produce better results because each classifier would be classifying objects that had a higher degree of similarity.

The cascade was set up such that the classification was returned as soon as the corresponding classifier returned an affirmative. It was found that the classifiers commonly classified one sub-class as another; for example, a partial person was classified as a full person. This indicated that there was a significant degree of similarity between the sub-classes, so the three classifiers were combined. The combined classifier is computationally simpler and produced better results than the cascade.

Training the classifier on the full 3300 image training set indicated that reducing the number of components used for the PCA from 80 to 45 provided the best classification results.

Using this method each input image is scaled to 48×20 pixels and then transformed into a feature vector using the PCA feature vector calculated using the training data. Since the images are rescaled the original size of the image is lost; the aspect ratio, height and width of the image are, therefore, added to the feature vector with the PCA features. The feature vector is then normalised using the mean and standard deviation of the features in the training data.

4.3 Tracking

Tracking of the people detected in the images begins by determining the 3D positions of the people relative to the vehicle, using the information from the 3D camera. Determining the 3D positions of the people in the thermal image requires the projection of the 3D image into the 2D thermal image. Once the 3D image is projected into thermal image coordinates, finding the position of each object involves finding the 3D pixel that projected closest to the pixel of interest. The projection of the 3D points into the thermal image frame requires a calibration to determine the intrinsic parameters of the thermal camera and the relative transform from the Kinect reference frame to that of the thermal camera. The Kinect which is used for the 3D imaging produces a depth map which requires a further transform to calculate the 3D points.

4.3.1 Calibration

Kinect Calibration

The depth map produced by the Kinect is a map of the perpendicular distance (z-distance) from the Kinect to the point and not the distance along the ray like a TOF camera. From *Equations* 2.45 and 2.46 we have

$$x = \frac{z\,(u - t_u)}{f_x} \quad \text{and} \quad y = \frac{z\,(v - t_v)}{f_y} \; . \tag{4.1}$$

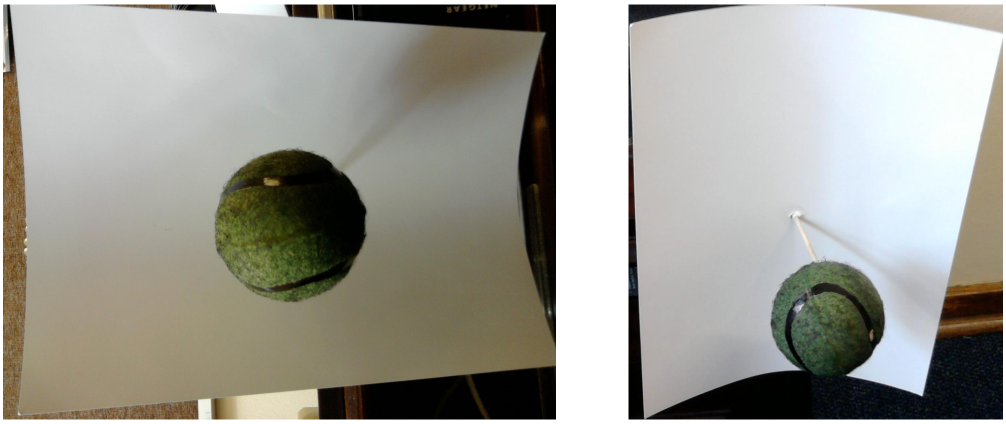
The focal length is different for the x and y axes to take into account the possibility that the pixels are not square. The calibration does not contain any distortion parameters because the addition of distortion correction was shown by [90] not to improve the re-projection error of the Kinect depth camera. The OpenNI API (`www.openni.org`) works together with the PrimeSense driver for the Kinect and calculates the depth from the disparity images produced by the Kinect, so the actual z distance can be used.

Kinect to FLIR Calibration

Once the Kinect has been calibrated it provides 3D positions for each pixel relative to its focal point. If the focal point of the Kinect is defined as a world reference frame then each 3D point can be projected into the thermal image using *Equation* 2.53. The calibration requires the identification of points in both the thermal and 3D images, which means that the standard checker-board calibration cannot be used. The initial calibration used a heated ball, as shown in *Figure* 4.3, which could be identified in both the thermal and depth images. A sphere always projects to a circle in an image so the ball can be identified by identifying circles in the images from the two cameras. The identification of the circles was achieved using a random sample consensus (RANSAC) circle fitting code developed by Price [91].

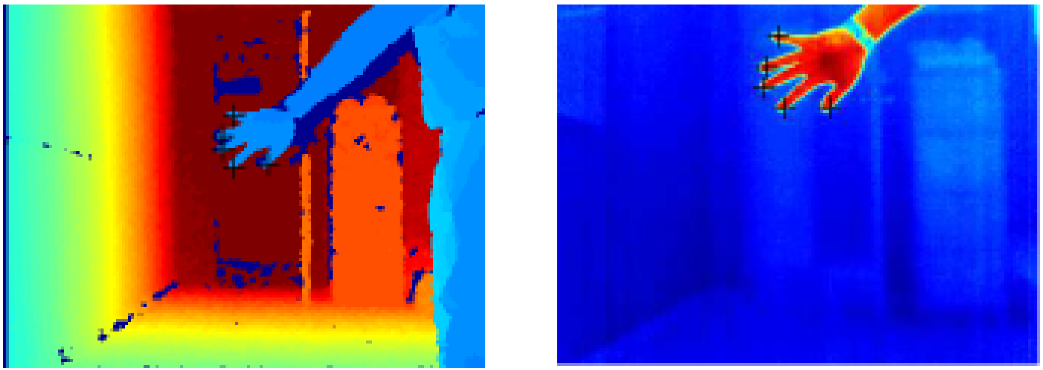
It was found that finding the ball in all images was difficult and required the tuning of the circle detector to accurately identify the ball. Owing to the difficulties associated with automatically detecting the ball's position it was decided to manually identify points that could be identified in both cameras. For the calibration the tips of a person's fingers were used because they are easily identified in both images, as shown in *Figure* 4.4.

A future improvement may be to produce a checker-board with squares that have significantly different emissivities in the long-wavelength IR band. The different emissivities would allow the corners of the blocks to be identified. Using the knowledge of the positions of the corners and the fact that the points all lie on the same plane allows the intrinsic and extrinsic calibration



(a) (b)

Figure 4.3: The initial spherical calibration object



(a) (b)

Figure 4.4: An example of a set of calibration images, a) the depth image and b) the thermal image. The corresponding points are indicated by crosses

parameters to be calculated as performed by Herrera, Kannala and Heikkila [90] for a depth and colour camera.

4.3.2 Trajectory Estimation

For the prediction of a collision the velocity of the person relative to the vehicle is sufficient. The problem of calculating the velocities of the people in the video sequence is a matter of associating each person in the current frame with the correct person in previous frames. The association is achieved using a nearest neighbour search based on the estimated positions of where the people should be in the current frame.

The estimation starts by identifying all the people in the current frame and determining their coordinates relative to the camera. For the first frame the velocity of all the people is set to zero, otherwise the velocity is calculated as follows. The position and the velocity of the previously detected people are used to estimate where they would be in the current frame. The distances between each current and previous position are calculated. The current person that is closest to the estimated position of the previous person is accepted as corresponding. The actual distance the person has moved between frames is used to estimate their instantaneous velocity.

If the distance between the current person and all of the previous people is greater than a threshold (1.5 m) then the person is assumed to be new in the scene and their velocity is initialised to zero.

A running buffer of previous velocities is filled and when sufficient estimates have been collected the estimate of the time to collision based on the average velocity is calculated. This is achieved by estimating the time to a collision (using the z-component of the distance and velocity) and determining the position of the person at the time when they would be in line with the vehicle. Two example cases are shown in $Figures$ 4.5 and 4.6. The first case is someone crossing the tracks slowly, such that they will be in the path of the vehicle when it gets to their position. The second case is where someone is crossing the tracks and will be safely across by the time the train gets to them.

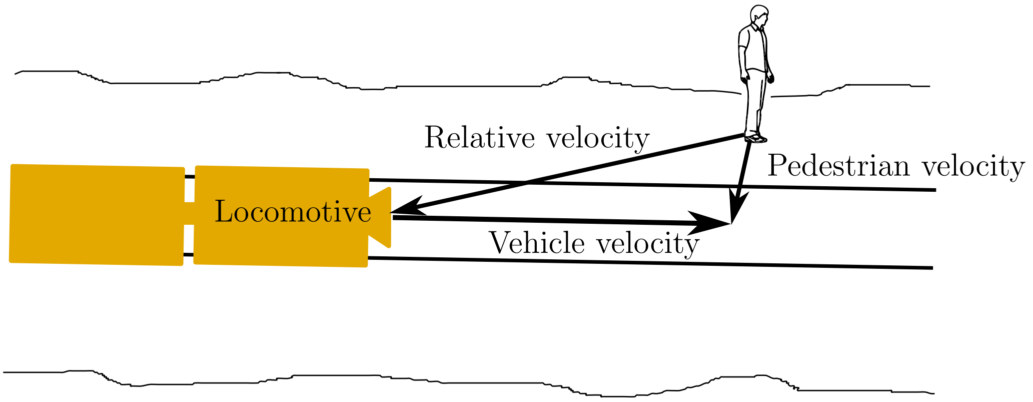


Figure 4.5: Figure showing a collision trajectories of a locomotive and a pedestrian

A difficulty arises in predicting a collision when the vehicle is on a curved trajectory; a simple collision estimate based on the relative velocity will not work in general because it is based on the instantaneous velocity of the

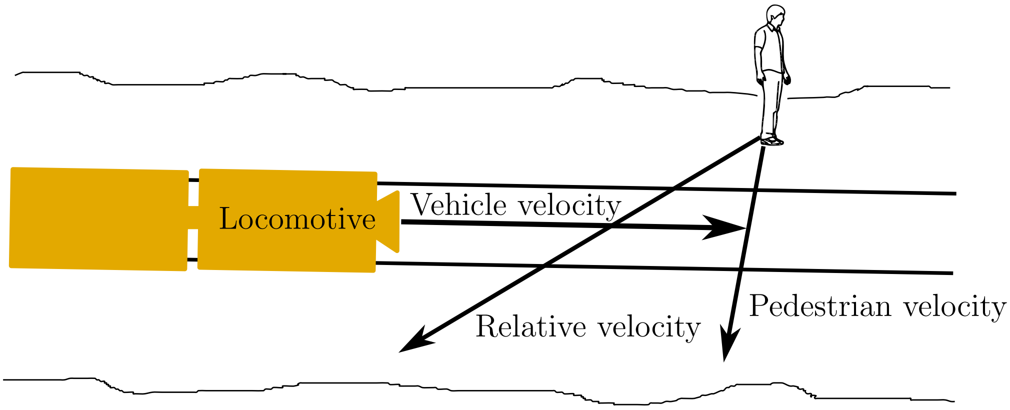


Figure 4.6: Figure showing safe locomotive and pedestrian trajectories

vehicle, which changes along the curve.

Figure 4.7 shows the problem with the locomotive travelling on a curved track. The instantaneous velocities of the locomotive and the pedestrian are shown; these will produce the relative velocity which the system will measure. According to the relative velocity the person will pass the vehicle well to its left; however, it is evident that the person will in fact end up on the tracks in the vehicle's path.

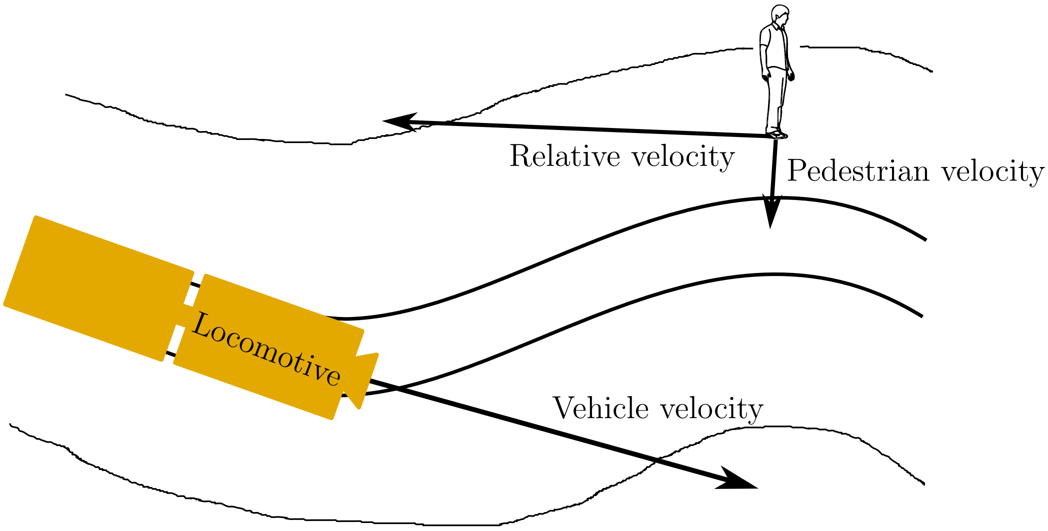


Figure 4.7: Figure showing the problem with using instantaneous velocities to predict collisions when the vehicle is on a curved track

Figure 4.8 shows that the error in the estimated x position at the time

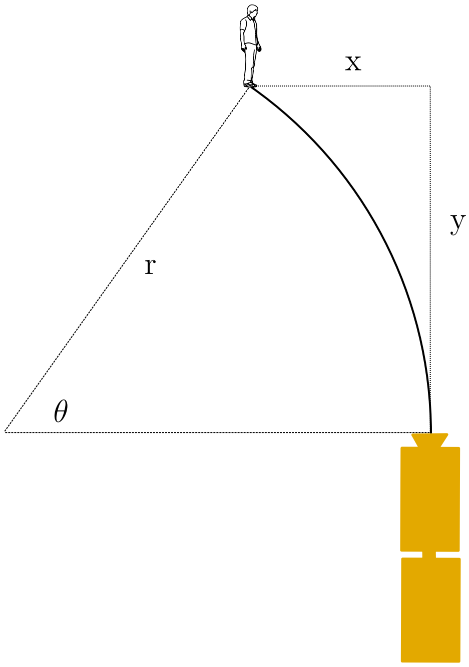


Figure 4.8: Geometry of a locomotive following a curved path

of impact will be incorrect by an amount that depends on the track radius of curvature and the distance between the locomotive and the person. It is evident from the diagram that

$$\theta = \arcsin\left(y/r\right) \tag{4.2}$$

and

$$x = r\left(1 - \cos(\theta)\right) \ . \tag{4.3}$$

Therefore the error in the estimated x position of an object is

$$x = r\left(1 - \cos(\arcsin\left(y/r\right))\right) \ . \tag{4.4}$$

As the distance between the vehicle and the pedestrian decreases it is evident that the error will decrease. According to Vermaak et al. [43], the minimum radius of curvature for underground rails is 15 m. This will give an error due to the curved path as shown in $Figure$ 4.9.

In order to improve the estimate of the trajectory of the vehicle and get a better estimate of the time to a collision, it is necessary to determine when the locomotive is on a curved trajectory. This could be achieved by detecting the tracks or by determining the angular velocity of the locomotive.

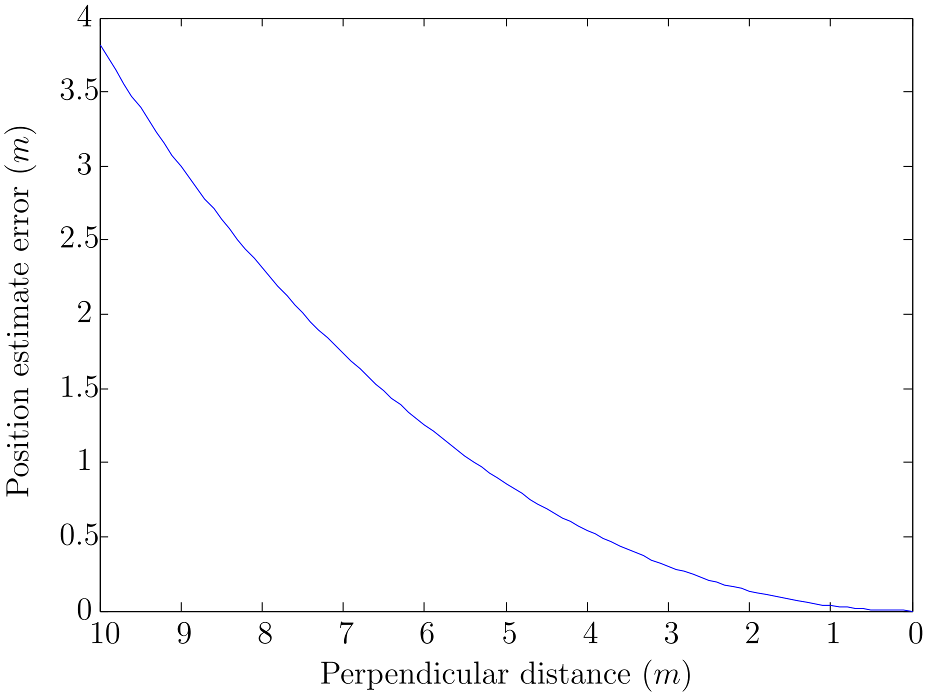


Figure 4.9: Plot of the x error versus the distance between vehicle and person, for a 15 m radius of curvature

Detecting the tracks is non-trivial using thermal or 3D imaging. The locomotive tracks are in thermal equilibrium with the rest of the tunnel footwall; therefore, the only detectable difference in a thermal image would be due to differences in emissivity of the rails and surroundings. Using the 3D image and extracting the tracks that are raised above the ground plane may work in some cases; however, it is common to have locomotive rails embedded in concrete in underground tunnels.

A simpler method would be to use the angular velocity of the locomotive to determine the curvature of the rails. It can be shown that for a vehicle travelling at a speed of V along a curved path, the radius of curvature of the path (r) is

$$r = \frac{V}{\omega} \ . \tag{4.5}$$

So if the locomotive is equipped with speed and yaw-rate sensors, the radius of curvature of its path can be determined and hence the error due to the curved path can be corrected.

The problem is that there is no way to determine the future trajectory of a person exactly. If, for example, the person is moving along the inside of a curved track at the same speed as a train on the track, their instantaneous velocity will project their path as crossing the tracks at some future time. So if we correct for the curvature of the track using the vehicle's velocity and angular speed the vehicle's trajectory will cross that of the person and a collision will be predicted.

If the vehicle moves much faster than the pedestrian, then correcting the path of the vehicle such as in [92] will work. The speed of underground locomotives typically ranges from 5 km/h to 16 km/h, which is not significantly faster than the average human walking speed of 5 km/h.

For a 15 m radius curve the locomotive speed should be less than 5 km/h (1.4 m/s) [43], so $Figure$ 4.9 can be replotted, as shown in $Figure$ 4.10, in terms of the time to collision.

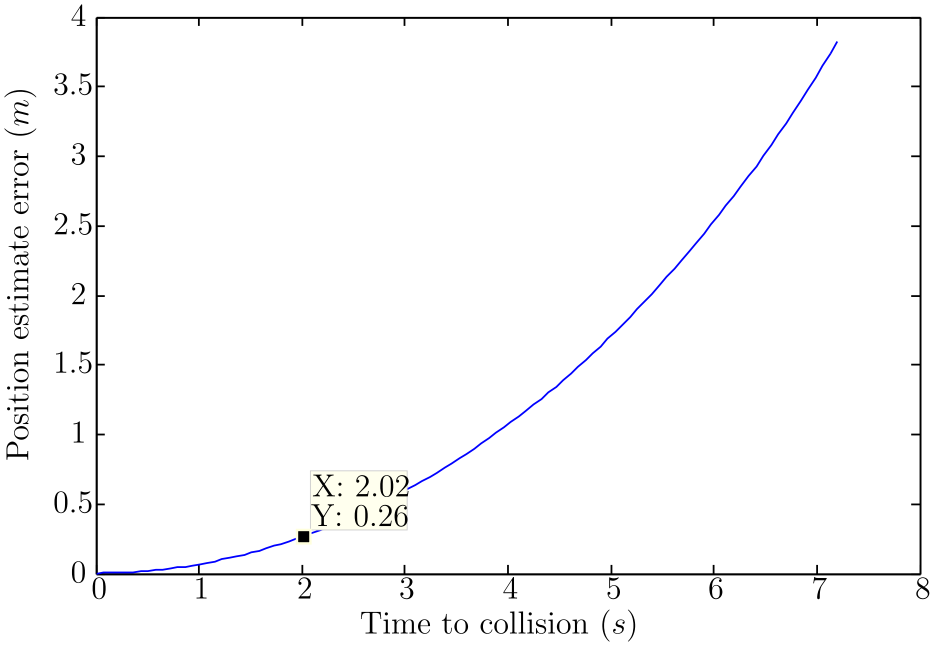


Figure 4.10: Plot of the x error versus the time to collision, for locomotive travelling at 1.4 m/s on a 15 m radius of curvature track

It can be seen from $Figure$ 4.10 that the error due to a curved path when the time to collision is 2 s is about 26 cm, which is less than the width of a person and therefore negligible. At the distance that would provide 2 s of warning of a collision the error due to a curved track is negligible. Only the instantaneous velocity is used for predicting a collision because correcting for the curvature of the tracks will not help without a better model of the pedestrians' trajectories. Additionally the curvature of the tracks becomes negligible when a collision is imminent.

Chapter 5

Results

5.1 Datasets

The data used for the training of the classifiers consists of datasets captured in an indoor passage environment, not unlike a mine tunnel, as well as in a tunnel of an out-of-service mine. The Kinect and the FLIR A300 are both streaming cameras; they cannot be synchronously triggered. Since the cameras cannot be triggered both are streamed at their maximum rate and a time stamp is saved with each frame when it is received by the capture computer. The Kinect frame that corresponds closest to each thermal image frame is used; since we do not know which Kinect frame will be closest to the FLIR frame we record every Kinect frame even though only one in 10 is actually used (the Kinect runs at 30 fps while the FLIR runs at 3 fps). The bandwidth of the Kinect data is immense. It streams 640×480, 16 bit images at 30 fps, which equates to a bandwidth of 147 Mb/s which is too high to capture directly to a laptop hard-drive without introducing delays that result in skipped frames. To handle the data bandwidth the images are cached in RAM and then written to a file when the capture is complete. The available RAM of the capture computer (a Dell Latitude 2110 notebook with 2 GB RAM) limited the capture duration to 40 s.

The final dataset was captured with the third sensor version. The new sensor consists of the Kinect mounted above a FLIR TAU320 thermal imaging core.

