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 $°C$ however ventilation and other cooling brings the temperature within working areas down to below 30 $°C$ 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 $°C$ (heat stress management programmes are required for $T_{wb} > 27.5\ °C$) [11, 12]. Therefore, it is assumed that the rock temperature within the mine tunnels will be below 28 $°C$.

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

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

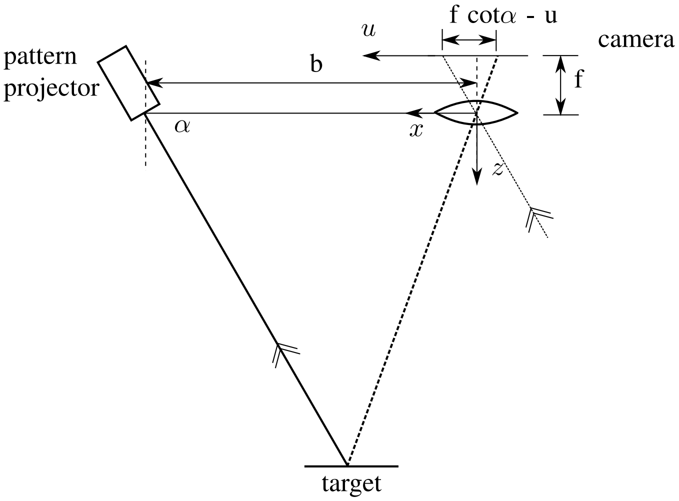

Fig. 1. Schematic showing the principle of structured light triangulation (Adapted from Siegwart and Nourbakhsh [20])

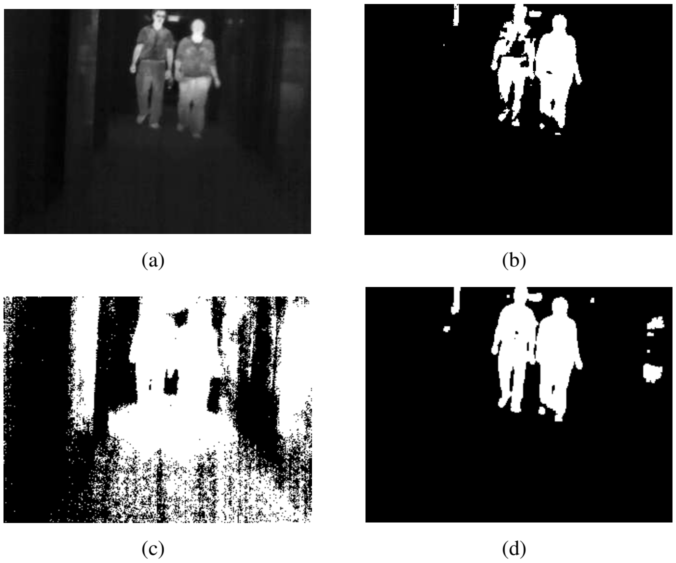

Fig. 2. Results of segmentation tests: (a) is the input image; (b) shows the result of edge and intensity segmentation; (c) is the result using Otsu's method; and, (d) is the result using temperature threshold-based segmentation

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 $^{\circ}C$ and 37 $^{\circ}C$ 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

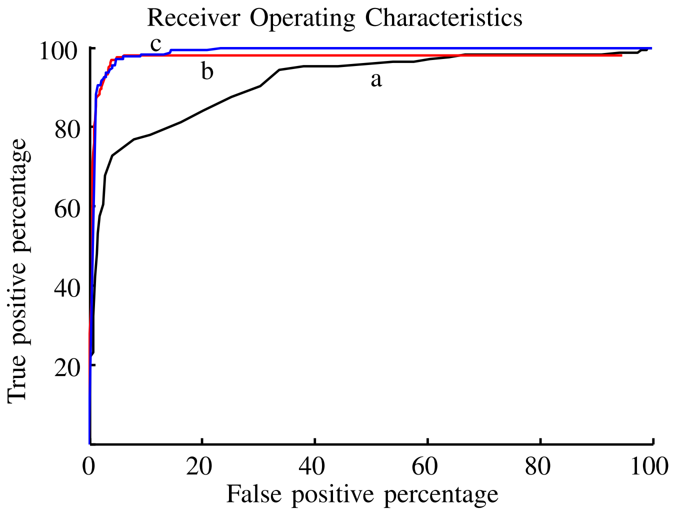

Fig. 3. The Receiver Operating Characteristics for a) the Template classifier, b) Parzen classifier and c) Neural Network

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.

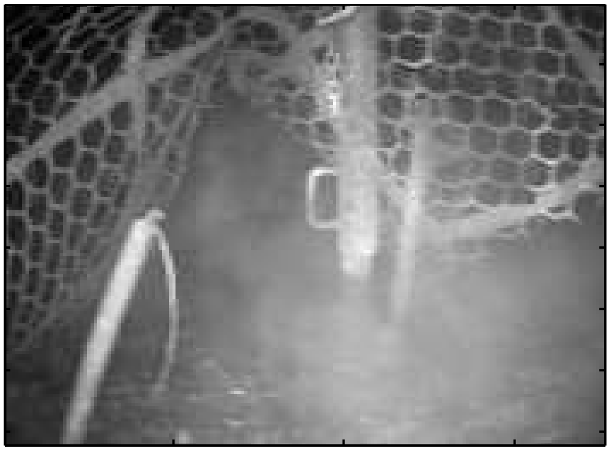

Fig. 4. Time of flight camera amplitude image through mist

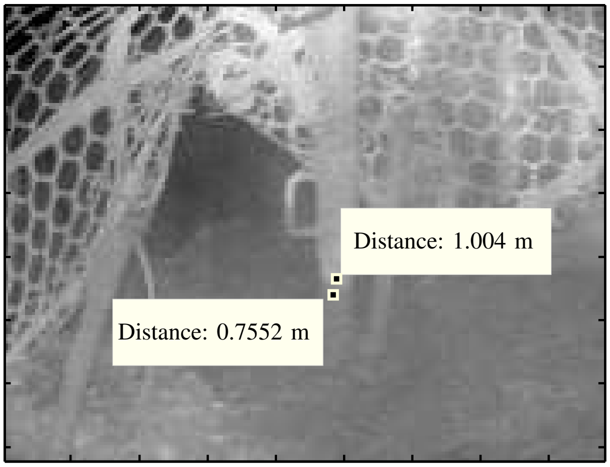

Fig. 5. Time of flight camera distance image through mist

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.