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.:

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

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 (O_j), i.e.:

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

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. $P(O_j|D)$.

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

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

$P(O_j)$ 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:

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

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

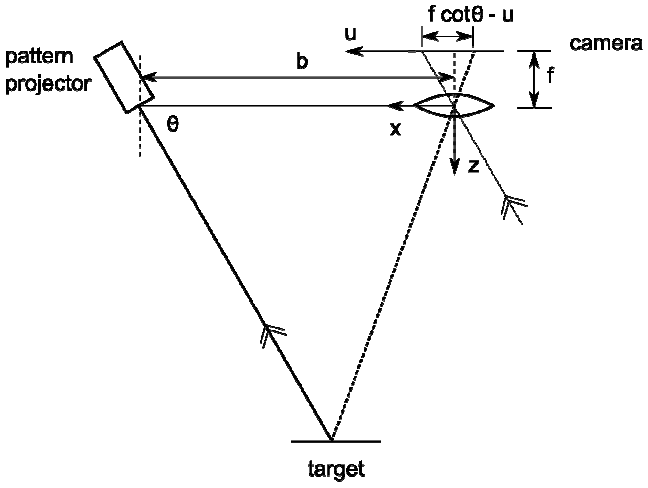

Figure 2: Schematic showing the principle of structured light triangulation (adapted from [11])

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:

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

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.

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

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.

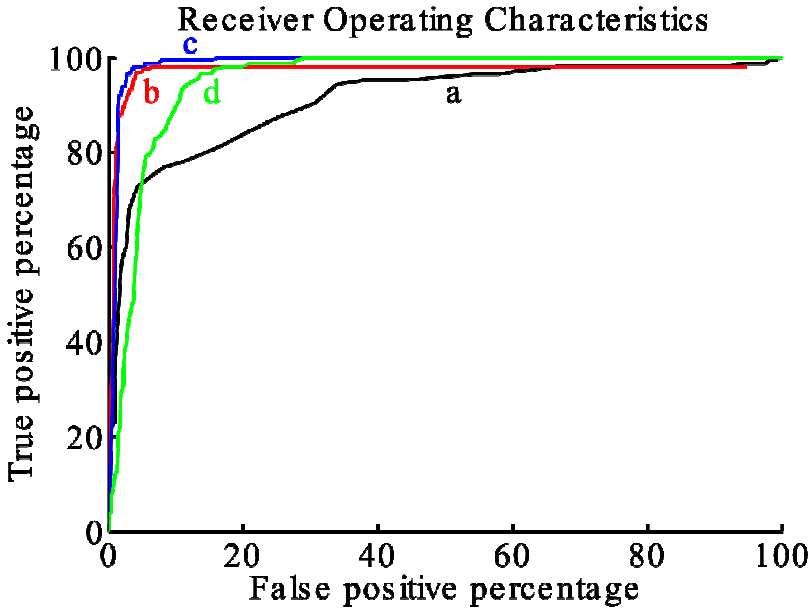

Figure 5: The Receiver Operating Characteristics of a) the template classifier, b) the Parzen classifier, c) the neural network and d) the support vector classifier

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

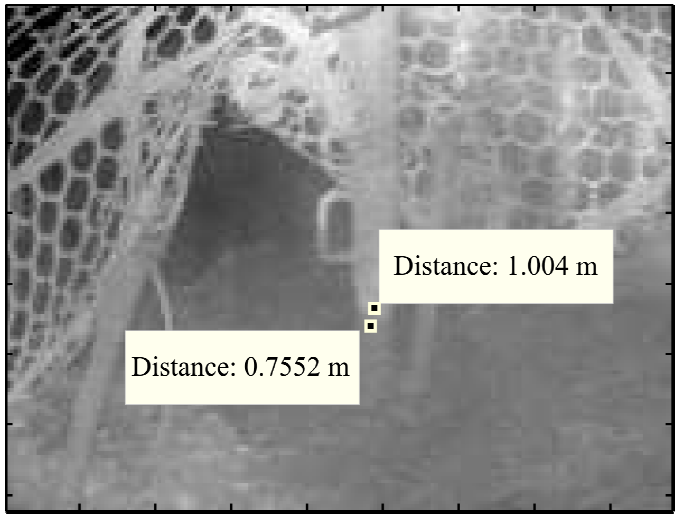

Figure 7: Time of Flight camera distance image through mist

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.