To test the robustness of the classifier two unseen datasets were used to test the classification. The first was captured in the tunnel of the mine at Gold Reef City (GRC), at a depth of 220 m, with an air temperature of approximately 16 $°C$. The dataset was taken using the second sensor with the A300 thermal camera. The dataset consists of 12 video sequences, with a total of 570 images.

The final dataset was captured using the third sensor version in First Uranium's Ezulwini gold and uranium mine. The dataset was captured in a small side tunnel on the 38^{th} level, at a depth of approximately 1000 m, with an air temperature of approximately 20 $°C$. The data was captured using the third sensor with the FLIR TAU320 thermal camera.

5.2 Segmentation Algorithm

The segmentation algorithm was used to extract the images for the training and test datasets. For all of the images the threshold was never such that a person in the image was not extracted. The minimum error thresholding does not always extract the entire person and does segment objects that are not people, hence the need for classification of the sub-images. The problem comes in when a person is in proximity with a background object. In cases such as this, the person and the background object are extracted as a single extended object. Having a single extended object consisting of foreground and background objects affects the classification results negatively.

Figure 5.1 shows an example of an image in which a person is segmented along with the background.

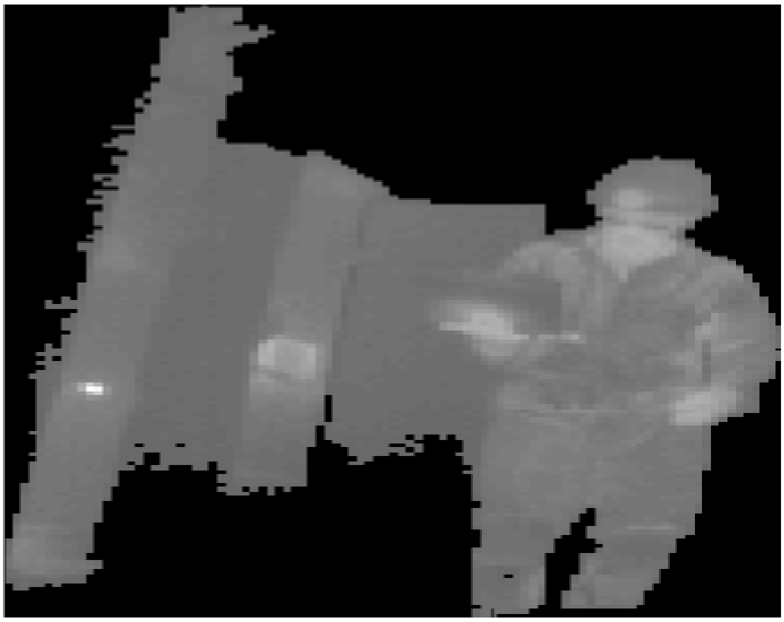


Figure 5.1: An example of a problematic segmentation containing both a person and background

Six out of the 13 missed detections for the GRC and Ezulwini datasets were images similar to *Figure* 5.1, which indicates that an improvement in the classification performance is possible with improved segmentation.

5.3 Classifier

The classifier was tested using four-fold cross-validation on a dataset of 3300 manually classified images. The training images consist of: 834 images of isolated people, 412 of partially occluded people, 178 of groups of people and 1876 images of non-human objects. The data was produced from 17 sequences containing a number of people with and without personal protective equipment. The background temperature varies from 11 $°C$ to 26 $°C$, in the different sequences. All the people in the dataset were aware that they were being videoed and gave verbal consent.

The classifier was validated using a four-fold cross-validation. K-fold cross-validation is performed by dividing the data into k equal (or approximately equal) sized subsets. For each of the subsets the classifier is trained on the $k - 1$ other sets. The classifier is then tested on the chosen subset. The classifier is tested on all of the k subsets; the performance for each test is averaged to determine the overall performance of the classifier.

As discussed previously the training data consists of a number of video sequences recorded in different environments with varying temperatures. Simply dividing the sequence into four equal sized subsets, consisting of images 1-825, 826-1650, 1651-2475 and 2476-3300, would not be acceptable. This is because there would be a high probability that the test datasets could come from a different environment to all of the training datasets and therefore bias the cross-validation results. In order to get a good spread of images from all of the environments, it was decided to select test and training images randomly. After further thought it was realised that randomly selecting the subsets would also introduce a bias.

Adjacent frames in a video sequence are correlated; there is a significant similarity between adjacent frames. Randomly selecting training frames from the data would result in a test subset that contained images that were very similar to those in the training subset.

To balance the requirements of separating the data into training and testing sets that are not too similar but still have training data from all the environments, the data set is divided into sets of 100 images. This is achieved by extracting 100 images from the data and then randomly assigning them to a subset. Cross-validation is then performed on the four subsets.

The performance of the classification algorithm is evaluated using its sensitivity, specificity and accuracy. The sensitivity (S_n) is the proportion of the human images in the training set that the classifier correctly identifies as human, i.e.

$$S_n = \frac{N_{TP}}{N_P} \tag{5.1}$$

where N_{TP} is the number of images correctly classified as people and N_P is the total number of images of people in the dataset.

The specificity (S_p) indicates the ability of the classifier to correctly identify background objects as not being human

$$S_p = \frac{N_{TN}}{N_N} \tag{5.2}$$

where N_{TN} is the number of correctly classified non-human images and N_N is the number of images of non-human objects.

The accuracy (A) is simply the fraction of images correctly classified

$$A = \frac{N_{TN} + N_{TP}}{N_N + N_P} \ . \tag{5.3}$$

As discussed in *Section* 4.2, a feature vector is created from each thresholded sub-image by performing a PCA and retaining the 45 most important components. The SV classifier is trained with a range of cost values and class weights. The cost applied to a positive feature point that is on the negative side of the decision plane is equal to the cost multiplied by the positive class weight. To reduce the number of variables that need to be optimised, the negative class weight is equated to the reciprocal of the positive class weight. The cross-validation results for the various class weights and cost values are shown in *Figures* 5.2 and 5.3.

The highest accuracy classification is achieved when the weights are 1.7 and 1 for the class weight and cost respectively; however, this corresponds to a dip in the sensitivity. The consequence of a false negative is significantly worse than the consequence of a false positive, so a high sensitivity is considered more important than a high specificity. *Figure* 5.2 shows the point that is considered the optimum, the point with weights that achieve the highest sensitivity without sacrificing too much accuracy.

The results of running a four-fold cross-validation on the classifier with a cost value of 5, a positive class weight of 9 and a negative class weight of 0.11 show the classifier achieves a good true positive rate and a fairly low rate of false positives (*Table* 5.1). A summary of the average performance of the classifier for the five tests is shown in *Table* 5.2. The figure shows that the classifier achieves a sensitivity (true positive percentage) of 97% and an overall accuracy of 95%.

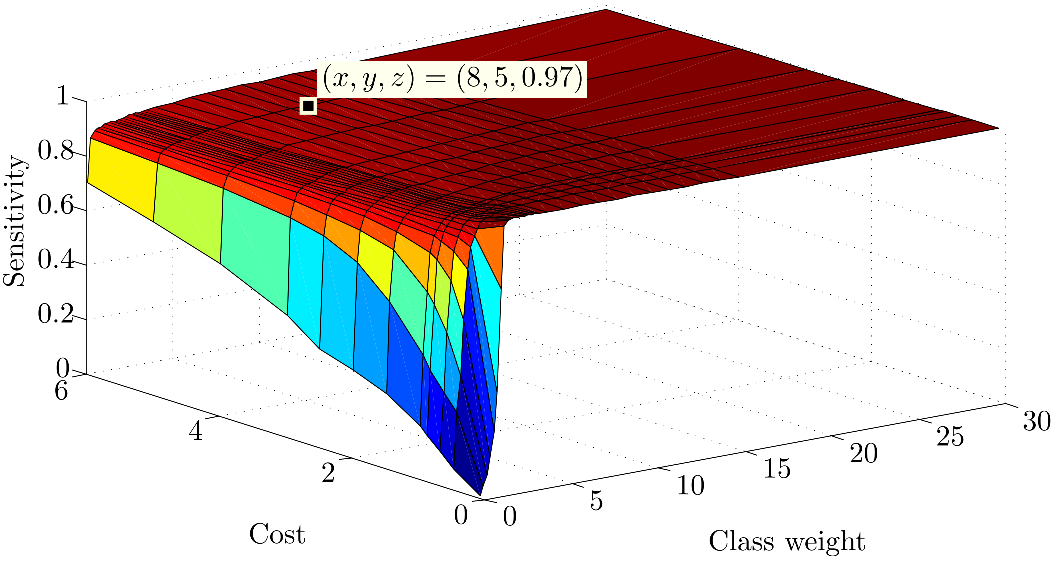


Figure 5.2: SV classifier sensitivity versus parameters

To test the robustness of the classifier, a further test was performed on a smaller unseen test dataset. The testing dataset consists of two datasets, the first from the GRC mine and the second from the Ezulwini mine. The two datasets contained examples of people who were crouching or lying down. These types of examples did not exist in the training dataset and were excluded from the classifier test, so only 937 of the 1030 images were used.

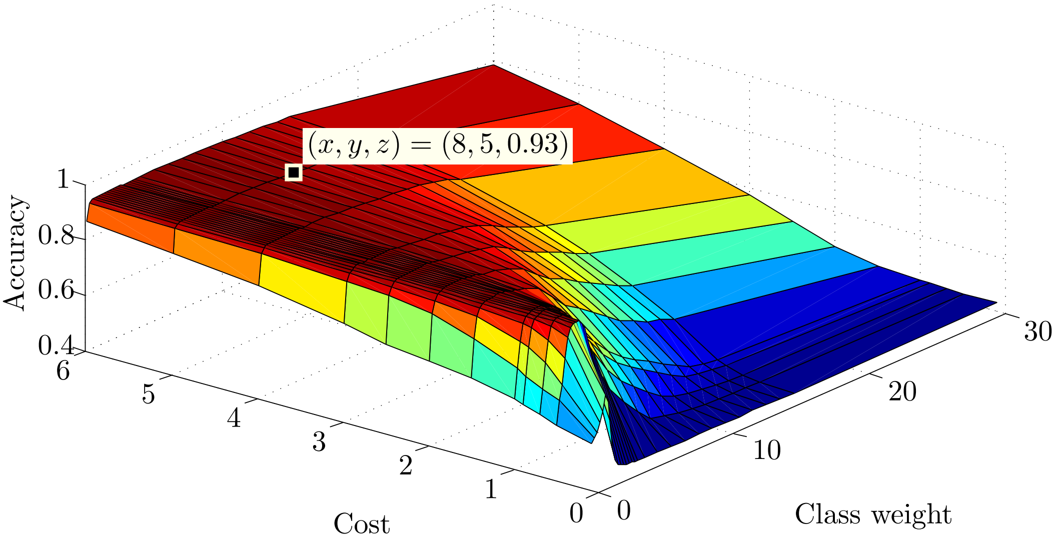


Figure 5.3: SV classifier accuracy versus parameters

Table 5.1: cross-validation results

Test	Humans	Non-human objects	True Positives	True Negatives
1	1424	1876	1390	1775
2	1424	1876	1387	1759
3	1424	1876	1388	1760
4	1424	1876	1388	1758
5	1424	1876	1394	1732

Table 5.2: Summary of the classifier's accuracy

Sensitivity	Specificity	Accuracy
0.976	0.937	0.953

The classifier that was trained using the training dataset was tested on the unseen dataset and the results are shown in *Table* 5.3.

Table 5.3: Unseen dataset test

Test	Humans	Non-human objects	True Positives	True Negatives
GRC	299	212	295	112
Ezulwini	126	300	117	207
Combined	425	512	412	319

It is evident from *Table* 5.4 that the performance of the classifier is reduced on the unseen dataset. It is interesting to note that the sensitivity is almost unaffected, only decreasing marginally, but the false positive rate has increased significantly. This result would indicate that the features of people do not change much in varying environments but the non-human objects vary with the environment. This is not unexpected since different environments have different types of warm background objects, such as lights, warm equipment, reflection off smooth surfaces etc.

The high specificity that can be achieved by the classifier both for the cross-validation results and the unseen data indicates that the chosen classifier can be used to classify pedestrians in thermal imagery effectively. The classifier maintains a good specificity even on the Ezulwini data, which is not only captured in a different environment but also with a different sensor. The reduced specificity of the classifier on the unseen data indicates that the

Table 5.4: Summary of the classifier's accuracy on the unseen dataset

Test	**Sensitivity**	**Specificity**	**Accuracy**
GRC	0.987	0.528	0.795
Ezulwini	0.929	0.690	0.761
Combined	0.969	0.623	0.780

background objects vary in various environments and therefore the classifier will need to be trained on data from the environment in which it will actually operate.

The system will operate on video data so it is necessary to determine the correlation between images and what the actual performance will be like on real video. If the classification of each image was independent the probability of missing multiple detections in a row would rapidly decrease. Unfortunately each video image is similar to those that are temporally close. What we would like to know is: what is the probability of missing a detection given that the classifier missed the previous one. If the classifications were independent this would be equal to the false positive rate of the classifier, approximately 3%. Looking at the missed detection, it is easy to see that the classifications are definitely not independent.

Appendix E contains tables showing the result of repeating the cross-validation three times and determining which of the misclassified images are consecutive. It can be seen from the tables in *Appendix* E that there are 29 false negatives where the previous image, in the video, was also a false negative. Given that the previous image was misclassified, the probability of misclassifying this image is approximately 0.26, so the probability of misclassifying two consecutive images is 0.6%. The worst observed case consists of nine consecutive misclassified images and occured once in the tests. The number of false negatives is too small to conclude much other than it is very unlikely that the classifier will miss more than nine or 10 images in a row. Using the TAU camera running at 8 fps it is, therefore, very unlikely that a person will not be detected within 1 s.

5.4 Tracker

For obvious safety reasons it is not possible to test the tracking system on an actual mine locomotive; therefore, the tracker was tested by mounting the sensor on an ERA-MOBI mobile robot. The robot was driven towards a stationary person and its velocity, angular velocity and position were recorded together with a time stamp. The velocity of the robot is used as a ground truth to test the accuracy of the tracker.

The distance between the person and the sensor, where the robot was approaching the person, is shown in *Figure* 5.4. The distance when the robot was retreating (a positive relative velocity) from the person is shown in *Figure* 5.5. It is evident from *Figures* 5.4 and 5.5 that there is significant variance in the measured distance for distances over approximately 5 m. This does make sense since the random error of structured light stereo is quadratically related to the distance and the Kinect is only designed to work up to a range of approximately 5 m. In *Figure* 5.5, the variation in the close measurements is due to the fact that the person is only partially visible in the image and is, therefore, missed by the classifier.

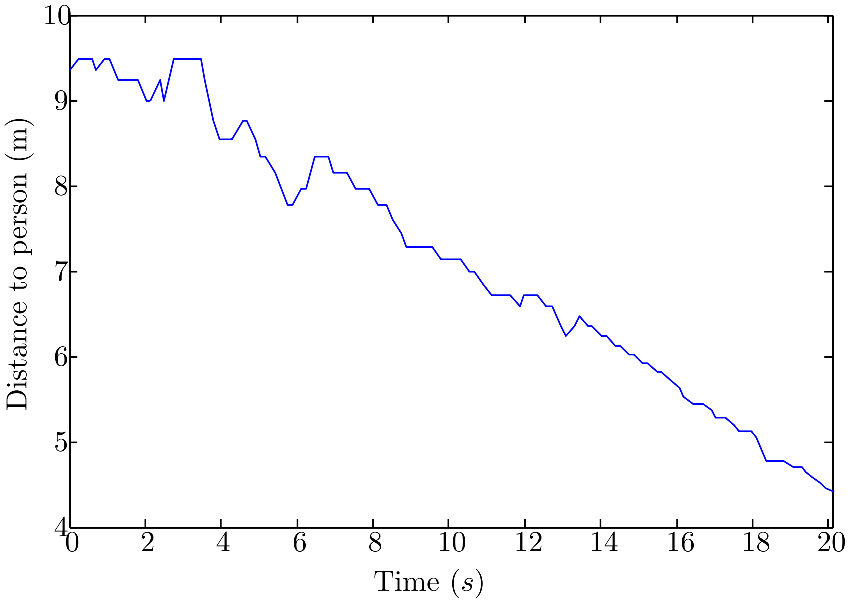


Figure 5.4: Plot of the measured distance to the pedestrian over time

The tracker estimates the relative velocity of the person from the vehicle using a moving average of the instantaneous measured velocities. The size of the filter can be adjusted based on the speed of the vehicle and the camera frame-rate. The filter for these tests averages approximately 2 s worth of measurements (16 frames). The calculated and measured velocities for the approaching and receding cases are shown in *Figures* 5.6 and 5.7 respectively.

Figures 5.6 and 5.7 show that even with a significant amount of filtering there is still a large amount of variation in the calculated velocity. It has been shown that the random error in the Kinect depth measurement increases quadratically with increasing depth [93]. The standard deviation of the depth

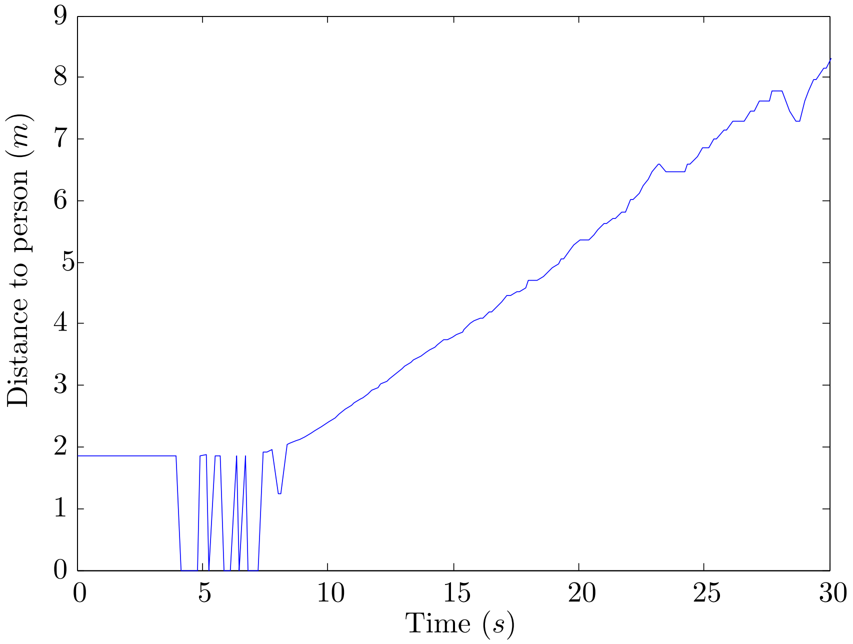


Figure 5.5: Plot of the measured distance to the pedestrian over time

measurements reaches approximately 4 cm at 5 m range. The quadratic relationship means that this would increase to over 14 cm at the maximum measured distance of 9.5 m. The high variance in the depth measurements is a significant component of the error in the estimated velocity, but it is not the whole story.

Let us look at the calculation of the velocity and the associated uncertainty. The velocity calculated before the running average filter is calculated using the distance between two positions of the person (Δx) at two different times (Δt), i.e.

$$V = \frac{\Delta x}{\Delta t} \ . \tag{5.4}$$

Using the propagation of uncertainty, the standard deviation in the velocity is

$$\begin{aligned}\sigma_V &= \sqrt{\left(\frac{1}{\Delta t}\right)^2 \sigma_x^2 + \left(\frac{\Delta x}{\Delta t^2}\right)^2 \sigma_t^2} \\ &= \sqrt{\left(\frac{1}{\Delta t}\right)^2 \sigma_x^2 + \left(\frac{V}{\Delta t}\right)^2 \sigma_t^2} \\ &= \left(\frac{1}{\Delta t}\right) \sqrt{\sigma_x^2 + V^2 \sigma_t^2} \qquad (5.5)\end{aligned}$$

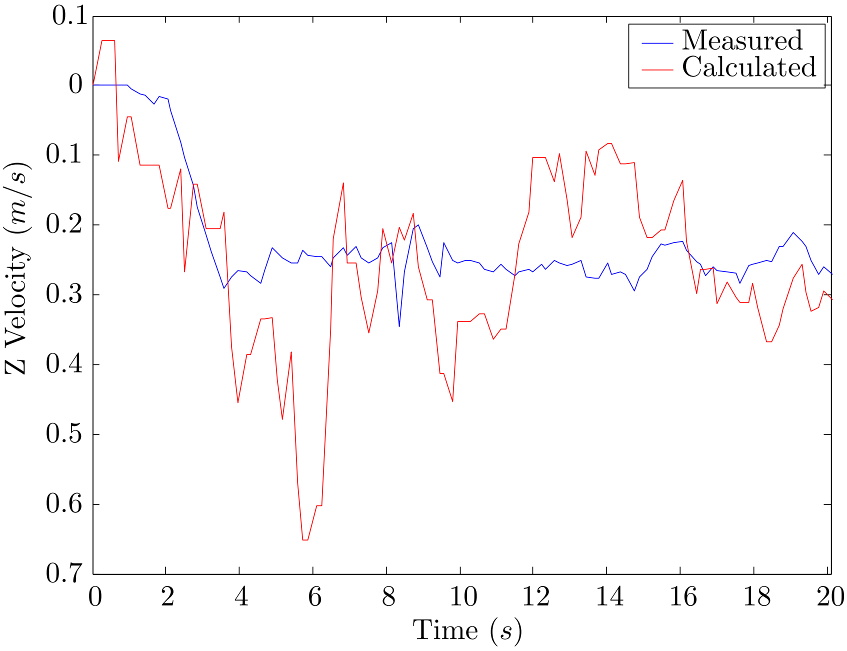


Figure 5.6: Plot of the measured and calculated velocities for the robot approaching the person

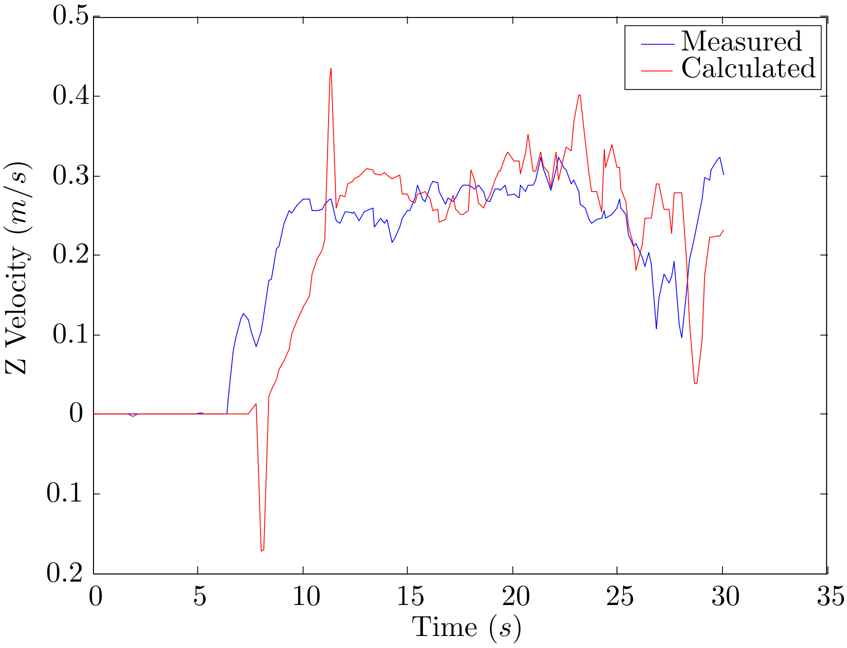


Figure 5.7: Plot of the measured and calculated velocities for the robot reversing

where σ_x is the standard deviation of Δx and σ_t is the standard deviation of Δt.

From *Equation* 5.5 it is evident that the faster the vehicle travels the more significant the variation in the capture frequency (or period) becomes. The current hardware has a large variance in the measured distance due to the accuracy of the Kinect and vibrations of the sensor.

The Kinect capture software runs a thread that constantly checks for new Kinect frames, and a second thread that waits for thermal images. When a new thermal image is received, it is saved to the frame buffer, along with the most recently captured Kinect frame. When each frame is received a time stamp is recorded. The time between when a Kinect frame is captured and when it is received by the computer depends on a number of factors such as the Kinect hardware and the load on the capture computer's processor. If we assume that the Kinect captures frames at a very well-defined rate, then the variance in the period calculated using the time stamps is equal to the variance of Δt. If we assume that the errors in the frame-rate and measured distance are independent of the velocity of the vehicle, then the standard deviation of the calculated velocity changes with the mean velocity of the vehicle as shown in *Figure* 5.8.

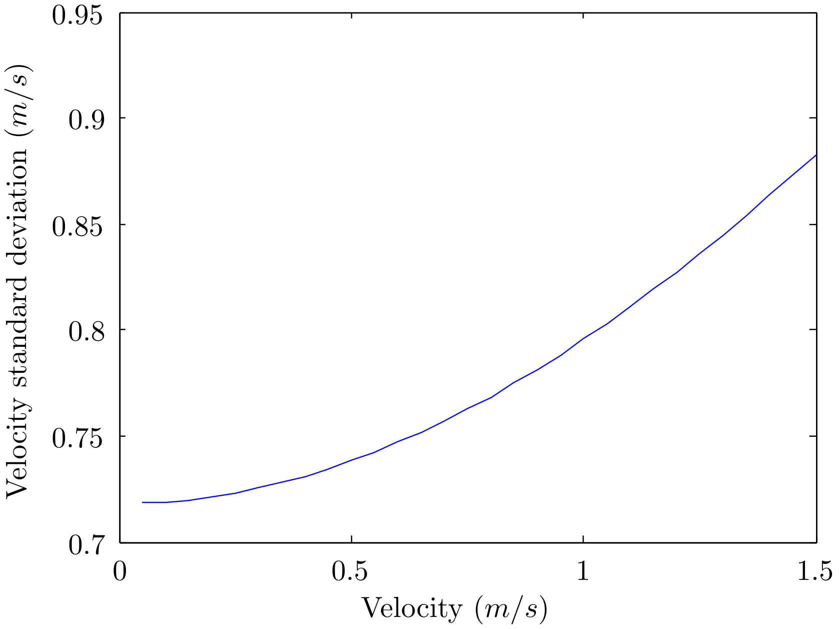


Figure 5.8: A plot showing the variance in the velocity estimate versus the mean velocity of the vehicle (assuming $\sigma_t = 60\ ms$ and $\sigma_x = 0.13\ m$)

From *Figure* 5.8 it is evident that the variance will increase with increasing velocity; however, as a proportion of the actual velocity the error in the measured velocity will decrease. A ten-fold increase in velocity only produces a 20% increase in the error.

The Kinect sensor is not acceptable for a final collision avoidance system because the useful range is less than 5 m, which is insufficient for anything but the slowest locomotive. Also it is not possible to know the exact time when the Kinect captures an image, which adds to the velocity estimate error. The two sources of error, being the error in the sampling frequency and the distance measurement, affect the velocity estimate. The effect of the sources of error is an important consideration for the replacement sensor for the Kinect. From *Equation* 5.5, it is obvious that the error in the distance measurement is important; what is interesting is that as the velocity to be estimated increases the error in the time-period becomes more significant. To ensure the best velocity estimate the Kinect replacement will need to provide an accurate time stamp of when the depth image was captured and produce depth values with a small error.

Another interesting observation from *Equation* 5.5 is that the error in the velocity estimate is proportional to the frame-rate that is used for tracking. This is not entirely unexpected. However, further consideration does produce a fairly counter-intuitive result. Consider two sensors that capture at different frequencies but are otherwise identical. Let us say the first sensor captures at a frequency of f and the second x times faster (xf). From *Equation* 5.5 we can see that the standard deviation of the measured velocity will be

$$\begin{aligned} \sigma_{Vf} &= xf\sqrt{\sigma_x^2 + V^2\sigma_t^2} \\ &= x\sigma_{Vs} \qquad (5.6) \end{aligned}$$

where σ_{Vf} is the standard deviation of velocity calculated using the fast sensor and σ_{Vs} is the standard deviation of velocity calculated using the slow sensor.

Now let us consider taking x velocity measurements from the fast sensor and averaging them so that they produce an average velocity measurement at the same rate as the slow sensor. It can be shown that the error of the averaged velocity measurements is

$$\begin{aligned} \sigma_{Va} &= \frac{1}{\sqrt{x}}\sigma_{Vf} \qquad (5.7) \\ &= \sqrt{x}\sigma_{Vs} \qquad (5.8) \end{aligned}$$

where σ_{Va} is the standard deviation of velocity calculated by averaging x measurements from the fast sensor.

Equation 5.8 shows that even taking multiple samples with the faster sensor and averaging them produces a worse result than the slow sensor would. For this result to hold the velocity needs to be constant for the period being used to calculate the velocity; therefore, we need to estimate the velocity as slowly as possible without the velocity changing in the period.

It was thought that tracking people using the full 30 Hz frame-rate of the Kinect would improve the tracking results. The thermal image could be used to identify people and then they would be tracked using the Kinect until the next thermal image was received. Based on *Equation* 5.8 it can be seen that it is actually unlikely that tracking with a higher frame-rate will improve the results.

5.5 Processing Time

All the processing times are for the algorithms running in MATLAB 2010b on a 3.4 GHz Core i7-2600 desktop PC, with 4 Gb of RAM and running Windows 7, 64 *bit*. *Table* 5.5 shows the system takes an average of 43 ms to process each frame, so even with the current implementation it is capable of operating at a frame-rate of 23 frames per second (fps) (without displaying the images). This speed is more than sufficient for the system to operate in real-time with the current hardware, which runs at 8 fps.

Subsystem	**Average time per frame** (ms)
Segmentation	13.6
Connected component	14.3
Classification	0.9
Calibration	10.1
Tracking	0.6
Other	3.5
Total	43

Table 5.5: Subsystem timing for the pedestrian detection system

The computer used for testing the algorithm processing times does not represent the processing power that would be available on a mobile platform; however, it is believed that the speed increase that would be gained by using a low-level programming language such as C or C++ would allow the same algorithms to be implemented on less powerful systems and still run at the same or higher speeds. A C implementation of the thresholding algorithm takes approximately 15 ms on a single board PC running an Intel Atom Z530, ultra-low power 1.6 GHz processor. The classification time is very short due to the efficient C/C++ implementation of the SVM library and it is believed that the processing times of the other subsystems can also be significantly reduced by implementing them in a language like C or C++.

Chapter 6

Conclusions and Recommendations

6.1 Conclusions

Underground mine vehicles are responsible for a significant percentage of mine fatalities. There is a need for a system that can warn a driver whether his vehicle is on a collision course with a pedestrian. A computer vision-based pedestrian detection system is proposed for pedestrian detection in underground mines because it allows the exact position of the people, near the vehicle, to be determined without the need for special equipment to be worn by workers.

The proposed system uses the fusion of thermal and 3D sensor data to detect and track people in the vicinity of rail-bound mine vehicles. The system uses a three-stage approach for detecting and tracking personnel. Initially the thermal image is segmented to produce regions of interest. The regions of interest are then classified to determine those that are actually people and should be tracked. The third step is to project the 3D image into the thermal image and determine the actual position of the person. Using multiple position estimates, the current relative velocity of the person with respect to the vehicle can be estimated.

It is shown that in a mine of up to 1000 m deep, there is still sufficient thermal difference between people and the surroundings to segment them. The segmentation algorithm performs acceptably to provide regions of interest for the classifier. The segmentation algorithm is shown to be able to segment people in environments where the background temperatures vary from 11 $°C$ to 26 $°C$ coupled with typical airflow. Environments with a high background temperature coupled with high velocity ventilation result in poor segmentation. This is because the high velocity ventilation reduces the temperature difference between people and the environment. The segmentation does not entirely miss people in the thermal image, but the regions of interest are fragmented. The air velocity is not expected to be high enough in an operating mine for this to be a significant issue (the problem was noticed in an underground training area of a mine).

The thresholding does not completely miss any people in the manually classified images but it does over segment regions, which causes the occasional missed detection of a person.

The classifier achieves an accuracy of over 95% and specificity of over 97%. The correlation between sequential video frames results in the chance of a false negative (missed detection) given that a previous one is higher than the overall 3% false negative rate. The chance of missing two consecutive frames is approximately 0.6% and the chance of missing someone for more than 1 s (8 frames) is very remote. The pedestrian detection system has a 99.4% chance of detecting a person within two frames (0.25 s), which compares favourably with the 99% detection rate shown for certain RFID tag systems.

A sensor built with currently available hardware has indicated that in principle a thermal imaging and 3D system can be used to detect and track people in an underground mine. The accuracy of the Kinect depth sensor is shown to be unacceptably low for effectively predicting a collision between a vehicle and pedestrian. The short range and poor accuracy of the current depth sensor does not allow the pedestrian detection system to provide an acceptable early warning of a collision trajectory.

The current implementation runs in real-time on an Intel i7 desktop PC. The conversion of the segmentation algorithm from MATLAB to C shows that the performance gained by implementation in a more efficient programming language will allow the system to also run in real-time on a mobile platform with limited computing resources.

6.2 Future Work

The most important future work is obviously the replacement of the Kinect with an accurate 3D imager. A multi-beam laser scanner, such as the Velodyne HDL-32E, may work well. The Velodyne, however, scans a full 360° , which is unnecessary and it is expensive (approximately R250 000). A 3D imager with a field of view similar to the Kinect but with a longer range is desired but a sensor with the desired characteristics at a feasible price has not been found. An alternative to a 3D imaging system could be to use a point laser distance sensor which is scanned to measure the distance to the objects of interest in the scene, i.e. the people. Such a system would be required to take far fewer measurements and would likely be more cost effective. A commercial system that could take spot distance measurements is not available and the design effort required means that this is not currently a viable option.

Future work may involve the incorporation of position information from the 3D imager to aid in the segmentation and classification of people. Currently the segmentation and classification are carried out entirely on the thermal image and then the depth image is used to track the objects identified as pedestrians. The addition of 3D information for segmentation will be a significant advantage for segmentation. One of the current limitations of the system is that if there is a person some distance in front of another group, the entire group will be segmented as a single entity. The addition of 3D information for segmentation would allow people that are close in the thermal image but are physically separated to be segmented separately. A similar problem also occurs with background objects that are segmented as part of a person, which results in images of people that contain a person and background, negatively affecting the classification results. Adding the 3D information for classification is also likely to improve the classification results because the classifier will be able to take into account the actual size of the object in the image.

Another possible improvement would be to use the vehicle's velocity and the measured relative velocity to predict the position where the pedestrian and the vehicle will meet. The width of the tunnel at the collision point can then be determined to check that there will be enough clearance on either side of the locomotive for the person to move out of the way without being crushed against the side wall.

A further improvement would be to use a more advanced state estimation technique for determining the velocity of the detected pedestrians. A popular option would be to use a Kalman filter which allows the unobservable states in a system to be estimated using noisy measurements. The performance of a Kalman filter is, however, dependent on the accuracy of the system model and modelling the dynamics of a moving person is very challenging.

Future work could include a warning based on a proximity alone, as a final safety warning. The system needs a number of frames to determine the trajectory of the person, so if someone suddenly moves out in front of the vehicle it would be useful to sound a warning immediately without first calculating the velocity.

The implementation of the system in a lower-level programming language, such as C or C++, will be required for the system to run in real-time on a mobile system.

Designing a calibration object that can be detected in both the 3D and thermal images to make automated calibration should be investigated. Automatically calibrating the cameras will simplify the calibration and probably improve the results.

Another future improvement could be implementing a motion model for the pedestrians that can predict if they are on a curved trajectory. If we can predict whether a pedestrian is on a curved trajectory, then correcting for the curved trajectory of the vehicle would improve the accuracy of the collision prediction.

References

[1] M. Creamer, "South African mining industrys 2013 zero-harm attainment unlikely - social scientist," Mining Weekly.com, December 2010. [Online]. Available: http://goo.gl/9FfO0

[2] A. Seccombe, "Decline in mine deaths 'too good to be true'," Business Day, January 2011, 2011/01/07.

[3] REUTERS. (2011, November) Mine health and safety audits Africa to be heightened. [Online]. Available: http://www.sabc.co.za/news/a/ea274900491857efbd9afde7f7089e4b/Mine-health-and-safety-audits-Africa-to-be-heightened-20111117

[4] First National Battery. (2012, Oct) Millenium loco batteries. [Online]. Available: http://www.battery.co.za/

[5] T. M. Ruff, "Advances in proximity detection technologies for surface mining equipment," in *Proc. of 34th AIMHSR*, Salt Lake City, 2004.

[6] National Institute of Occupational Safety and Health, "Proximity detection," August 2010. [Online]. Available: http://www.cdc.gov/niosh/mining/topics/topicpage58.htm

[7] T. M. Ruff and T. P. Holden, "Preventing collisions involving surface mining equipment: a gps-based approach," *Journal of Safety Research*, vol. 34, pp. 175–181, 2003.

[8] Becker Mining Systems, "Proximity detection and collision avoidance system," Product Information, June 2011. [Online]. Available: http://www.vale.com.au/runtime/cms.run/doc/English/1206/Collision_Avoidance_System.html

[9] P. Laliberté, "Summary study of underground communications technologies," CANMET Mining and Mineral Sciences Laboratories, Canada, Tech. Rep. CANMET-MMSL 09-004, May 2009.

[10] Mine Site Technologies, 2006. [Online]. Available: http://www.minesite.com.au/applications/proximity-detection/

[11] Minecom. (2008) Tracking and collision avoidance. [Online]. Available: http://www.minecom.net.au/solutions/tracking.html

[12] Trysome Auto Electrical, "Jannatec advanced warning system - 2-way collision avoidance," Product Information, May 2011. [Online]. Available: www.trysome.cc

[13] GoldFields, "Fact sheet as at march 2011," March 2011.

[14] F. Xu, H. V. Brussel, M. Nuttin, and R. Moreas, "Concepts for dynamic obstacle avoidance and their extended application in underground navigation," *Robotics and Autonomous Systems*, vol. 42, pp. 1–15, 2003.

[15] J. J. Green, P. Bosscha, L. Candy, K. Hlophe, S. Coetzee, and S. Brink, "Can a robot improve mine safety," in *25th International Conference of CAD/CAM, Robotics & Factories of the Future*, Pretoria, South Africa, July 2010.

[16] J. M. Roberts, E. S. Duff, and P. I. Corke, "Reactive navigation and opportunistic localization for autonomous underground mining vehicles," *Information Sciences*, vol. 145, pp. 127–146, 2002.

[17] J. J. Green and D. Vogt, "A robot miner for low grade narrow tabular ore bodies: The potential and the challenge," in *3rd Robotics & Mechatronics Symposium*, Pretoria, South Africa, November 2009.

[18] T. M. Ruff, "Innovative safety interventions: Feasibility of using intelligent video for machinery applications," National Institute of Occupational Safety and Health, Spokane, Washington, Tech. Rep., 2010.

[19] FLIR Commercial Vision Systems, "Avoiding accidents with mining vehicles," Application story, 2008.

[20] R. Paschotta, *Encyclopedia of Laser Physics and Technology*, 1st ed. Wiley-VCH, 2008. [Online]. Available: http://www.rp-photonics.com/infrared_light.html

[21] OptoIQ, "IR imagin: Short-wave IR offers unique remote sensing solutions," Apr 2006. [Online]. Available: http://goo.gl/X8uCH

[22] G. Brooker, *Introduction to Sensors for Ranging and Imaging*, 1st ed. SciTech Publishing, 2008, no. 978-1-901121-74-6.

[23] C. P. Sherman Hsu, *Handbook of Instrumental Techniques for Analytical Chemistry.* Prentice Hall, 1997, ch. 15, pp. 247–283.

[24] FLIR Systems Inc., "Thermal imaging cameras for automation & safety," Catalogue, June 2010.

[25] A. Fernández-Caballero, J. C. Castillo, J. Martínez-Cantos, and R. Martínez-Tomás, "Optical flow or image subtraction in human detection from infrared camera on mobile robot," *Robotics and Autonomous Systems*, vol. 58, pp. 1273–1281, December 2010. [Online]. Available: http://dx.doi.org/10.1016/j.robot.2010.06.002

[26] J.-J. Yon, L. Biancardini, E. Mottin, J. L. Tissot, and L. Letellier, "Infrared microbolometer sensors and their application in automotive safety," in *Proceedings of 7th International Conference on Advanced Microsystems for Automotive Applications*, Berlin, 2003.

[27] W. L. Fehlman and M. K. Hinders, *Mobile Robot Navigation with Intelligent Infrared Image Interpretation*, 1st ed. London: Springer, 2009.

[28] E. Goubet, J. Katz, and F. Porikli, "Pedestrian tracking using thermal infrared imaging," in *Infrared Technology and Applications XXXII*, B. F. Andresen, G. F. Fulop, and P. R. Norton, Eds., vol. 6206, no. 1. Orlando, FL, USA: SPIE, April 2006, p. 62062C. [Online]. Available: http://link.aip.org/link/?PSI/6206/62062C/1

[29] N. Pandya and J. van Anda, "Across the spectrum," in *SPIE oemagazine.* SPIE, Sept 2004, pp. 28–31.

[30] FLIR Commercial Vision Systems, "Uncooled detectors for thermal imaging cameras," Technical Note, October 2008.

[31] T. Enukova, N. Ivanova, Y. Kulikov, V. Malyarov, and I. Khrebtov, "Amorphous silicon and germanium films for uncooled microbolometers," *Technical Physics Letters*, vol. 23, pp. 504–506, 1997. [Online]. Available: http://dx.doi.org/10.1134/1.1261727

[32] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 288 – 299, 2007, special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[33] A. Yilmaz, K. Shafique, and M. Shah, "Target tracking in forward looking infrared imagery," *Image and Vision Computing*, vol. 21, no. 7, pp. 623–635, July 2003.

[34] H. L. Hartman and J. M. Mutmansky, *Introductory Mining Engineering*, 2nd ed. Wiley Publishing, August 2002, ch. 1, pp. 1–24.

[35] (2011, December) Glossary of mining terminology. Thompson Falls, MT, USA. [Online]. Available: http://www.rocksandminerals.com/glossary.htm

[36] J. J. Green, "Robots in mining south africas unique opportunity," Mobile Robotics Symposium, November 2010.

[37] J. R. F. Handley, *Historic Overview of the Witwatersrand Goldfields.* Howick: Handley, 2004, ch. 4.

[38] F. H. Von Glehn and S. J. Bluhm, "The flow of heat from rock in an advancing stope," in *Proceedings of the International Conference on Gold*, vol. 1. Johannesburg: SAIMM, 1986.

[39] S. Bluhm and M. Biffi, "Variations in ultra-deep, narrow reef stoping configurations and the effects on cooling and ventilation," *Journal of The South African Institute of Mining and Metallurgy*, vol. 101, no. 3, pp. 127–134, May 2001.

[40] W. M. Marx and R. M. Franz, "Determine appropriate criteria for acceptable environmental conditions," CSIR: Division of Mining Technology, DeepMine Research Task 6.1.1, June 1999.

[41] R. H. Clarke, D. Twede, J. R. Tazelaar, and K. K. Boyer, "Radio frequency identification (rfid) performance: the effect of tag orientation and package contents," *Packaging Technology and Science*, vol. 19, no. 1, pp. 45–54, 2006. [Online]. Available: http://dx.doi.org/10.1002/pts.714

[42] M. Jo, H. Y. Youn, S.-H. Cha, and H. Choo, "Mobile rfid tag detection influence factors and prediction of tag detectability," *Sensors Journal, IEEE*, vol. 9, no. 2, pp. 112 –119, feb. 2009.

[43] P. Vermaak, T. P. T. Page, A. G. du Plessis, W. J. Kempson, and C. F. P. Smythe, "Performance requirements for locomotive braking systems," Turgis Technology (Pty) Ltd, Safety in Mines Research Advisory Committee Report SIMGAP 635, February 2000.

[44] AngloGold Ashanti, "Mineral resource and ore reserve report," 2011. [Online]. Available: http://www.anglogold.com/Home

[45] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, "Pedestrian detection in infrared images," in *Intelligent Vehicles Symposium, 2003. IEEE*, 2003, pp. 662 – 667.

[46] Y. L. Guilloux, J. Lonnoy, R. Moreira, M.-P. Bruyas, A. Chapon, and H. Tattegrain-Veste, "Paroto project: the benefit of infrared imagery for obstacle avoidance," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, 2002, pp. 81–86.

[47] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Who? when? where? what? real time system for detecting and tracking people," in 3^{rd} *International Conference on Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 222–227.

[48] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, June 2002, pp. 15 – 20.

[49] S. M. Thornton, M. Hoffelder, and D. D. Morris, "Multi-sensor detection and tracking of humans for safe operations with unmanned ground vehicles," in *Workshop on Human Detection from Mobile Platforms*, Pasadena, May 2008.

[50] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 1, pp. 63 – 71, March 2005.

[51] N. Otsu, "A threshold selection method from grey-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.

[52] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, 2004. [Online]. Available: http://dx.doi.org/doi/10.1117/1.1631315

[53] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41 – 47, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320386900300

[54] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0734189X85901252

[55] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225 – 236, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320399000552

[56] D. Olmeda, C. Hilario, A. de la Escalera, and J. Armingol, "Pedestrian detection and tracking based on far infrared visual information," in *Advanced Concepts for Intelligent Vision Systems*, ser. Lecture Notes in Computer Science, J. Blanc-Talon, S. Bourennane, W. Philips, D. Popescu, and P. Scheunders, Eds. Springer Berlin, 2008, vol. 5259, pp. 958–969. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88458-3_87

[57] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121–167, 1998.

[58] L. E. Navarro-Serment, C. Mertz, and M. Hebert, "Pedestrian detection and tracking using three-dimensional ladar data," *The International Journal of Robotics Research*, vol. 29, no. 12, pp. 1516–1528, October 2010.

[59] Y. Le Cun, L. Jackel, B. Boser, J. Denker, H. Graf, I. Guyon, D. Henderson, R. Howard, and W. Hubbard, "Handwritten digit recognition: applications of neural network chips and automatic learning," *Communications Magazine, IEEE*, vol. 27, no. 11, pp. 41 –46, November 1989.

[60] H. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 1, pp. 23 –38, January 1998.

[61] V. Turchenko, I. Paliy, V. Demchuk, R. Smal, and L. Legostaev, "Coarse-grain parallelization of neural network-based face detection method," in *4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems*, September 2007, pp. 155 –158.

[62] L. Zhao and C. E. Thorpe, “Stereo- and neural network-based pedestrian detection,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, pp. 148–154, 2000.

[63] L. E. Navarro-Serment, C. Mertz, N. Vandapel, and M. Hebert, “Ladar-based pedestrian detection and tracking,” in *Proc. 1st. Workshop on Human Detection from Mobile Robot Platforms, IEEE ICRA 2008.* IEEE, May 2008.

[64] R. Siegwart and I. R. Nourbakhsh, *Introduction to Autonomous Mobile Robots*, 1st ed. Cambridge, Massachusetts: The MIT Press, 2004, ch. 4, pp. 122–128.

[65] J. Ens and P. Lawrence, “An investigation of methods for determining depth from focus,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 15, no. 2, pp. 97 –108, February 1993.

[66] A. P. Pentland, “A new sense for depth of field,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-9, no. 4, pp. 523 –531, July 1987.

[67] N. R. Gans, “3D machine vision tutorial,” University of Johannesburg tutorial, November 2009.

[68] T. S. Huang and A. N. Netravali, “Motion and structure from feature correspondences: A review,” in *Proceedings of the IEEE*, vol. 82, February 1994, pp. 252–268.

[69] D. Scaramuzza, F. Fraundorfer, and R. Siegwart, “Real-time monocular visual odometry for on-road vehicles with 1-point ransac,” in *ICRA’09: Proceedings of the 2009 IEEE international conference on Robotics and Automation.* Piscataway, NJ, USA: IEEE Press, 2009, pp. 488–494.

[70] G. Brooker, R. Hennessey, C. Lobsey, M. Bishop, and E. Widzyk-Capehart, “Seeing through dust and water vapor: Millimeter wave radar sensors for mining applications: Research articles,” *J. Field Robot.*, vol. 24, pp. 527–557, July 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1285289.1285291

[71] A. Jansson, “Sea-lynx gated camera - active laser camera system,” Datasheet, May 2004.

[72] J. Battaglia, R. Brubaker, M. Ettenberg, and D. Malchow, “High speed short wave infrared (swir) imaging and range gating cameras,” in *SPIE Thermosense XXIX*, 2007.

[73] P. Pencikowski, “A low cost vehicle-mounted enhanced vision system comprised of a laser illuminator and range-gated camera,” *SPIE*, vol. 2736, pp. 222–227, 1996.

[74] P. Andersson, “Long-range three-dimensional imaging using range-gated laser radar images,” *SPIE Optical Engineering*, vol. 45, no. 3, p. 034301, 2006. [Online]. Available: http://dx.doi.org/doi/10.1117/1.2183668

[75] P. Sphikas, *SR4000 User Manual*, 1st ed., MESA Imaging AG, Technoparkstrasse 1 8005 Zurich, March 2010.

[76] T. Ringbeck and B. Hagebeuker, “A 3d time of flight camera for object detection,” ETH Zürich, Tech. Rep., July 2007, optical 3-D Measurement Techniques.

[77] D. Droeschel, D. Holz, and S. Behnke, “Multi-frequency phase unwrapping for time-of-flight cameras,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 1463 –1469.

[78] L. Zhang, B. Curless, and S. M. Seitz, “Rapid shape acquisition using color structured light and multi-pass dynamic programming,” in *Proceedings of the First International Symposium on 3D Data Processing Visualization and Transmission*, November 2002, pp. 24–36.

[79] J. Salvi, J. Pags, and J. Batlle, “Pattern codification strategies in structured light systems,” *Pattern Recognition*, vol. 37, no. 4, pp. 827 – 849, 2004, agent Based Computer Vision. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320303003303

[80] B. Freedman, A. Shpunt, M. Machline, and Y. Arieli, “Depth mapping using projected patterns,” U.S. Patent US20 100 118 123A1, 2010.

[81] Y. Zhang, “A survey on evaluation methods for image segmentation,” *Pattern Recognition*, vol. 29, no. 8, pp. 1335 – 1346, August 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320395001697

[82] M.-P. Dubuisson and A. Jain, “A modified hausdorff distance for object matching,” in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, oct 1994, pp. 566 –568 vol.1.

[83] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[84] L. I. Smith, "A tutorial on principle component analysis," University of Otago, Tech. Rep., February 2002.

[85] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[86] J. Strand and T. Taxt, "Two-dimensional phase unwrapping using robust derivative estimation and adaptive integration," *Image Processing, IEEE Transactions on*, vol. 11, no. 10, pp. 1192 – 1200, Oct. 2002.

[87] PrimeSense, "Products - chip," online, 2010. [Online]. Available: http://www.primesense.com/?p=488

[88] Y. Liu. (2011, June) Kinect accuracy. ROS Wiki. [Online]. Available: http://www.ros.org/wiki/openni_kinect/kinect_accuracy

[89] M. Price, "Thermal imaging in 3d ii," Cogency, Tech. Rep., March 2011.

[90] D. C. Herrera, J. Kannala, and J. Heikkila, "Accurate and practical calibration of a depth and color camera pair," in *Conference on Computer Analysis of Images and Patterns, CAIP 2011*, Seville, Spain, August 2011, pp. 437–445, part II.

[91] M. Price, "Thermal imaging in 3d," Cogency, Tech. Rep., January 2011.

[92] T. Tsuji, H. Hattori, M. Watanabe, and N. Nagaoka, "Development of night-vision system," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 3, no. 3, pp. 203 – 209, sep 2002.

[93] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *sensors*, vol. 12, no. 2, pp. 1437–1454, Feb 2012.

Appendix A

Mining Fatalities

Table A.1 shows a breakdown of the major causes of mining fatalities in South African mines between May 2005 and March 2010. For the division between vehicle-related and machinery-related fatalities: machinery is equipment that is stationary or not self-propelled. Falls of ground and rock bursts involve the hazardous displacement of material from an in situ position.

Table A.1: A breakdown of the causes of mining fatalities for May 2005 to March 2010

Cause	**Number**	**Percentage**
Fall of ground / Rock burst	323	38.6
Vehicle related	152	18.2
Other	79	9.4
Fall from a height	57	6.8
Struck by machinery	50	6.0
Caught in machinery	38	4.5
Inundated by material	28	3.3
Explosion	27	3.2
Unspecified	23	2.7
Gassing	21	2.5
Struck by material	19	2.3
Electrocuted	18	2.2
Fall into machinery	2	0.2
Total	837	100

It is evident from *Table* A.2 that the cause of the majority of vehicle related fatalities is unknown or undocumented. *Table* A.3 shows the average number of workers killed annually in the various vehicle related accidents in mines.

Table A.2: Causes of vehicle related fatalities (May 2005 to March 2010)

Cause	Number	Percentage
Unknown	38	25
Frontal impact	27	17.7
Other	21	13.8
Two vehicles involved	19	12.5
Worker caught between vehicle and side wall	16	10.5
Worker caught between vehicle and obstacle	11	7.2
Vehicle rolled	11	7.2
Vehicle hit an obstacle	5	3.3
Reverse impact	4	2.6
Total	152	100

Table A.3: The average number of people killed annually in various vehicle related accidents (during May 2005 to March 2010)

Cause	Average annual fatalities
Unknown	7.1
Frontal impact	5.1
Other	3.9
Two vehicles involved	3.6
Worker caught between vehicle and side wall	3
Worker caught between vehicle and obstacle	2.1
Vehicle rolled	2.1
Vehicle hit an obstacle	0.9
Reverse impact	0.8
Total	28.5

Table A.4 shows the division of fatalities by mining sector. It is evident the the majority of mine deaths occur in gold mines, there are probably a number of factors that cause gold mine to have a high proportion of fatalities. Some of these reasons may be the depth at which gold is mined and the fact that gold mining relies mostly on manual labour meaning there is a large workforce being exposed to risk.

Table A.4: A table showing the number of deaths per mining sector between May 2005 to March 2010

Mining Sector	**Number**	**Percentage**
Gold	448	53.5
Platinum	185	22.1
Coal	81	9.7
Diamond	19	2.3
Iron ore	17	2.0
Chrome	14	1.7
Other	73	8.7
Total	837	100

Appendix B

Otsu's Threshold Selection

Let there be L grey-levels and the number of pixels at each grey-level, i, is n_i. The normalised grey-scale histogram is considered a probability distribution, such that

$$p_i = n_i/N \tag{B.1}$$

Where:
p_i is the probability that a pixel belongs to the i^{th} grey level
N is the total number of pixels

The zeroth- and first-order cumulative moments of the image histogram up to the k^{th} grey-level are:

$$\omega(k) = \sum_{i=1}^{k} p_i \tag{B.2}$$

and

$$\mu(k) = \sum_{i=1}^{k} ip_i \tag{B.3}$$

The total mean level of the original picture is:

$$\mu_T = \mu(L) = \sum_{i=1}^{L} ip_i \tag{B.4}$$

It can be shown that the between-class variance, σ_b^2, is:

$$\sigma_b^2 = \frac{(\mu_T\omega(k) - \mu(k))^2}{\omega(k)\,(1-\omega(k))} \tag{B.5}$$

Otsu's method selects the optimal threshold k_{opt} in order to maximise the between-class variance. The optimal threshold is the value of k that maximises *Equation* B.5, ie.

$$k_{opt} = \underset{k}{argmax}\ \sigma_b^2(k) \tag{B.6}$$

Appendix C

Time-of-flight Cameras

Initial tests performed underground show some significant disadvantages of using TOF cameras on a moving platform underground.

The first problem with the TOF camera is that a single depth measurement requires the camera to take four samples. The four images are used to determine the phase shift of the return signal and hence the depth. Since all four samples are need for a single depth measurement, an accurate measurement can only be achieved if the camera does not move during the acquisition. The actual integration time depends on the actual camera (the intensity of the modulated light). For the SR4000 camera, using a reasonable integration time of 12 ms (the range is 0.3 ms to 25.8 ms for the SR4000), the camera was found to produce blurred images for all but the slowest movements. For rotation it can be shown that the rotation speed must be below 5 $^{\circ}/s$.

Another problem with the TOF camera is that its accuracy is severely affected by aerosol obscurants. The drilling of blast holes in a mine gives off a fine water spray; coupled with high humidity this causes a mist in active areas of the mine. The TOF camera's amplitude image, in $Figure$ C.1, shows the water mist near the base of the support in the centre of the image. The distance image, shown in $Figure$ C.2, shows a significant jump in measured distances near the base of the support due to the mist there.

Investigating the operation of the TOF camera makes the reasons for this performance obvious. The modulated signal returning to the camera after reflection off an object will have the form of $sin(\omega t + \phi)$ where ϕ is the phase shift due to the time-of-flight to the object and back. If there are partially transmissive objects in the scene then the returning light is the sum of the light reflected off all the objects in the scene. So, assuming no phase wrapping, the returning light is

$$R = \int_{0}^{l_{max}} \rho\left(l\right) \sin\left(\omega t + 2\omega l/c\right) dl \tag{C.1}$$

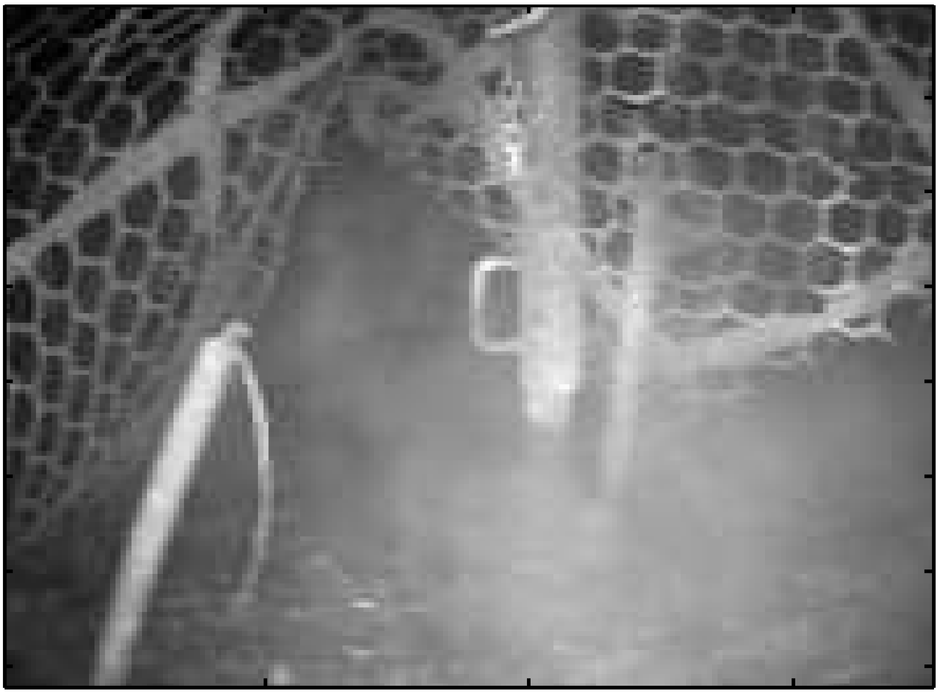


Figure C.1: TOF camera amplitude image through mist

where $\rho\left(l\right)$ is the reflectivity as a function of the distance from the camera (l). If we have an object with a reflectivity of ρ_0 at a distance of (l_0) then the returning light is

$$R = \int_0^{l_{max}} \rho_0 \delta\left(l - l_0\right) \sin\left(\omega t + 2\omega l/c\right) dl \tag{C.2}$$

$$= \rho_0 \sin\left(\omega t + 2\omega l_0/c\right) \ . \tag{C.3}$$

So the measured phase shift is $2\omega l_0/c$ and therefore the measured distance is l_0 as expected. Now if in addition to the above reflector at l_0, we have a dispersed reflector with a uniform per meter reflectivity of ρ_d dispersed from l_1 to l_2 then the reflectivity function (ignoring the attenuation through l_1 to l_2) is

$$\rho\left(l\right) = \rho_0 \delta\left(l - l_0\right) + \rho_d \left(u\left(l - l_1\right) - u\left(l - l_2\right)\right) \tag{C.4}$$

where $u(x)$ is the unit step function.

Substituting *Equation* C.4 into *Equation* C.1 and simplifying a slightly

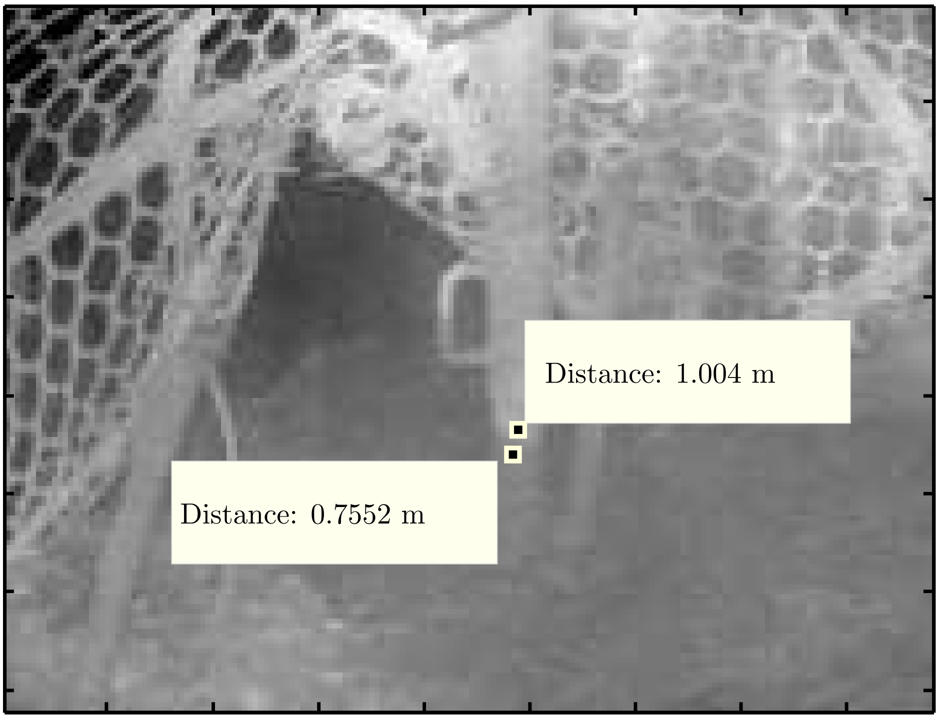


Figure C.2: TOF camera distance image through mist

gives

$$\begin{aligned}
R &= \rho_0 \sin(\omega t + 2\omega l_0/c) + \int_{l_1}^{l_2} \rho_d \sin(\omega t + 2\omega l/c)\, dl \\
&= \rho_0 \sin(\omega t + 2\omega l_0/c) + \rho_d \int_{l_1}^{l_2} \sin(\omega t)\cos(2\omega l/c) + \cos(\omega t)\sin(2\omega l/c)\, dl \\
&= \rho_0 \sin(\omega t + 2\omega l_0/c) + \rho_d \sin(\omega t)\, c/2\omega \left[\sin(2\omega l_2/c) - \sin(2\omega l_1/c)\right] \\
&\quad + \rho_d \cos(\omega t)\, c/2\omega \left[\cos(2\omega l_1/c) - \cos(2\omega l_2/c)\right] \\
&= \rho_0 \sin(\omega t + 2\omega l_0/c) + \rho_d c/2\omega \left[\cos(\omega t + 2\omega l_1/c) - \cos(\omega t + 2\omega l_2/c)\right] \\
&= \rho_0 \sin(\omega t + 2\omega l_0/c) \\
&\quad - \rho_d c/\omega \sin(2\omega l_1/c - 2\omega l_2/c) \sin\left(\omega t + \frac{2\omega l_1/c + 2\omega l_2/c}{2}\right) \ .
\end{aligned} \tag{C.5}$$

Substituting $\rho_1 = \rho_d c/\omega \sin(2\omega l_1/c - 2\omega l_2/c)$, $\alpha = 2\omega l_0/c$ and $\beta = \frac{2\omega l_1/c + 2\omega l_2/c}{2}$

into *Equation* C.5 gives

$$\begin{aligned} R &= \rho_0 \sin\left(\omega t + \alpha\right) - \rho_1 \sin\left(\omega t + \beta\right) \\ &= \rho_0 \left[\sin \omega t \cos \alpha + \cos \omega t \sin \alpha\right] - \rho_1 \left[\sin \omega t \cos \beta + \cos \omega t \sin \beta\right] \\ &= \left[\rho_0 \cos \alpha - \rho_1 \cos \beta\right] \sin \omega t + \left[\rho_0 \sin \alpha - \rho_1 \sin \beta\right] \cos \omega t \ . \end{aligned} \tag{C.6}$$

Defining

$$A \cos \theta = \rho_0 \cos \alpha - \rho_1 \cos \beta \tag{C.7}$$

and

$$A \sin \theta = \rho_0 \sin \alpha - \rho_1 \sin \beta \tag{C.8}$$

then *Equation* C.6 becomes

$$\begin{aligned} R &= A \cos \theta \sin \omega t + A \sin \theta \cos \omega t \\ &= A \sin(\omega t + \theta) \ . \end{aligned} \tag{C.9}$$

From *Equation* C.9 the net resulting phase shift is θ. Using the definitions above we know

$$\begin{aligned} \tan \theta &= \frac{A \sin \theta}{A \cos \theta} \\ &= \frac{\rho_0 \sin \alpha - \rho_1 \sin \beta}{\rho_0 \cos \alpha - \rho_1 \cos \beta} \\ \therefore \theta &= \arctan\left(\frac{\rho_0 \sin \alpha - \rho_1 \sin \beta}{\rho_0 \cos \alpha - \rho_1 \cos \beta}\right) \ . \end{aligned} \tag{C.10}$$

Instead of having a phase shift that depends on the distance to the object, we now have an equation that is a complex function of the distance to the object of interest, the thickness of the dispersed reflector, the reflectivity of the object and the dispersed reflectance. All these interactions make it impossible to correct the error introduced by the obscurants without additional information which cannot be measured.

Appendix D

Conference Publications

Below are conference papers presented as part of this work. The first was presented at IEEE Africon 2011 and the second was presented at the 4th Robotics and Mechatronics Conference of South Africa and the final paper was presented at the 26th International Conference on CAD/CAM, Robotics and Factories of the Future.

Pedestrian detection for underground mine vehicles using thermal images

J. S. Dickens
CSIR Centre for Mining Innovation
PO Box 91230
Auckland Park 2006
Johannesburg, South Africa
Email: jdickens@csir.co.za

M. A. van Wyk
University of the Witwatersrand
Private Bag 3
Wits 2050
Johannesburg, South Africa
Email: anton.vanwyk@wits.ac.za

J. J. Green
CSIR Centre for Mining Innovation
PO Box 91230
Auckland Park 2006
Johannesburg, South Africa
Email: jgreen@csir.co.za

Abstract—**Mine vehicles are a leading cause of mining fatalities. A reliable anti-collision system is needed to prevent vehicle-personnel collisions. The proposed collision detection system uses the fusion of a three-dimensional (3D) sensor and thermal infrared camera for human detection and tracking. In addition to a thermal camera, a distance sensor will provide depth information and allow the calculation of the vehicle and pedestrian velocities. The results of subsystem tests show that a simple temperature range is sufficient for segmentation and a neural network shows the best classification results in terms of speed and accuracy. Results of initial tests performed on two different 3D sensors show a significant disadvantage to the use of time of flight cameras in a mine environment.**

Index Terms—**mining, obstacle detection, human tracking, segmentation, thermal imaging, classification**

I. Introduction

Transportation machinery is responsible for a large portion of mine deaths in South Africa. After rock falls, vehicles are the second leading cause of mining fatalities. A reliable system for detecting people near mining vehicles is needed to prevent collisions between vehicles and personnel. The South African mining industry has committed itself to reducing the vast majority of serious of mine accidents and striving for zero fatalities by 2013 [1]. Given that the number of mining fatalities for 2010 was over one hundred, zero fatalities by 2013 is going to require significant improvements in mine safety systems.

The pedestrian detection system described in this paper is intended to assist mine vehicle operators by detecting a possible collision with a pedestrian and alerting the operator.

There are a number of existing of proximity warning systems for mining vehicles, using a number of detection technologies such as ultrasonic, laser, radar, global positioning systems (GPS), radio-frequency identification (RFID) tags, cameras or some combination of these [2–5].

Radar-based proximity detection is used for surface mining equipment as an aid drivers of dump trucks to detect people and small vehicles behind the truck. The system is fairly effective for surface mining equipment with only occasional false alarms [5]. The close proximity of tunnel walls in an underground mine makes the use of radar problematic owing to frequent false alarms [3].

GPS proximity detection has been proposed for surface mining operations. Each vehicle and worker broadcasts its position to nearby vehicles. A display in the vehicle shows the position of nearby people, vehicles and stationary objects and alarms if they are within a predetermined range [5]. The reliance on GPS signals precludes its use in a GPS-deprived underground environment.

RFID tags are popular for collision avoidance systems owing to their very low false alarm rates. RFID tag-based systems operating at various frequencies are used for a number of collision avoidance systems. The Becker NCS Collision Avoidance System and the Dynamic Anti Collision System (DACS600) use RFID tags operating in the 400 MHz frequency range while the HazardAvert Proximity Detection System and the Nautilus International Buddy system use low frequency magnetic fields [2, 4]. These RFID systems all operate on the same basic principle; each miner has an RFID tag (usually active) embedded in their cap-lamp. A transmitter mounted on the vehicle determines whether the tag is within a certain range of the vehicle and alarms or stops the vehicle if so. Some of the systems such as the HazardAvert system provide multiple zones, which provides a discrete distance measure. None of the systems provide the exact location of the personnel, merely how close they are.

A machine vision based pedestrian tracking system can address some of the shortcomings of current systems. Vision provides a way of detecting people and determining exactly where they are in relation to a vehicle. Machine vision has been investigated as a method for detecting people who are dangerously close to vehicles [5]. Thermal infrared (IR) imaging provides the advantages of vision based detection without the problems of sensitivity to illumination and obscuring dust. The illumination for thermal images is radiated by people and the long wavelength (7-14 μm) allows it to penetrate dust and smoke [6].

The IR spectrum can be divided into four main regions. The main regions are near-infrared, short-wavelength, mid-wavelength and long-wavelength IR [7]. Near-infrared (0.7 to 1.4 μm) is commonly used for light-based distance sensors such as laser scanners and Time Of Flight (TOF) cameras. Near-infrared illumination is also often used for night-vision surveillance since it can be detected using the same imaging sensor used for visible light. Short-wavelength IR is used for various process monitoring and inspection tasks such as hot furnace monitoring [8]. Mid-wavelength IR can be used for gas spectroscopy [7]. Long-wavelength IR (or thermal IR) is the region of interest for this paper and is used for thermal imaging.

In Section II of this paper the basic architecture of the proposed pedestrian detection system and the major subsystems is described. The results of tests to evaluate the segmentation and classification algorithms and the distance sensors are presented in Section III. Finally the results are discussed and conclusions are drawn.

II. System Architecture

The detection system first extracts regions of interest (ROIs); these are regions that have a temperature that would possibly allow them to be human. The ROIs are then classified as being human or background objects. A distance sensor provides the three-dimensional (3D) position of the person for the tracking system. The tracking system provides the trajectory of the people in the camera's field of view. A sensor head consisting of a FLIR A300 thermal camera, a SwissRanger SR4000 TOF camera and an Xbox Kinect was used for data gathering. The background excluding pedestrians is assumed to be stationary and is used to determine the trajectory of the vehicle. The vehicle trajectory will be estimated using the established iterative closest point surface matching algorithm. Using the trajectory of the vehicle and the pedestrians the system calculates whether a collision will occur.

A. Thermal Image Segmentation

The system first extracts Regions Of Interest (ROIs) that could be human which are then classified. The thermometric image provided by the A300 allows segmentation of the image based on an empirically determined temperature threshold. As discussed in Section III-A the temperature based segmentation outperforms more complex algorithms on the indoor data.

Virgin rock temperatures of deep South African gold mines are in the region of 60 $^{\circ}C$ however ventilation and other cooling brings the temperature within working areas down to below 30 $^{\circ}C$ to allow work to be done [9]. Work conducted to model the heat flow from advancing stopes shows that the rock surface temperature can be assumed to be equal to the ventilation air wet-bulb temperature (T_{wb}) [10]. Significant work has been performed to design ventilation systems to ensure the air T_{wb} remains below 28 $^{\circ}C$ (heat stress management programmes are required for $T_{wb} > 27.5$ $^{\circ}C$) [11, 12]. Therefore, it is assumed that the rock temperature within the mine tunnels will be below 28 $^{\circ}C$.

B. Classification

There are a number of methods used to classify humans in thermal images. To the authors' knowledge, there has not been a quantitative comparison of methods for human classification in thermal imaging. In the absence of a clear choice, it was decided to compare three different classification modalities. The three classification methods are: 1) an appearance-based classifier using a template match. 2) A feature-based classifier which uses a number of features extracted from the image which are classified using a Parzen classifier and 3) a neural network classifier. Each of these are discussed in turn below.

1) Template classifier: Template-based classification has been used for human detection in thermal images from moving vehicles [13, 14] Nanda and Davis [13] use a probabilistic template created from training images while Bertozzi et al. [14] use a greyscale morphological template. It was decided to use a method similar to Bertozzi et al.'s except to use a template created from training images. The images of humans in the training data are rescaled to form a $M \times N$ image (in this case 30×12). A template is created by taking the mean of the scaled images. The candidate regions are rescaled to the same dimensions as the template and the two are compared using an absolute difference distance measure, ie.

$$Difference = \sum_{i=1}^{M} \sum_{j=1}^{N} abs(T_{ij} - I_{ij}) \tag{1}$$

Where:
T is the template image.
I is the image to be classified.

If the difference between the image and the template is less than a threshold value then the candidate image is classified as human.

2) Parzen classifier: The second method tested is a Parzen classifier, using some simple statistical features. The features used with the Parzen classifier are the mean, standard deviation, aspect ratio, the entropy and fill ratio (the ratio of foreground pixels to the total) of the images. Fehlman and Hinders [15] use 15 features and a committee of classifiers for the classification of non-heat generating objects in thermal images. To reduce the computational requirements, a smaller number of features was chosen to test the Parzen classifier. A Parzen classifier is a statistical classifier that uses a Parzen density estimate. The Parzen density estimate estimates the conditional probability of getting a given feature vector (D) given the image is of class j (O_j) [15], ie:

$$P(D|O_j) = \frac{1}{N_j h^d} \sum_{q=1}^{N_j} H\left(\frac{D - Dqj}{h}\right) \tag{2}$$

Where:
h is the length of one side of a d dimensional hypercube
d is the dimensionality of the feature space.
D_{qj} is the q^{th} training feature of class j.
N_j is the number of feature vectors belonging to class j.

H is the Parzen window function:

$$H\left(u\right) = \begin{cases} 1 & |u_p| \leq \frac{1}{2} \; p = 1, ..., d \\ 0 & otherwise \end{cases} \tag{3}$$

Where:
$|u_p|$ is the magnitude of the p^{th} component of u.

The Parzen classifier uses Bayes' theorem and the Parzen density estimation in Equation 2 to determine the probability that the image belongs to a certain class given the observed feature vector. The posterior probability given by the Parzen classifier is

$$P(O_j|D) = \frac{P(D|O_j)P(O_j)}{P(D)} \tag{4}$$

$$= \left[\frac{1}{N_j h^d} \sum_{q=1}^{N_j} H\left(\frac{D - Dqj}{h}\right) \right] \frac{P(O_j)}{P(D)} \tag{5}$$

Where:
$P(O_j)$ is the prior probability of getting an object of class j.
$P(D)$ is called the evidence and normalises the posterior probabilities so that they sum to one.

Normally a decision is made based purely on the posterior probability: an image is classified as human if the probability that it is human is greater than the probability that it is not. For this work an offset is added which allows the adjustment of the sensitivity and false positive rates. An offset, in the range of -1 to 1 exclusive, is added to the probability of not being human. A negative offset will increase the probability that an image is classified as a human, i.e. it will result in an increased number of true positives but also increase the number of false positives. A positive offset has the opposite effect, biasing the classifier towards returning fewer false positives.

3) Neural network classifier: The third classifier investigated is a neural network classifier. Neural networks have been used for a wide variety of computer vision applications, including: vision-based vehicle driving [16], face detection [17] and pedestrian detection [18].

The network chosen for evaluation is a single hidden layer perceptron with a sigmoidal activation function. The network has 80 input nodes, 12 hidden nodes and a single output. The network is trained three times using back propagation and the weights giving the smallest error out of the three runs are saved.

The input images from the segmentation algorithm are resampled to produce 20×48 pixel images. The high dimensionality of the input is reduced using a principal component analysis. Using the magnitude of the eigenvalues, it can be shown that the first 80 components capture the majority of the significant information about the images. For classification, the rescaled input image is projected onto the lower dimensional space using the 80 chosen components. The 80 resulting features are then classified by a network with 80 input nodes. Initial tests showed that a network with 12 hidden nodes gave good results.

C. Distance Sensors

In order to predict the trajectory of the people identified by the classification step correctly, the distance from the camera to the people needs to be determined. There are a number of ways of determining the distance to objects of interest. Some of the common ways of determining distances are: structure from motion, depth from focus or defocus, stereo vision, scene geometry and fusion of the thermal camera with a separate 3D camera. It was decided that a 3D camera is necessary in addition to the thermal camera owing to limitations of using a single camera for depth estimation. Monocular depth estimation methods such as depth from focus require a number of images to determine distance and are too slow for collision avoidance. The high cost of thermal cameras does not make stereo IR a viable option so fusion of the thermal and distance images is required

There are a number of possible depth sensors that could be used, such as TOF cameras, laser scanners or structured light cameras. For this work a TOF camera and structured light camera have been used.

TOF cameras measure the phase shift of light returning from a scene to calculate the distance to each point. Unlike a laser scanner which scans a single beam across a scene a TOF camera has an array of receiving elements and measures the distance to all points simultaneously. Commercial TOF cameras use a modulated near-infrared light source and measure the phase shift between the transmitted and received light [19]. The maximum unambiguous distance (D_{unamb}) to a target would be:

$$D_{unamb} = c/2f \tag{6}$$

$$D_{unamb} = \lambda/2 \tag{7}$$

Where:
f is the modulation frequency.
λ is the modulation wavelength.
c is the speed of light.

Any distance less than D_{unamb} is calculated by measuring the ratio of the phase shift (ϕ) to a full cycle and multiplying it by the maximum distance.

$$d = (\phi/2\pi)D_{unamb} \tag{8}$$

$$d = (\lambda/4\pi)\phi \tag{9}$$

One of the problems with TOF cameras is caused by the phase shift ambiguity. A phase shift of slightly over 2π would be measured as a shift of just greater than zero and according to Equation 9 the calculated distance would be close to zero.

Structured light sensors project a known pattern onto a surface and record the pattern using a camera a certain distance from the projector. The projected pattern can be a series of lines, a grid of lines or matrix or dots. Fig. 1 shows the principle used to calculate the distance by triangulation. It can be shown using similarity of triangles that the x and z coordinates of the target are:

$$x = bu/(fcot\alpha - u) \tag{10}$$

and

$$z = bf/(fcot\alpha - u) \tag{11}$$

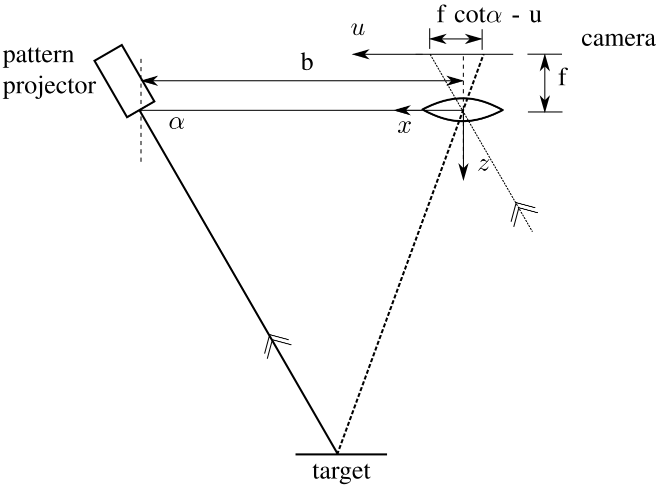


Fig. 1. Schematic showing the principle of structured light triangulation (Adapted from Siegwart and Nourbakhsh [20])

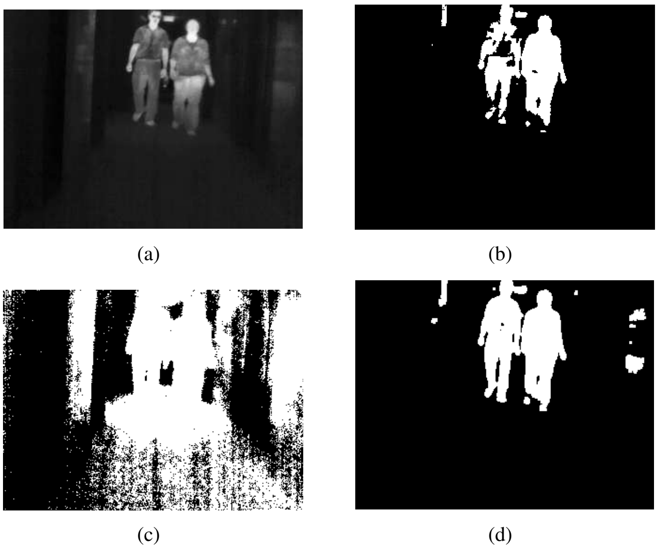

(a)
(b)
(c)
(d)

Fig. 2. Results of segmentation tests: (a) is the input image; (b) shows the result of edge and intensity segmentation; (c) is the result using Otsu's method; and, (d) is the result using temperature threshold-based segmentation

III. Results

This section describes the results of subsystem testing using preliminary indoor data. A dataset was taken in a corridor environment using a FLIR A300 thermal camera. The thermal images from the A300 camera were segmented to extract ROIs that could possibly be humans. The ROIs were classified by hand to provide ground truth data. The regions were classified as containing a single standing person, multiple overlapping people, a partial image of a person or as not containing a person. The classification resulted in a training set containing sub-images of 332 people, 55 groups of people, 126 sub-images of partially occluded people and 1287 sub-images not containing any people. This ground truth data was used for the training and verification of the classification algorithms.

The SR4000 TOF camera and a Microsoft Xbox Kinect structured light 3D sensor have been tested in a working mine and the results are discussed.

A. Segmentation

Fig. 2 shows an image from the A300. Ideally the ROIs should only be the two people in the image. It is shown in Fig. 2 that a simple temperature threshold ROI extraction performs better than two more complex algorithms.

The first ROI extraction algorithm uses a combination of intensity and edge information. The algorithm extracted regions with a certain intensity surrounded by strong edges. The addition of edge information reduced the number of noise regions, however it was found that objects in the thermal images are invariably surrounded by edges that are incomplete. A robust integration was used that could highlight regions surrounded by incomplete edges but it is computationally intensive and does not improve the segmentation results significantly. As people get closer to the camera additional edges are detected across their bodies due to, for example, clothing. This causes the addition of edge information to degrade the segmentation performance at shorter ranges.

A histogram-based segmentation algorithm, using Otsu's threshold selection method [21], was also tested for segmentation. Otsu's method is commonly used for grayscale image thresholding. Otsu's method assumes a bimodal distribution of intensities and attempts to optimally divide the distribution into two. Otsu's threshold selection does not work on the thermal images. This is because the temperature distribution is unimodal due to the uniformity of the background temperature.

It was found that a simple temperature threshold-based segmentation performed better than the two above mentioned thresholding algorithms. The temperature threshold extracts regions that have a temperature of between 26.8 $°C$ and 37 $°C$ and then performs a morphological opening, on the binary image created, to remove small noise regions. The ROIs extracted using the temperature threshold are shown in Fig. 2.

B. Classification

For testing the classifiers only a binary classification was considered, whether the region contains a single person or not. The 1800 manually classified regions are randomly divided into training and evaluation data sets, each of approximately the same size (a random division with equal chance of being in each set). Each classifier is trained and then run three times, the first time it is run using the data from the evaluation data. The two subsequent tests are run using a new randomly chosen dataset. Each classifier is evaluated in terms of classification accuracy and speed.

The classification rates are for the classifiers run in MATLAB R2010b on a 2.8 GHz Pentium 4 PC. The number of classifications per second for each classifier is averaged over the three tests and the results are shown in Table I.

TABLE I
COMPARISON OF CLASSIFIER SPEEDS. (RUNNING IN MATLAB R2010B)

Classifier	Speed (classifications/s)
Template	4830
Parzen	552
Neural Network	1227

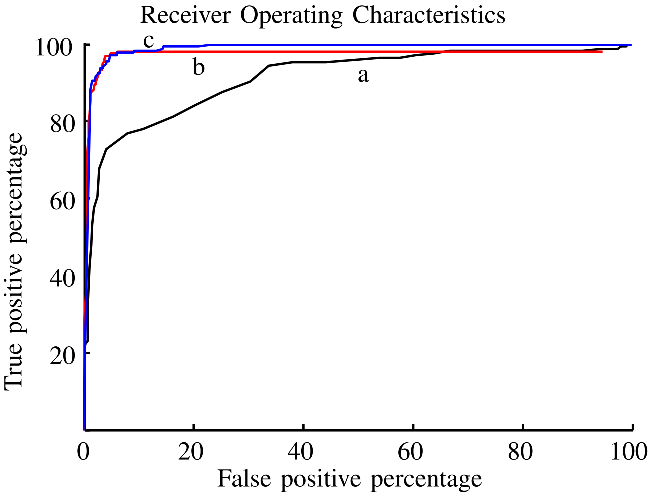


Fig. 3. The Receiver Operating Characteristics for a) the Template classifier, b) Parzen classifier and c) Neural Network

Fig. 3 shows typical Receiver Operating Characteristic (ROC) curves for each of the classifiers.

The performance of the template classifier is significantly poorer than the other two and does not warrant further consideration despite its speed.

The neural network achieves very similar classification performance to the Parzen classifier. The main difference between the two is that the Parzen classifier achieves a maximum true positive rate of 98% while the neural network can detect 100% of the targets (albeit with a high false positive rate). The reason the Parzen does not reach 100% true positive is the finite extent of the Parzen window. So if all the features fall just outside the window, the classifier will return a zero probability of being human.

The classifier is required to detect people without missing any, ie. the true positive rate needs to be as close to 100% as possible. The effect of false positives is less severe, simply adding to the number of objects that need to be tracked.

The neural network classifier achieves slightly better detection performance and significantly faster classification than the Parzen and is therefore the classifier chosen for development as part of the human detection system.

C. Distance Sensors

Testing of the two 3D sensors underground has shown a significant disadvantage of using TOF camera technology in a harsh underground environment. The drilling of blast holes in a mine gives off a fine water spray; coupled with high humidity this causes a mist in active areas of the mine. The TOF camera's amplitude image, in Fig. 4, shows the water mist near the base of the support in the centre of the image. The distance image, shown in Fig. 5, shows a significant jump in measured distances near the base of the support due to the mist there.

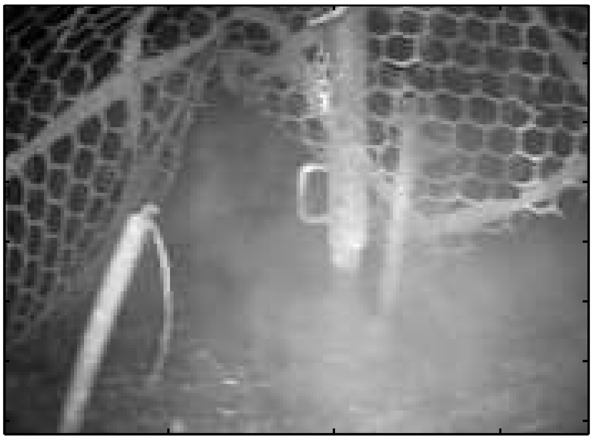


Fig. 4. Time of flight camera amplitude image through mist

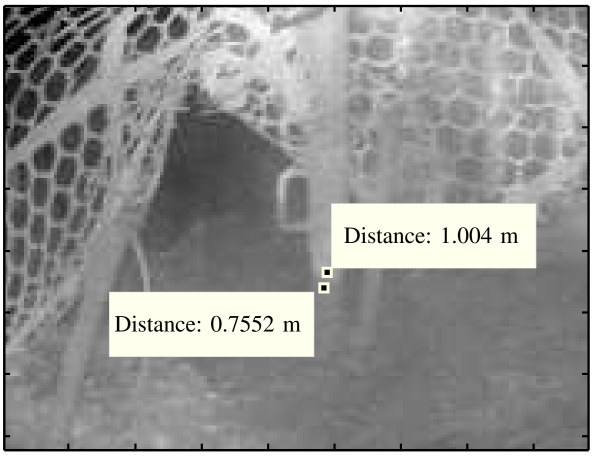


Fig. 5. Time of flight camera distance image through mist

The reason for the poor performance of the TOF camera is that the camera is receiving a reflection off the object of interest as well as multiple reflections off the intervening water droplets. The reflection off the mist causes the received phase shift to be less than the true value and therefore the measured distance is shortened. It is expected that dust, which will be more of a problem in the tunnels where the pedestrian detection system will operate, will have a similar effect to the mist.

The TOF camera was also found to suffer from significant motion blurring due to the fact that a single range image is measured using four phase measurements. Decreasing the integration time will reduce the blurring but will decrease the accuracy and range of the camera.

The structured light Kinect sensor seems unaffected by the mist but without a known ground truth distance the effect of the mist on the accuracy of the Kinect is not known.

IV. Conclusion

The current state of the development of a pedestrian detection system for underground mine vehicles is described in this paper. Some current pedestrian detection systems are listed and their limitations described. The system architecture and major subsystems are outlined. It is shown that as a result of the thermometric nature of the IR images, a temperature range based segmentation is superior to other more complex segmentation methods. It is shown that a neural network classifier outperforms a template classifier and a Parzen classifier. An evaluation of two distance sensors shows that a TOF cameras suffer from motion blurring and inaccuracies due to obscuring mist. Future work involves the acquisition of a large underground dataset from a moving platform to test the velocity estimation methods. Further work is also required to verify whether the effect of dust on the TOF camera is similar to the effect of the mist. Work is also required to determine the quantitative effect of dust on the accuracy of the time of flight and structured light 3D sensors.

References

[1] M. Creamer, "South african mining industrys 2013 zero-harm attainment unlikely - social scientist," Mining Weekly.com, December 2010. [Online]. Available: http://goo.gl/9FfO0

[2] Mine Site Technologies, 2006. [Online]. Available: http://www.minesite.com.au/applications/proximity-detection/

[3] National Institute of Occupational Safety and Health, "Proximity detection," August 2010. [Online]. Available: http://www.cdc.gov/niosh/mining/topics/topicpage58.htm

[4] P. Laliberté, "Summary study of underground communications technologies," CANMET Mining and Mineral Sciences Laboratories, Tech. Rep., May 2009.

[5] T. M. Ruff, "Advances in proximity detection technologies for surface mining equipment," in *Proc. of 34th AIMHSR*, Salt Lake City, 2004.

[6] FLIR Commercial Vision Systems, "Avoiding accidents with mining vehicles," Application story, 2008.

[7] R. Paschotta, *Encyclopedia of Laser Physics and Technology*, 1st ed. Wiley-VCH, 2008. [Online]. Available: http://www.rp-photonics.com/infrared_light.html

[8] OptoIQ, "IR imagin: Short-wave IR offers unique remote sensing solutions," Apr 2006. [Online]. Available: http://goo.gl/X8uCH

[9] J. R. F. Handley, *Historic Overview of the Witwatersrand Goldfields*. Howick: Handley, 2004, ch. 4.

[10] F. H. Von Glehn and S. J. Bluhm, "The flow of heat from rock in an advancing stope," in *Proceedings of the International Conference on Gold*, vol. 1. Johannesburg: SAIMM, 1986.

[11] S. Bluhm and M. Biffi, "Variations in ultra-deep, narrow reef stoping configurations and the effects on cooling and ventilation," *Journal of The South African Institute of Mining and Metallurgy*, vol. 101, no. 3, pp. 127–134, May 2001.

[12] W. M. Marx and R. M. Franz, "Determine appropriate criteria for acceptable environmental conditions," CSIR: Division of Mining Technology, DeepMine Research Task 6.1.1, June 1999.

[13] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, June 2002, pp. 15 – 20.

[14] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, "Pedestrian detection in infrared images," in *Intelligent Vehicles Symposium, 2003. IEEE*, 2003, pp. 662 – 667.

[15] W. L. Fehlman and M. K. Hinders, *Mobile Robot Navigation with Intelligent Infrared Image Interpretation*, 1st ed. London: Springer, 2009.

[16] D. A. Pomerleau, *Robot Learning*, 2nd ed. Kluwer Academic Publishers, 1997, ch. 2, pp. 19–44.

[17] V. Turchenko, I. Paliy, V. Demchuk, R. Smal, and L. Legostaev, "Coarse-grain parallelization of neural network-based face detection method," in *4th IEEE Workshop on Intelligent Data Acquisition and Advanced Computing Systems*, September 2007, pp. 155 –158.

[18] L. Zhao and C. E. Thorpe, "Stereo- and neural network-based pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 3, pp. 148–154, 2000.

[19] P. Sphikas, *SR4000 User Manual*, 1st ed., MESA Imaging AG, Technoparkstrasse 1 8005 Zurich, March 2010.

[20] R. Siegwart and I. R. Nourbakhsh, *Introduction to Autonomous Mobile Robots*, 1st ed. Cambridge, Massachusetts: The MIT Press, 2004, ch. 4, pp. 122–128.

[21] N. Otsu, "A threshold selection method from grey-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.

Segmentation techniques for extracting humans from thermal images

J. S. Dickens
CSIR Centre for Mining Innovation
PO Box 91230
Auckland Park 2006
Johannesburg, South Africa
Email: jdickens@csir.co.za

J. J. Green
CSIR Centre for Mining Innovation
PO Box 91230
Auckland Park 2006
Johannesburg, South Africa
Email: jgreen@csir.co.za

***Abstract*—A pedestrian detection system for underground mine vehicles is being developed that requires the segmentation of people from thermal images in underground mine tunnels. A number of thresholding techniques are outlined and their performance on a number of thermal images is investigated. The thresholding techniques are evaluated on images in various ambient conditions and it is shown that a minimum error thresholding technique is the most effective.**

I. Introduction

A pedestrian detection system for underground mine vehicles is being developed to address the high number of fatalities caused by mine vehicles [1]. The system makes use of a combination of thermal and 3D imaging to identify and track people near the mine vehicle. The system will help improve drivers' awareness of people near their vehicles and also allow for the safe operation of autonomous mine vehicles in the presence of humans.

The system detects, classifies and tracks humans in the thermal images and then combines the thermal images with 3D images to provide actual position information. It will need to determine how far away from the vehicle the people are and track them to determine whether they are on a collision course with the vehicle. A commonly used paradigm for object detection and tracking in video is to first extract regions of interest and then classify or validate them [2–7], which is the methods that is being used for the pedestrian detection system, as shown in the system diagram, Fig. 1. This paper deals with the segmentation subsystem, the regions that have been segmented by this system will be further processed to remove small noise regions and then the remaining regions will be classified. Various methods for segmenting people from thermal images will be reviewed and compared.

Image thresholding takes in a multi-valued input image and outputs a binary image where one of the states represents foreground objects and the other represents the background. Image thresholding is used for a wide variety of applications from extracting printed characters for optical character recognition through identification of defects in automated inspection tasks, to segmenting computed tomography x-ray images.

At first glance segmenting humans from thermal images may seem trivial because we know that human core body temperature remains in a very narrow range. However human surface temperatures vary significantly depending on a number of factors such as the clothes worn and the naturally lower temperature of the extremities (arms and legs).

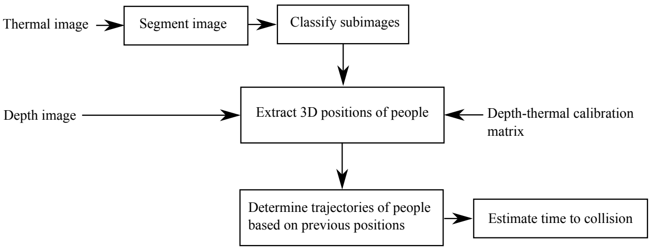

Thermal image
Segment image
Classify subimages
Depth image
Extract 3D positions of people
Depth-thermal calibration matrix
Determine trajectories of people based on previous positions
Estimate time to collision

Fig. 1. A block diagram showing the subsystems making up the pedestrian detection system.

It is assumed that the temperature of people within a mine tunnel will always exceed the temperature of the tunnel itself. In deep South African gold mines the virgin rock temperatures can be as high as 60 $^{\circ}C$ however ventilation and other cooling brings the temperature within working areas (stopes) down to below 30 $^{\circ}C$ to allow work to be done [8]. Work conducted to model the heat flow from advancing stopes shows that the rock surface temperature can be assumed to be equal to the ventilation air wet-bulb temperature (T_{wb}) [9]. The wet-bulb temperature takes into account the relative humidity of the air and therefore the effects of evaporative cooling. Since the wet-bulb temperature takes into account the effect of evaporative cooling there will always be a positive temperature gradient between people and the environment to allow the dissipation of metabolic heat.

Since the people in the thermal images will always be warmer than the background the segmentation of the thermal images involves determining an optimal threshold to extract only the people as foreground objects. The camera used to capture the images used for evaluating the thresholding methods is a FLIR A300 providing a thermometric image.

II. Thresholding methods

There are a very large number of thresholding algorithms belonging to a number of categories, a good survey of a large number of them is provided by Sezgin and Sankur [10]. The methods evaluated here will be those identified as the best performing by Sezgin and Sankur as well as a number of other techniques chosen for certain characteristics. Thresholding methods falling into the following categories; clustering-based thresholding, entropy-based thresholding, locally adaptive thresholding and model-based thresholding will now be discussed.

For all of the following discussions the following notation will be used. Each picture has a total of N pixels that fall into L grey-levels. The number of pixels that fall into each grey-level (i) of the image histogram is denoted by n_i. The normalised grey-scale histogram can be considered an estimate of the probability distribution of pixel intensities i.e.

$$p_i = n_i/N \tag{1}$$

Where:
p_i is the probability that a pixel belongs to the i^{th} grey level
N is the total number of pixels

The cumulative probability function for the k^{th} grey-level is defined as

$$P(k) = \sum_{i=1}^{k} p_i \tag{2}$$

A. *Clustering-Based Thresholding*

1) Otsu's Method: The first thresholding method that is evaluated is Otsu's threshold selection method [11]. Otsu's method is evaluated due to its popularity as a thresholding method, being one of the most cited thresholding methods [10]. Otsu's method finds a threshold that minimises the within-class variances of the foreground and background classes. Minimising the within class variance is equivalent to maximising the between class variance.

The zeroth- and first-order cumulative moments of the image histogram up to the k^{th} grey-level are:

$$\omega(k) = \sum_{i=1}^{k} p_i \tag{3}$$

and

$$\mu(k) = \sum_{i=1}^{k} ip_i \tag{4}$$

The total mean level of the original picture is:

$$\mu_T = \mu(L) = \sum_{i=1}^{L} ip_i \tag{5}$$

It can be shown that the between class variance, σ_b^2, is:

$$\sigma_b^2 = \frac{(\mu_T\omega(k) - \mu(k))^2}{\omega(k)\,(1-\omega(k))} \tag{6}$$

Otsu's method selects the optimal threshold T_{opt} in order to maximise the between class variance. The optimal threshold is the value of k that maximises Equation 6, ie.

$$T_{opt} = \underset{k}{argmax}\ \sigma_b^2(k) \tag{7}$$

2) Iterative Clustering: Iterative clustering assumes that the intensity histogram has two peaks, one for the foreground objects and another for the background objects. The algorithm starts with the threshold set to the centre intensity level, the peak of the histogram on either side of the threshold is then determined. The threshold value is moved to the midpoint of the two peaks and the peaks are found again. The process is repeated until the change in the threshold is sufficiently small.

3) Minimum Error Thresholding: Minimum error thresholding assumes that the image is made up of foreground and background objects with normally distributed intensities. The method of minimum error thresholding is that of Kittler and Illingworth [12], their method minimises a criterion function which gives the approximate minimum error threshold. The criterion function derived by Kittler is

$$\begin{aligned} J(k) \quad = \quad & 1 + 2\left[P(k)ln\left(\sigma_1(k)\right) + (1-P(k))\,ln\left(\sigma_2(k)\right)\right] \\ & -2\left[P(k)ln\left(P(k)\right) + (1-P(k))\,ln\left(1-P(k)\right)\right] \end{aligned} \tag{8}$$

Where:
$\sigma_1(k)$ is the standard deviation of the background up to grey level k
$\sigma_2(k)$ is the standard deviation of the foreground, from k to L

The criterion function shown in Equation 8 gives a measure of the overlap of the two distributions, so the method estimates the parameters of the two normal distributions on either side of the threshold and then calculates the overlap of the two estimated distributions. Using Equation 8 the optimal threshold is easily determined.

$$T_{opt} = \underset{k}{argmin}\ J(k) \tag{9}$$

Since the distributions overlap, the estimation of the parameters will contain a bias, however this is assumed to be small. The bias does indeed appear to have little effect of the result. Another advantage of the minimum error thresholding technique is that the criterion function will not have a minimum for a unimodal distribution, so an image that does not contain any people can be detected and not segmented.

B. *Entropy-Based Thresholding*

Entropic thresholding methods exploit the entropy distribution of the grey-levels in the scene. Maximising the entropy of the thresholded image maximises the information between the foreground and background distributions in the image [10, 13]. For a threshold at grey-level k the entropy of the background up to grey-level k is

$$H_b = -\sum_{i=1}^{k} \frac{p_i}{P(k)} ln \frac{p_i}{P(k)} \tag{10}$$

and the entropy of the foreground is

$$H_f = -\sum_{i=k+1}^{L} \frac{p_i}{(1-P(k))} ln \frac{p_i}{(1-P(k))} \tag{11}$$

Defining the sum of the two entropies as $\phi(k)$ we get

$$\phi(k) = -\sum_{i=1}^{k} \frac{p_i}{P(k)} ln \frac{p_i}{P(k)} - \sum_{i=k+1}^{L} \frac{p_i}{(1-P(k))} ln \frac{p_i}{(1-P(k))} \tag{12}$$

Maximising $\phi(k)$ gives the maximum information between the two distributions. So the optimal threshold is

$$T_{opt} = \underset{k}{argmax}\ \phi(k) \tag{13}$$

C. Locally Adaptive Thresholding

Locally adaptive thresholding adapts the threshold for each pixel in the image, instead of having one threshold (T) the threshold is an matrix the same size as the image ($T(x, y)$). The adaptive thresholding method is that of Sauvola and Pietikäinen [14] which is adapted based on the mean and standard deviation of the pixels in a window around each pixel. The threshold is calculated according to the formula

$$T(x, y) = m(x, y) \cdot \left[1 + k\left(\frac{s(x, y)}{R} - 1\right)\right] \tag{14}$$

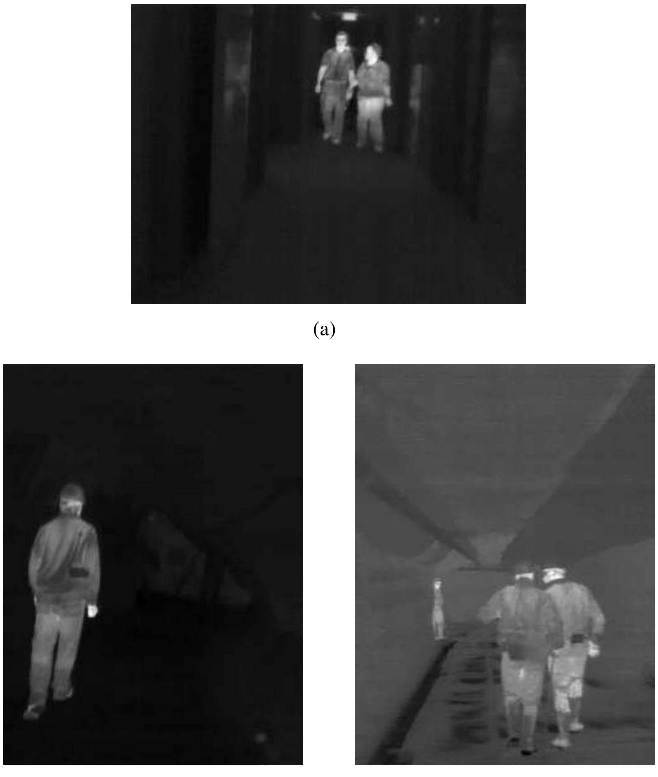
Where:
$m(x, y)$ is the mean of the window centred on pixel xy
$s(x, y)$ is the standard deviation of the window centred on pixel xy
R is the range of the standard deviation
k is a user defined constant

In our experiments the value of k was chosen to be $k = -0.02$ and the window for calculating the mean and standard deviation is 15×15 pixels. The value of k is negative because we are attempting to extract higher intensity (warmer) objects from a darker background while Sauvola was attempting to extract dark text from a light background.

III. Thresholding results

The methods were evaluated on thermal images containing people in a variety of conditions. The background temperature of the images varies from about 11 $°C$ to 25 $°C$. Due to the difficulty in establishing ground truth for testing the thresholding, the methods are tested qualitatively. Qualitative testing is sufficient due to the fact that the results are very sensitive to the threshold chosen so mostly the results are binary, the method provides an acceptable threshold or not. The test images used for the testing of the thresholding methods are shown in Fig. 2 below.

The images in Fig. 2 represent typical images from three datasets. The corridor provided a good dataset to test the classification algorithm because of the presence of warm objects that were not people (the lights and reflections off doors). The mine in b provides one end of the spectrum, it is a shallow mine with a cold air temperature. The tunnel in c shows an example of a problem case, the air temperature was fairly high but there was a very high ventilation air velocity, this high velocity air reduces the temperature difference between the people and surroundings. The area in image c was part of the training area of the mine and does not represent the typical conditions that would be present in the mine.


(a)

(b)

(c)

Fig. 2. Test images for the thresholding algorithms: (a) a corridor at 25 $°C$; (b) a mine tunnel at 11 $°C$ and (c) a tunnel at 21 $°C$.

A. Clustering-Based Thresholding

All the clustering based thresholding methods suffer from a similar problem, they assume that the foreground and background objects have intensity distributions that are well separated which is not the case in the thermal images in this work.

1) Otsu's Method: Otsu's method produces acceptable results for images where the number of foreground and background pixels are approximately equal [10]. This is not the case in the thermal images investigated where the number of background pixels is significantly larger than the number of foreground pixels. When there are a significantly larger number of pixels in one class than the other, then Otsu's method tends to split the larger mode in half [12], which is exactly what is seen in Fig. 3, the background has been split by a threshold dividing the background mode of the histogram.

2) Iterative Clustering: The results of the iterative clustering method test are shown in Fig. 4, the results for image a are acceptable and the results on image b are good but the result on image c is unacceptable. The reasons for the difference in the performance between the different images can be seen by looking at the image histograms shown in Fig. 5 and Fig. 6.

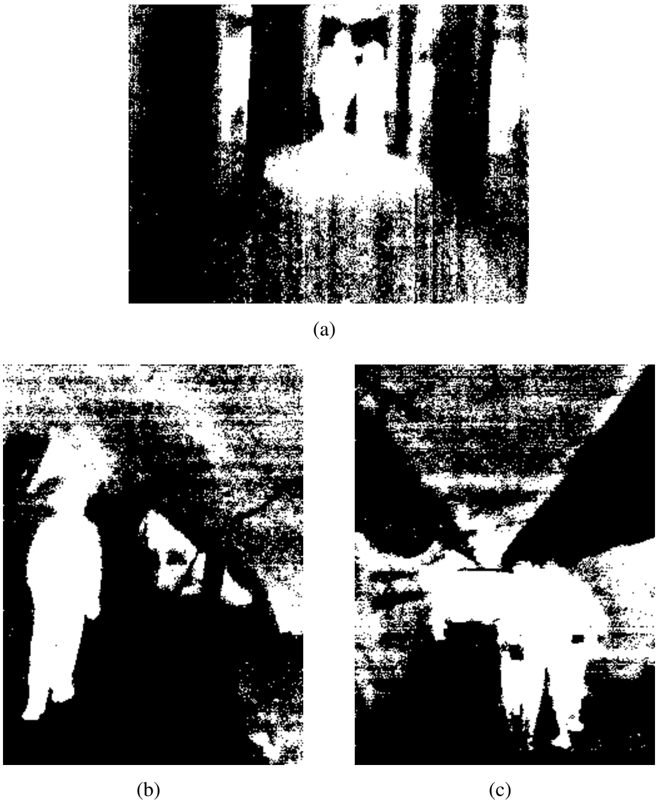


Fig. 3. Thresholding results using Otsu's method

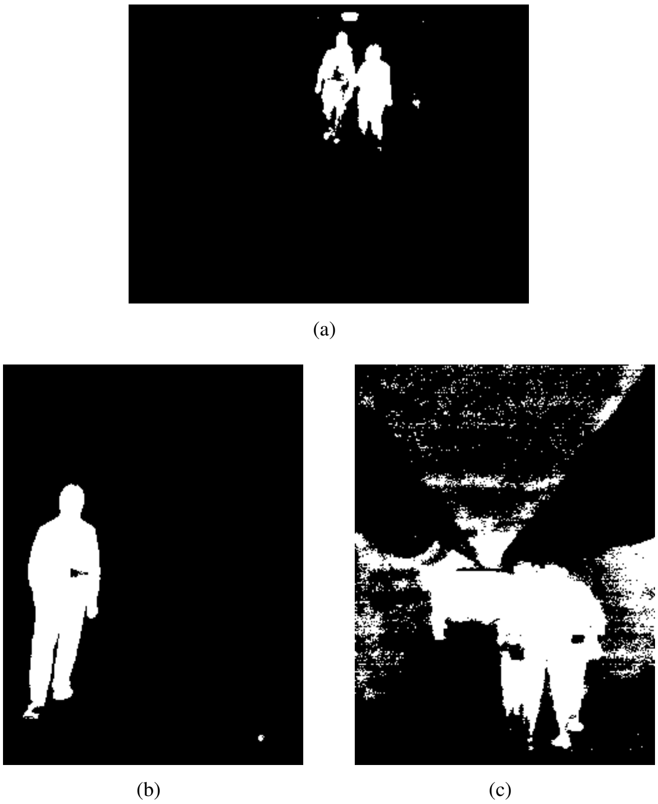


Fig. 4. Thresholding results using iterative clustering method

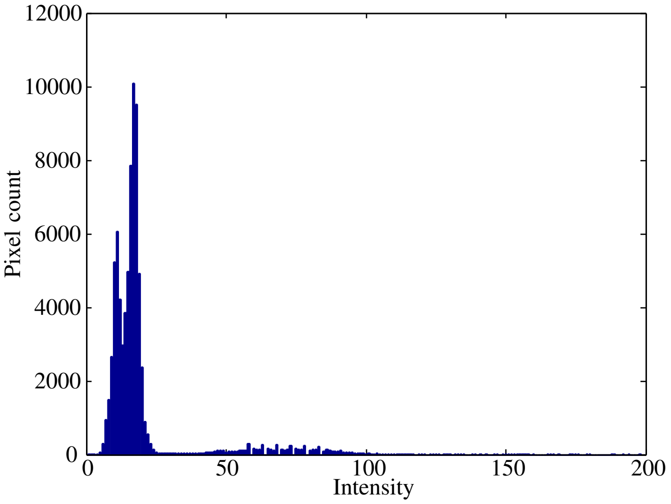


Fig. 5. Histogram for image in the cold mine tunnel (image *b*)

It is evident in Fig. 5 that the histogram consists of two distributions that are fairly well separated, while in Fig. 6 the distribution appears to simply taper off to the right of the main peak. Without a well separated second peak, this method will obviously not work.

3) Minimum Error Thresholding: The results of the minimum error thresholding algorithm, shown in Fig. 7, indicate that the minimum error thresholding technique performs well on all of the input images. The result on image c shows incomplete segmentation of the two people close to the camera. While unfortunate it is not possible for a single threshold method to perform better since parts of the people (their hard-hats, gum-boots and cap-lamp batteries) are at the same temperature as the background.

B. Entropy-Based Thresholding

The entropy based threshold performs well on all of the images with only a small amount of noise, see Fig. 8. This makes sense since the entropy-based method is segmenting the images without making any assumptions about the underlying distributions of the foreground and background objects.

C. Locally Adaptive Thresholding

The locally adaptive thresholding method produces some interesting results. The method extracts part of the people and a fair amount of noise from the background. The reason for the poor performance of the adaptive thresholding method is that unlike text which the algorithm was originally intended for, the foreground objects in the thermal images are large in extent. In Sauvola's work each character being thresholded is smaller than the window used to calculate the mean. In the images used for these experiments the people (foreground objects) are larger than the window so the mean value is increased near the center of the object where the window encloses the whole object. The increasing mean towards the center

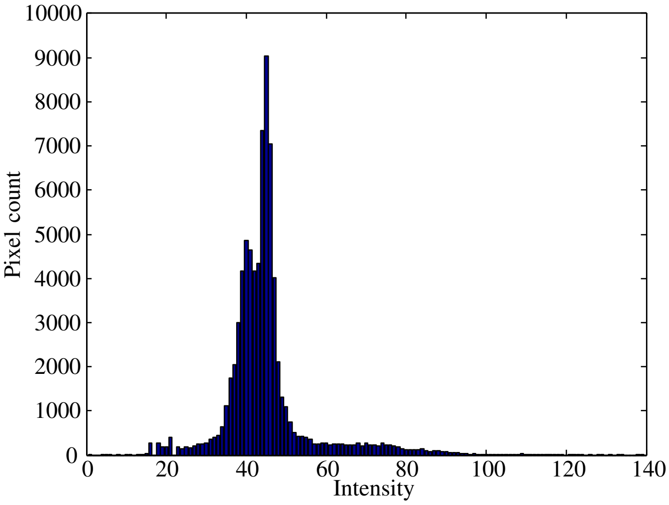


Fig. 6. Histogram for image in the warm mine tunnel (image c)

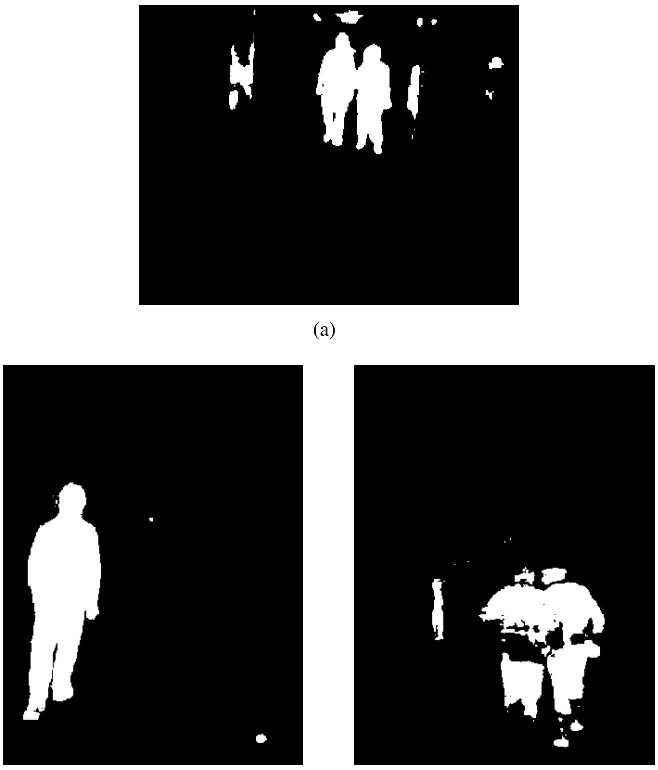


Fig. 7. Thresholding results using minimum error clustering method

of the foreground objects causes a commensurate increase in the threshold, which explains the tendency of the adaptive thresholding to extract only the edges of people.

D. Background-only Images

The two best performing segmentation methods have an additional advantage over the other methods presented, they provide a measure of the certainty of the threshold. This measure can be used to prevent the segmentation of an image

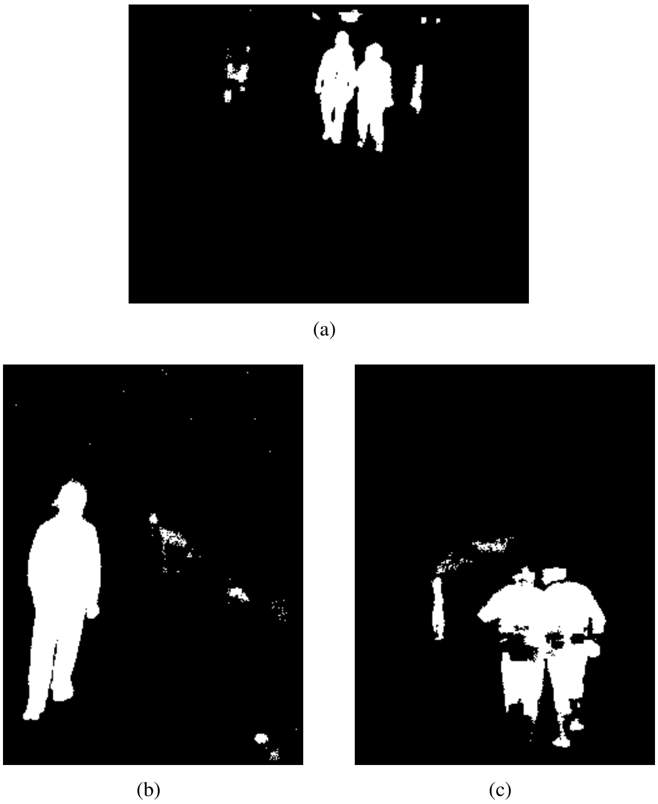


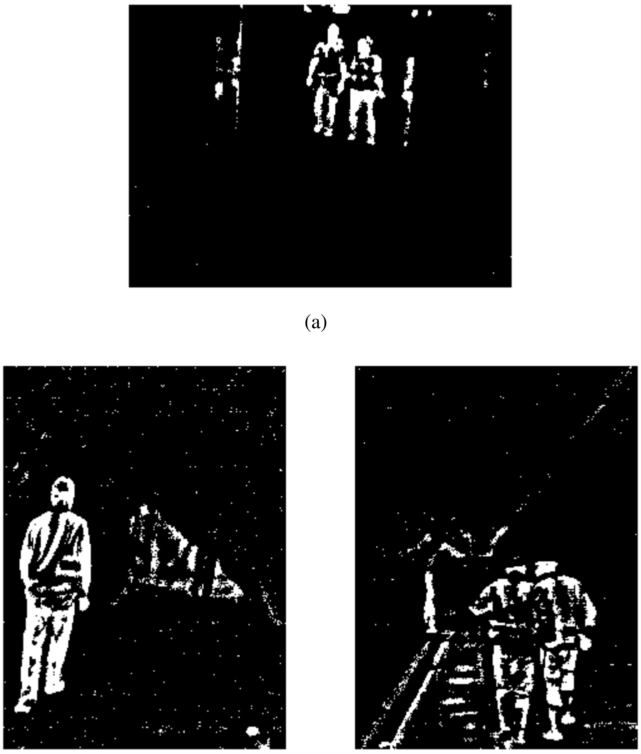
Fig. 8. Thresholding results using Entropy-based thresholding method

that does not contain any people. The entropic thresholding gives a measure of the information retained for each threshold value. For a uniform distribution of pixels the information content will remain unchanged for any threshold, while for an image containing two very different distributions the difference in the information content between the optimal threshold and the others will be significant. Fig. 10 shows a plot of the entropy versus the threshold, notice the difference in scale between the background-only image and the image containing people.

The criterion function for minimum error thresholding shows a similar effect where the scale can be used to determine whether the image contains only background. Minimum error thresholding also estimates the mean and standard deviation of foreground and background at each threshold value. A combination of the range of the criterion function and the difference between the means is currently being used to prevent the segmentation of the background.

IV. Conclusion

Thresholding techniques that have been shown to perform well on text and non-destructive testing (including thermal images) images have been evaluated for segmenting people in thermal images. Segmenting people from thermal images in mine tunnels is challenging due to a significant overlap in the distributions of the foreground and background and the relative difference in the number of foreground and background pixels.

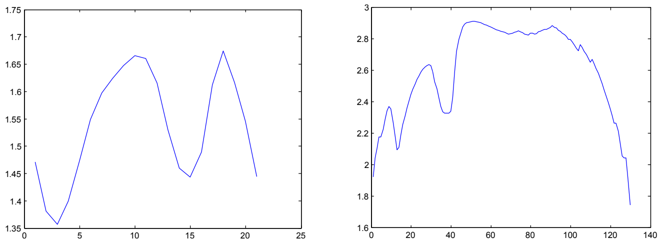


(a)

(b)

(c)

Fig. 9. Thresholding results using locally adaptive thresholding method



(a)

(b)

Fig. 10. Plots of the entropy versus threshold for: (a) an image containing only background, (b) an image containing people

The results show that the minimum error thresholding technique performs the best followed by the entropy based method. The other methods tested produced unacceptable results and the reasons for their performance is explored. Future work may involve the addition of 3D information to improve the segmentation by allowing parts of a single person, identified by the thermal segmentation, to be combined.

References

[1] J. S. Dickens, J. J. Green, and M. A. van Wyk, "Pedestrian detection for underground mine vehicles using thermal images," in *IEEE Africon 2011*. Livingstone, Zambia: IEEE, September 2011.

[2] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, "Pedestrian detection in infrared images," in *Intelligent Vehicles Symposium, 2003. IEEE*, 2003, pp. 662 – 667.

[3] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Computer Vision and Image Understanding*, vol. 106, no. 2-3, pp. 288 – 299, 2007, special issue on Advances in Vision Algorithms and Systems beyond the Visible Spectrum.

[4] Y. L. Guilloux, J. Lonnoy, R. Moreira, M.-P. Bruyas, A. Chapon, and H. Tattegrain-Veste, "Paroto project: the benefit of infrared imagery for obstacle avoidance," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, 2002, pp. 81–86.

[5] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, June 2002, pp. 15 – 20.

[6] S. M. Thornton, M. Hoffelder, and D. D. Morris, "Multisensor detection and tracking of humans for safe operations with unmanned ground vehicles," in *Workshop on Human Detection from Mobile Platforms*, Pasadena, May 2008.

[7] F. Xu, X. Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 1, pp. 63 – 71, March 2005.

[8] J. R. F. Handley, *Historic Overview of the Witwatersrand Goldfields*. Howick: Handley, 2004, ch. 4.

[9] F. H. Von Glehn and S. J. Bluhm, "The flow of heat from rock in an advancing stope," in *Proceedings of the International Conference on Gold*, vol. 1. Johannesburg: SAIMM, 1986.

[10] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic Imaging*, vol. 13, no. 1, pp. 146–168, 2004. [Online]. Available: http://dx.doi.org/doi/10.1117/1.1631315

[11] N. Otsu, "A threshold selection method from grey-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, January 1979.

[12] J. Kittler and J. Illingworth, "Minimum error thresholding," *Pattern Recognition*, vol. 19, no. 1, pp. 41 – 47, 1986. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0031320386900300

[13] J. Kapur, P. Sahoo, and A. Wong, "A new method for gray-level picture thresholding using the entropy of the histogram," *Computer Vision, Graphics, and Image Processing*, vol. 29, no. 3, pp. 273 – 285, 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0734189X85901252

[14] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225 – 236, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320399000552

HUMAN DETECTION FOR UNDERGROUND AUTONOMOUS MINE VEHICLES USING THERMAL IMAGING

J. S. Dickens[1], J. J. Green[2] and M. A. van Wyk[3]
[(1,2)]CSIR Centre for Mining Innovation
Johannesburg, South Africa
jdickens@csir.co.za[1], jgreen@csir.co.za[2]
[3]University of the Witwatersrand, Faculty of Engineering and the Built Environment
Johannesburg, South Africa
anton.vanwyk@wits.ac.za[3]

ABSTRACT

Underground mine automation has the potential to increase safety, productivity and allow the mining of lower-grade resources. In a mining environment with both autonomous robots and humans, it is essential that the robots are able to detect and avoid people. Current pedestrian detection systems and the reasons that they are inadequate for mining robots are discussed. A system for human detection in underground mines, using a fusion of three-dimensional (3D) information with thermal imaging, is proposed. The system extracts regions of interest and classifies them as human or background. The scene excluding the pedestrians is assumed to be static and is intended to be used to determine the ego motion of the vehicle. In addition to the thermal camera, a distance sensor will provide depth information and allow the calculation of the vehicle and pedestrian velocities. Various classification methods are compared and it is shown that a neural network provides the best results in terms of speed and accuracy. The results of tests on two 3D sensors indicate that further work is required to determine the effect of the harsh environment on the accuracy of the sensors.

Keywords: underground, mining, autonomous robots, obstacle detection, human tracking, thermal imaging, classification.

1 INTRODUCTION

Transportation machinery is responsible for a large portion of mine deaths in South Africa. After rock falls, vehicles are the second leading cause of mining fatalities. A reliable system for detecting people near mining vehicles is needed to prevent collisions between vehicles and personnel. The South African mining industry has committed itself to strive for zero fatalities by 2013 [1]. Given that the number of mining fatalities in 2010 was 128 [1], achieving zero fatalities by 2013 is unlikely to be possible without a fundamental change in mining methods. Automation in mines has the ability to improve human safety [2] and potentially enable the mining of resources that cannot be mined in the traditional way [3]. An autonomous mine vehicle operates in an area with people must be able to detect humans in order to operate without posing a threat to nearby personnel. As a step towards an underground autonomous mine vehicle, a pedestrian detection system is proposed that will assist vehicle operators by predicting collisions.

It is desirable that the detection system can be used in future to provide automated mine machines with the ability to operate safely in conjunction with humans. The system should be able to detect and localise people near an underground mine vehicle, which allows the system to be used for the planning of a safe path around people in an underground mine.

There are a number of existing proximity warning systems for mining vehicles, using technologies such as ultrasonic, laser, radar, GPS, Radio Frequency Identification (RFID) tags, cameras or some combination of these. Some of the strengths and weaknesses of these warning systems are outlined below.

Radar-based proximity detection is used for surface mining equipment as an aid to drivers of dump trucks for detecting people and small vehicles behind the truck. The system is fairly effective with only occasional false alarms [4]. The close proximity of tunnel walls in an underground mine causes frequent false alarms, making the use of radar problematic underground [5].

GPS proximity detection has been proposed for surface mining operations. Each vehicle and worker broadcasts its position to nearby vehicles. A display in the vehicle shows the position of nearby people, vehicles and stationary objects and alarms if they are within a predetermined range. The reliance on GPS signals precludes its use in a GPS deprived underground environment.

RFID tags are popular for collision avoidance systems owing to their very low false alarm rates. Each miner has an RFID tag embedded in their cap-lamp. A transmitter mounted on the vehicle determines the distance to each tag. RFID systems do not provide the exact location of the personnel, merely how close they are. RFID do not provide sufficient information for an autonomous vehicle. The fact that RFID cannot provide direction information implies that it cannot be used to plan a path around a pedestrian.

A machine vision based pedestrian tracking system can address some of the shortcomings of current systems. Vision provides a way of detecting people and determining exactly where they are in relation to a vehicle. Thermal infrared (IR) imaging provides the advantages of vision based detection without the problems of sensitivity to illumination and obscuring dust. Unlike visible range imaging, the illumination for thermal images is radiated by the objects being imaged, in this case people. The long wavelength (7-14 μm) of thermal IR allows it to penetrate dust and smoke [6].

The IR spectrum can be divided into four main regions. The main regions are near-infrared, short-wavelength, mid-wavelength and long-wavelength IR. Near-infrared (0.7 to 1.4 μm) is commonly used for light-based distance sensors such as laser scanners and Time of Flight (TOF) cameras. Near-infrared illumination is also often used for night-vision surveillance since this wavelength can be detected using the same imaging sensor used for visible light. Short-wavelength IR is used for various process monitoring and inspection tasks such as hot furnace monitoring. Mid-wavelength IR can be used for gas spectroscopy. Long-wavelength IR (or thermal IR) is the region of interest for this paper and is used for thermal imaging. It can be shown using, Wien’s displacement law, that objects at room temperature, around 300 K, emit IR radiation in the long wavelength IR region (peak wavelength of 9.7 μm).

In Section 2 of this paper the basic architecture of the proposed pedestrian detection system and the major sub-systems is described. The results of tests to evaluate the segmentation and classification algorithms and the distance sensors are presented in Section 3. The results are discussed and then conclusions are drawn and recommendations presented.

2 SYSTEM ARCHITECTURE

The proposed detection system uses the fusion of thermal imaging and a three-dimensional (3D) image for pedestrian detection. The sensor head consists of a FLIR A300 thermal camera, a SwissRanger SR4000 TOF camera and an Xbox Kinect, as shown Figure 1.

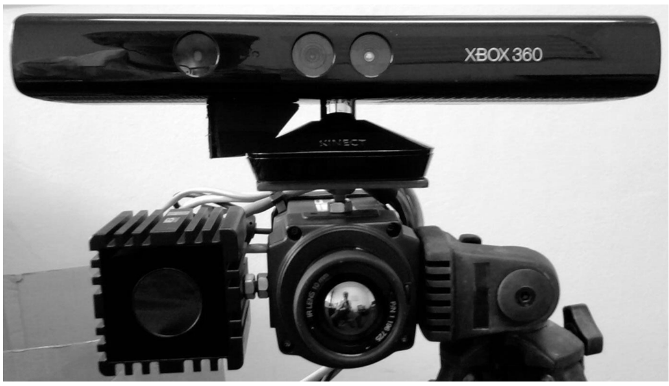


Figure 1: The sensor used for the detection system

A region that the sensor identifies as having a temperature that indicates the region could be human is defined as a Region of Interest (ROI). The detection system first extracts ROIs which are then classified as being human or background objects. The 3D points from the depth camera will be projected into the FLIR's thermal image. The humans identified in the thermal image can be extracted from the 3D image by determining which 3D points project the human regions of the thermal image.

The 3D position of the people will be used by the tracking system. The tracking system estimates the trajectory of the people in the camera's field of view. The background, excluding pedestrians, is assumed to be stationary and is used to determine the trajectory of the vehicle. The vehicle trajectory estimation will be done using the established iterative closest point surface matching algorithm. Using the trajectory of the vehicle and the pedestrians the system calculates whether a collision is likely to occur.

In order for the system to extract ROIs and classify them as human or background, thermal image segmentation and classification of the images take place. These steps are outlined and various classification methods compared below.

2.1 Thermal Image Segmentation

The system first extracts the ROIs and those confirmed as human by a classification step are tracked. The thermometric image provided by the FLIR camera allows segmentation of the image on the basis of an empirically determined temperature threshold. Tests performed show that the temperature based segmentation outperformed two more complex segmentation algorithms.

2.2 Classification

There are a number of methods for classifying humans in thermal images. To the authors' knowledge, there has not been a quantitative comparison of methods for human classification in thermal imaging. In the absence of a clear choice, it was decided to compare four different classification modalities. The classification methods compared are:

. An appearance based classifier using the difference between the candidate and a template.
a. A feature based classifier which uses a number of features extracted from the image which are classified using a Parzen classifier.
b. A neural network classifier.
c. A radial basis function support vector classifier.

A single binary classification was chosen for evaluation of the classifiers. The classifiers all indicate whether a sub-image is of a single standing pedestrian or not. The final system is intended to involve multiple classifiers to identify groups of pedestrians, occluded pedestrians and people in poses other than standing.

2.2.1 Template classifier

The first method tested was a template classifier. Template-based classification has been used for human detection in thermal images from moving vehicles. For example Nanda and Davis [7] use a probabilistic template created from training images. It was decided to create a template that represents the average appearance of a person, similar to the idea used by Nanda and Davis. The images of humans in the training data are rescaled to form an M×N pixel image. The template is the mean of the scaled images. The candidate regions are rescaled to the same dimensions as the template and the two are compared using an absolute difference distance measure, i.e.:

$$\text{Difference} = \sum_{\mathrm{i=1}}^{\mathrm{M}}\sum_{\mathrm{j=1}}^{\mathrm{N}} abs\left(T_{ij} - I_{ij}\right) \qquad (1)$$

where T is the template image and I is the image to be classified. If the difference between the image and the template is less than a threshold value then the candidate image is classified as human.

2.2.2 Parzen classifier

The second method tested was a Parzen classifier with image features. Fehlman and Hinders [8] use 15 features and a committee of classifiers for classification of non-heat generating objects in thermal images. A smaller number of features were chosen to test the Parzen classifier. The feature vectors used for classification are the mean, standard deviation, aspect ratio, the entropy and fill ratio of the images. The fill ratio is the ratio of the number of pixels extracted as foreground pixels to the total number of pixels in an enclosing rectangle. A Parzen classifier is a statistical classifier that uses Bayes' theorem and a Parzen density estimate. The Parzen density estimate, estimates the conditional probability of getting a given feature vector (D) given that the image is of class j ($\mathrm{O_j}$), i.e.:

$$\mathrm{P}\left(\mathrm{D} \mid \mathrm{O_j}\right) = \frac{1}{N_j h^d}\sum_{q=1}^{N_j} H\left(\frac{D - D_{qj}}{h}\right) \qquad (2)$$

where D_{qj} is the q^{th} training feature of class j, N_j is the number of feature vectors belonging to class j, h is the length of the sides of a hypercube with the dimensionality of the feature space (d) and H is the Parzen window function i.e.:

$$\mathrm{H}(\mathrm{u}) = \begin{cases} 1 & |u_p| \le 1/2 \quad p = 1 \ldots d \\ 0 & otherwise \end{cases} \tag{3}$$

The Parzen classifier uses Bayes' theorem and the Parzen density estimate, in Equation 2, to determine the posterior probability that the image belongs to a certain class given the observed feature vector i.e. $P(O_j|D)$.

$$\mathrm{P}(\mathrm{O}_\mathrm{j} \mid D) = \frac{P(D \mid O_j)P(O_j)}{P(D)} \tag{4}$$

$$\mathrm{P}(\mathrm{O}_\mathrm{j} \mid D) = \left[\frac{1}{N_j h^d} \sum_{q=1}^{N_j} H\left(\frac{D - D_{qj}}{h} \right) \right] \frac{P(O_j)}{P(D)} \tag{5}$$

$P(O_j)$ is the prior probability of getting an object of class j, which can be estimated from the frequency with which class j is observed. P(D) is called the evidence and normalises the posterior probabilities so they sum to one.

The image is classified as human if the probability that it is human is greater than the probability that it is not plus some offset. The offset allows the adjustment of the sensitivity and false positive rates.

2.2.3 Neural network classifier

The third classifier investigated was a neural network classifier. Neural networks have been used for a wide variety of computer vision applications including: vision based vehicle driving, handwritten digit recognition, face detection and pedestrian detection.

The network chosen for evaluation is a single hidden layer network with a sigmoidal activation function. The input images from the segmentation algorithm are re-sampled to produce 20×48 pixel images. The high dimensionality of the input is reduced using a principal component analysis. Using the magnitude of the eigenvalues, it can be shown that the first 80 components capture the majority of the significant information about the images. For classification the input image is scaled to 20×48 pixels and then projected onto the lower dimensional space using the 80 chosen components. The 80 resulting features are then classified by a neural network with 80 input nodes. Initial tests showed that a network with 12 hidden nodes gave good results. The neural network is trained three times using back propagation and the weights that give the smallest error are saved.

2.2.4 Support vector classifier

Support vector classification is a popular method for pedestrian detection. A support vector classifier was tested for classifying the test images. A support vector classifier finds a hyperplane in feature space that separates the two classes of objects with the maximum margin. The MATLAB SVM toolbox was used for the implementation of the support vector classifier [9].

A number of kernels were tested and it was found that the Radial Basis Function (RBF) kernel performed the best. As with the neural network the input images are scaled and then a principal component analysis is performed to produce 80 features that are used for classification. A soft margin (C value of 10) was used that allows the classifier to accept a small number of training errors. Allowing a small number of errors enables the classifier to generalise better by not over fitting the data. The receiver operating characteristic curve for the classifier was obtained by adjusting the bias of the hyperplane and evaluating the performance for each value of the bias.

2.3 Distance Sensors

In order to predict the trajectory of the people identified by the classification step, the distance from the vehicle to the people needs to be determined. It was decided that a 3D camera is necessary in addition to the thermal camera owing to the limitations of using a single camera for depth estimation. Monocular depth estimation methods such as depth from focus require a number of images to determine distance and are too slow for collision avoidance. The high cost of thermal cameras does not make stereo IR a viable option so a fusion of the thermal and distance images is required

There are a number of possible depth sensors that could be used, such as TOF cameras, laser scanners and structured light cameras.

Structured light sensors project a known pattern onto a surface and record the pattern using a camera a certain distance from the projector. The projected pattern can be a series of lines, a grid of lines or matrix or dots. Figure 2 shows the principle used to calculate the distance by triangulation. It can be shown using similarity of triangles that the x and z coordinates of the target are:

$$x = \frac{bu}{f\cot\theta - u} \tag{6}$$

$$z = \frac{bf}{f\cot\theta - u} \tag{7}$$

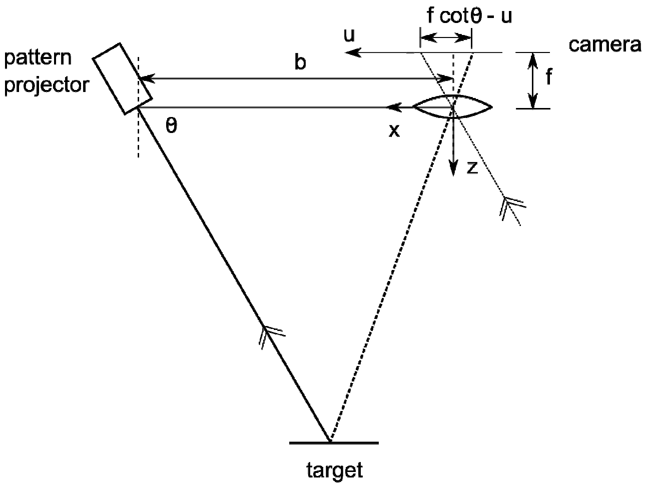


Figure 2: Schematic showing the principle of structured light triangulation (adapted from [11])

Laser scanners and TOF cameras operate on similar principles to each other. Both have of an emitter that emits a pulse of light and a receiver that measures the round trip time of the light. For typical measurement distances the round trip time is in the order of picoseconds and therefore the electronics required to measure the time directly are expensive. TOF cameras measure the phase shift of modulated light reflected off a target to calculate the distance for a grid of pixels simultaneously. Laser scanners have a single receiver that is mechanically scanned and uses pulse travel time or phase shift to measure distance.

Commercial TOF cameras use a modulated near-infrared light source and measure the phase shift between the transmitted and received light [10]. The maximum unambiguous distance (D_{unamb}) to a target would be:

$$D_{unamb} = \frac{c}{2f} \qquad (8)$$

where f is the modulation frequency of the light source. Any distance less than D_{unamb} is calculated by measuring the ratio of the phase shift (φ) to a full cycle and multiplying it by the maximum distance.

$$d = \frac{\varphi}{2\pi} D_{unamb} \qquad (9)$$

One of the problems with TOF cameras is due to phase shift ambiguity. A phase shift of slightly over 2π would be measured as a shift of just greater than zero and according to Equation 9 the calculated distance would be close to zero.

3 RESULTS

This section describes the results of subsystem testing using preliminary indoor data. A dataset was taken in a corridor environment using the FLIR A300 thermal camera. The thermal images from the FLIR were segmented to extract ROIs that could possibly be humans. The ROIs were classified by hand to provide a ground truth dataset. The regions were classified as containing: a single standing person, multiple overlapping people, a partial image of a person or no person. The classification resulted in a training set containing sub-images of 332 people, 55 groups of people, 126 partially occluded people and 1287 sub-images not containing a person. This ground-truth data was used for the training and verification of the classification algorithms.

The SwissRanger SR4000 TOF camera and a Microsoft Kinect structured light 3D sensor were tested in an operational mine and the results are discussed in Section 3.3.

3.1 Segmentation

Figure 3 shows an image from the FLIR camera. Ideally the ROIs should only be the two people in the image. It is shown that a simple temperature threshold-based ROI extraction performs better than two more complex algorithms.

The first ROI extraction algorithm uses a combination of intensity and edge information. The algorithm extracts regions with a certain intensity surrounded by strong edges. It was found that objects in the thermal images are invariably surrounded by edges that are incomplete. A robust integration was used that could highlight regions surrounded by incomplete edges but it is computationally intensive.

A histogram based segmentation algorithm, using Otsu's threshold selection method, was also tested for segmentation. Otsu's method is commonly used for greyscale image thresholding [12]. Otsu's method assumes a bimodal distribution of intensities and attempts to optimally divide the distribution into two. Otsu's threshold selection does not work on the thermal images. This is because the temperature distribution is uni-modal due to the uniformity of the background temperature.

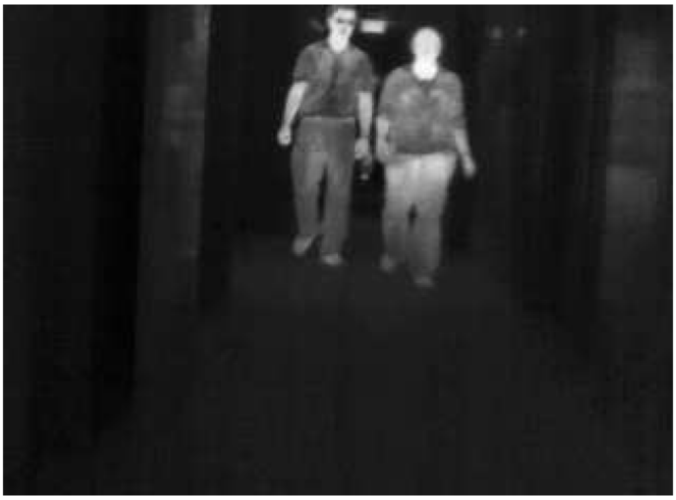


Figure 3: An example image for ROI extraction

It was found that a simple temperature threshold based segmentation performed better than the two above-mentioned algorithms. The temperature threshold extracts regions that have a temperature of between 26.8 and 37 °C and then performs a morphological opening, on the binary image created, to remove small noise regions. The ROIs extracted using the temperature threshold are shown in Figure 4.

Following thresholding, each region in the binary image is numbered using a connected component labelling method so that the regions can be classified separately.

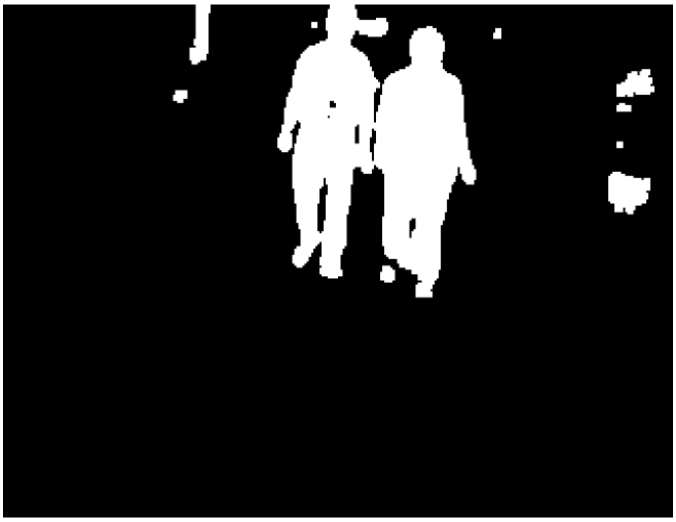


Figure 4: ROIs extracted with the temperature range threshold

3.2 Classification

Each classifier classifies the ROIs as a single standing person or something else. The dataset of 1800 manually classified regions is randomly divided into training and evaluation datasets, each of approximately the same size (a random division with equal chance of being in each set). Each classifier is trained and then run three times, the first time it is run using the data from the evaluation set. The two subsequent tests are run using a new randomly chosen sub-set of the data. Each classifier is evaluated in terms of its classification accuracy and speed.

The classifiers are all run in MATLAB R2010b on a 2.8 GHz Pentium 4 PC. The speed of each classifier is averaged over the three tests and the results are shown in Table 1.

Table 1: A comparison of classifier speeds. (running in MATLAB R2010b)

Classifier	Speed (classifications/s)
Template	4830
Parzen	552
Neural Network	1227
Support Vector	1677

Figure 5 shows typical Receiver Operating Characteristic (ROC) curves for each of the classifiers.

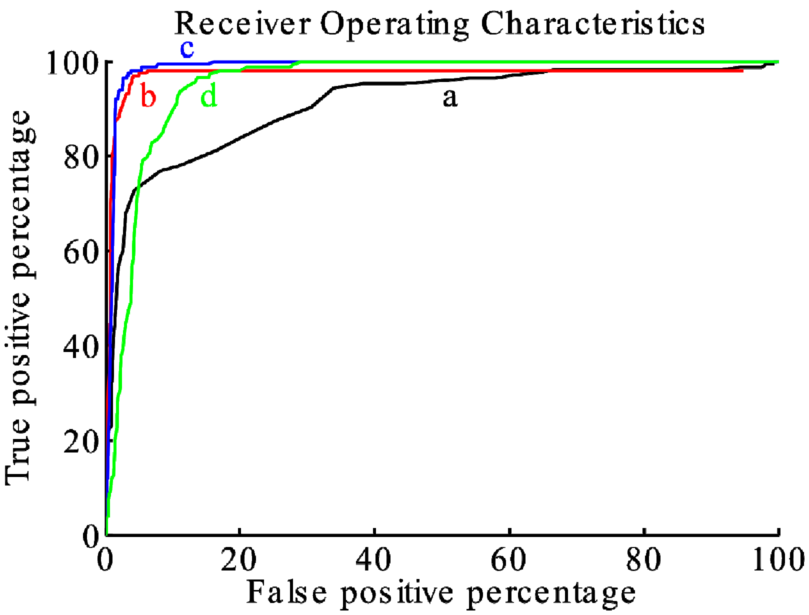


Figure 5: The Receiver Operating Characteristics of a) the template classifier, b) the Parzen classifier, c) the neural network and d) the support vector classifier

The performance of the template classifier is significantly poorer than the other two and does not warrant further consideration despite being the fastest.

The support vector classifier shows intermediate classification results but performs significantly worse than the Parzen and neural network classifiers. The support vector classifier is the second fastest because classification involves a single matrix multiplication, an addition and a sign check.

The neural network achieves very similar classification performance to the Parzen classifier. The main difference between the two is that the Parzen classifier achieves a maximum true positive rate of 98% while the neural network can detect 100% of the targets (albeit with a high false positive rate). The classifier is required to detect people without missing any, i.e. the true positive rate needs to be close to 100%. The effect of false positives is less severe simply adding to the number of objects that need to be tracked. Consequently achieving a 100% detection rate is an important characteristic of a classifier for pedestrian detection.

The neural network classifier achieves slightly better detection performance and a significantly faster classification than the Parzen classifier. The neural network classifier also achieves a significantly lower number of false positives compared to the support vector classifier. The higher speed of the support vector classifier is not sufficient to compensate for inferior performance. The neural network classifier is therefore the classifier of choice for the proposed human detection system.

3.3 Distance Sensors

Testing of the two 3D sensors underground showed a significant disadvantage of using TOF camera technology in a harsh underground environment.

The drilling of blast holes in a mine gives off a fine water spray; coupled with high humidity this creates a fine mist in active areas of the mine. The TOF camera's amplitude image in Figure 6 shows the water mist near the base of the support in the centre of the image. The distance image shown in Figure 7 shows a significant jump in measured distances near the base of the support due to the mist there.

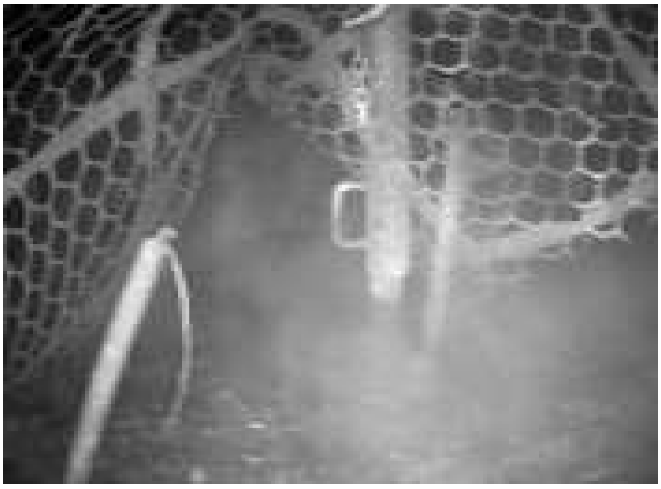


Figure 6: Time of Flight camera amplitude image through mist

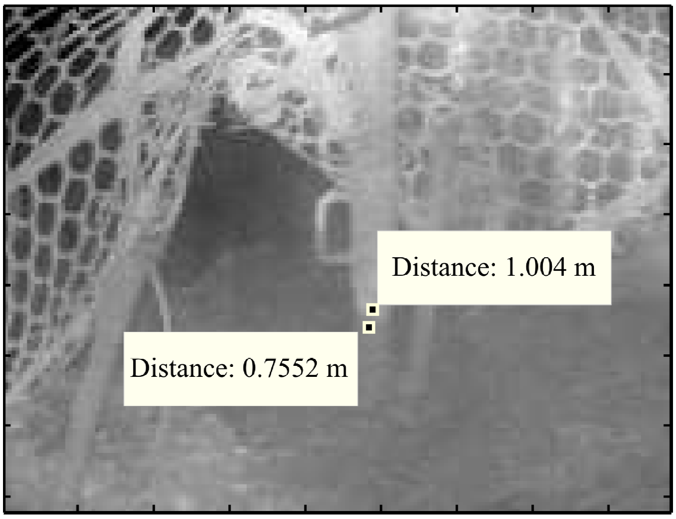


Figure 7: Time of Flight camera distance image through mist

The reason for the poor performance of the TOF camera is that the camera is receiving a reflection off the object of interest as well as multiple reflections off the intervening water droplets. The reflection off the mist causes the received phase shift to be less than the true value and therefore the measured distance is shortened. It is expected that dust, which will be more of a problem in the tunnels where the pedestrian detection system will operate, will have a similar effect as the mist.

The TOF camera was also found to suffer from significant motion blurring due to the fact that a single range image is calculated using four phase measurements. Reducing the integration time of the camera would reduce the blurring but would decrease the range of the camera.

The structured light Kinect sensor seems unaffected by the mist. This is probably because the processing hardware calculates the distance on the basis of the most intense reflection. Without a known ground-truth distance the effect of the mist on the accuracy of the Kinect remains undetermined.

4 CONCLUSION

This paper examines a proposed pedestrian detection system in underground mines using a fusion of 3D information with thermal imaging. This system is proposed in response to the high number of fatalities in the mining industry caused by underground transportation machinery and the fact that current pedestrian detection systems are limited. The architecture of the proposed system is outlined and the steps of segmenting images and classifying them described. It is shown that due to the thermometric nature of the images, temperature range-based segmentation is superior to other more complex segmentation methods. A neural network classifier is chosen for the detection system because of its superior performance on the test dataset. It is shown that a neural network classifier outperforms a Parzen classifier slightly in accuracy and significantly in speed. The neural network is slightly slower than a support vector classifier but achieves similar detection rates with far fewer false positives. An evaluation of two 3D cameras shows that TOF cameras suffer from inaccuracies due to obscuring mist. The structure light camera appears unaffected by the same obscuring mist but further work is needed to confirm this.

5 RECOMMENDATIONS

Further work required involves the acquisition of a large underground dataset for testing, including a dataset from a moving platform in order to test the calculation of vehicle velocity from the 3D data. The acquisition of a large dataset will enable the classifier to be tested and optimised for the mine environment.

Work is also required to determine whether the effect of dust on the TOF camera is similar to the effect of mist, as suspected. A quantitative analysis of the effect of dust on the accuracy of the TOF and structured light 3D sensors is also required.

6 ACKNOWLEDGEMENTS

The authors would like to acknowledge Mathew Price of Cogency for the acquisition software used for data gathering. We would also like to thank the Bafokeng Rasimone Platinum Mine (BRPM) for allowing us access to the mine to gather data.

7 REFERENCES

[1] Seccombe, *A., Decline in mine deaths 'too good to be true'*, Business Day, 07 January 2011.

[2] Green, J., Bosscha, P., Candy, L., Hlophe, K., Coetzee, S., Brink, S., *Can a Robot Improve Mine Safety*, 25th International Conference of CAD/CAM, Robotics & Factories of the Future, Pretoria, 2010.

[3] Green, J., Vogt, D., *A Robot Miner for Low Grade Narrow Tabular Ore Bodies: The Potential and the Challenge*, 3rd Robotics & Mechatronics Symposium, Pretoria, 2009, http://hdl.handle.net/10204/4115.

[4] Ruff, T., *Advances in Proximity Detection Technologies for Surface Mining Equipment*, in Proc. of 34th AIMHSR, Salt Lake City, 2004.

[5] *Proximity Detection* - National Institute of Occupational Safety and Health, [Online] http://www.cdc.gov/niosh/mining/topics/topicpage58.htm , August 2010.

[6] *Avoiding accidents with mining vehicles* - FLIR Commercial Vision Systems, Application Story, 2008

[7] Nanda, H., Davis, L., *Probabilistic template based pedestrian detection in infrared videos*, IEEE Intelligent Vehicle Symposium, Vol. 1, pp. 15 – 20, 2002.

[8] Fehlman, W., Hinders, M., *Mobile Robot Navigation with Intelligent Infrared Image Interpretation*, Springer, 1st ed., 2009.

[9] Gunn, S., *Support Vector Machines for Classification and Regression*, Technical Report, Image Speech and Intelligent Systems Research Group, University of Southampton, 1997, http://www.isis.ecs.soton.ac.uk/resources/svminfo/.

[10] Sphikas, P., *SR4000 User Manual*. MESA Imaging, Zurich, 2010

[11] Siegwart, R., Nourbakhsh, I., *Introduction to Autonomous Mobile Robots*, The MIT Press, Cambridge, Massachusetts, 1st ed., pp. 122-128, 2004.

[12] Otsu, N., *A Threshold Selection Method from Grey-level Histograms*, IEEE Transactions on Systems, Man and Cybernetics, Vol. 9, pp. 62-66, 1979.

Appendix E

Consecutive Missed Detections

Tables E.1 to E.3 shows which misclassified images, from the cross-validation tests, are consecutive. Knowing how many of the misclassification are consecutive is essential to determine the probability of making multiple consecutive misclassifications (since consecutive classification errors are not independent). It can be seen from the tables that subsequent detections are definitely not independent, this is not unexpected since the change in appearance of a person between consecutive frames is relatively small.

Table E.1: A table indicating which false negatives are consecutive for the first test

Test 1	
Image	Consecutive with previous
1	no
2	no
3	no
4	no
5	no
6	no
7	no
8	no
9	no
10	yes
11	yes
12	yes
13	yes
14	yes

Continued on next page

Table E.1 – continued from previous page

Test 1	
15	yes
16	yes
17	yes
18	no
19	yes
20	no
21	no
22	no
23	no
24	no
25	no
26	no
27	no
28	no
29	no
30	no
31	no
32	yes
33	no
34	no
35	no
36	no
37	no
38	no
39	no
40	yes
41	yes
42	yes
43	yes
44	yes
45	no

Table E.2: A table indicating which false negatives are consecutive for the second test

Test 2	
Image	Consecutive with previous
1	no
2	no
3	no
4	yes
5	no
6	no
7	no
8	yes
9	yes
10	no
11	no
12	no
13	no
14	no
15	no
16	no
17	no
18	no
19	no
20	no
21	no
22	no
23	no
24	yes
25	yes
26	yes
27	yes
28	no
29	no
30	no
31	no
32	yes
33	no
34	no

Table E.3: A table indicating which false negatives are consecutive for the third test

Test 3	
Image	Consecutive with previous
1	no
2	no
3	no
4	no
5	no
6	no
7	no
8	no
9	no
10	yes
11	no
12	yes
13	yes
14	yes
15	yes
16	no
17	no
18	yes
19	no
20	no
21	no
22	no
23	no
24	no
25	no
26	no
27	no
28	no
29	no
30	no
31	no
32	no
33	no
34	no