From ImageNet to Mining: Adapting Visual Object Detection with Minimal Supervision

Alex Bewley and Ben Upcroft

Abstract This paper presents visual detection and classification of light vehicles and personnel on a mine site. We capitalise on the rapid advances of ConvNet based object recognition but highlight that a naive black box approach results in a significant number of false positives. In particular, the lack of domain specific training data and the unique landscape in a mine site causes a high rate of errors. We exploit the abundance of background-only images to train a k-means classifier to complement the ConvNet. Furthermore, localisation of objects of interest and a reduction in computation is enabled through region proposals. Our system is tested on over 10km of real mine site data and we were able to detect both light vehicles and personnel. We show that the introduction of our background model can reduce the false positive rate by an order of magnitude.

1 Introduction

While the mining industry pushes for greater autonomy, there still remains a need for human presence on many existing mine sites. This places significant importance on the safe interaction between human occupied and remotely operated or autonomous vehicles. In this work, we investigate a vision based technique for detecting other vehicles and personnel in the workspace of heavy vehicles such as haul trucks.

Traditionally, methods for detecting light vehicles and personnel from heavy mining equipment have relied on radio transponder based technologies. Despite

Alex Bewley
School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia, e-mail: aj.bewley@qut.edu.au

Ben Upcroft
ARC Centre of Excellence for Robotic Vision, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia. http://www.roboticvision.org/, e-mail: ben.upcroft@qut.edu.au

transponder based sensors being mature and reliable for ideal conditions, in practise their reliability is circumvented by practical issues around their two way active nature, portable power requirements, limited spatial resolution and human error. Using computer vision offers a unique alternative that is passive and readily available on existing remotely operated vehicles.

Vision based object recognition has made tremendous progress as measured by standard benchmarks [4, 16]. The major advancements in this area can be attributed to both the availability of huge annotated datasets [7, 26, 4, 16] and developments in data driven models such as deep convolutional networks (ConvNets) [13, 24]. In this work we utilise the ConvNet of [13] which has shown astonishing performance on the ImageNet recognition benchmark [4] and extend it to data collected from mine sites with minimal training.

Using ConvNets in different domains requires a large training set relevant to the target task [29]. When the amount of training data is small, data driven approaches tend to over-fit the training samples and not generalise to unseen images. In this work we utilise a pre-trained ConvNet using millions of images from ImageNet and address how to map the original ImageNet classes to mining classes with minimal training effort.

Another consideration regarding this application is that cameras are rigidly coupled to the vehicles orientation and configured with a fixed focal length. This distinguishes it from the ImageNet recognition problem where typical images collected were implicitly pointed at regions of interest and appropriately zoomed. Additionally, due to the wide field of view the majority of the images are background with zero to potentially multiple objects of interest visible in any given frame. To locate the objects, we follow a similar strategy to [10] and apply an initial step for finding likely object locations through a region proposal process before performing object recognition with the ConvNet.

Given that the majority of the images collected in a mine site dataset have zero objects of interest in them, we can provide a standard classifier with a huge amount of labelled background data. Using this newly trained classifier in conjunction with the ConvNet ensures robustness and drastically reduces spurious detections. This classifier is based on k-means clustering offering a convenient way to partition the background data into different categories. This approach accurately captures the characteristics of the background, enabling the discovery of novel non-background objects.

The contributions of this paper are:

- adapting ConvNets to new scenes in a mining context,
- complementing the powerful classification provided by ConvNets with a simple classifier trained on background mine data for increased robustness,
- a novelty detector using ConvNet feature clustering.

This paper is organised with a short review of related literature before describing the proposed method in greater detail. We then analyse the performance of the proposed method on a challenging set of mining videos and conclude with a discussion of the learnt outcomes and avenues for future improvement.

2 Related Work

Here we briefly review object detection methods that are not reliant on two way communication before covering some related work using ConvNets for generic object detection. Early work has focused on range based techniques such as LiDAR [22, 17] commonly used for mapping fixed obstacles such as buildings or underground tunnel walls. Applying these sensors to detecting personnel and vehicles fitted with retro-reflectors, is found to be sensitive to the dynamics of the sensor platform [20]. In this work we focus specifically on detecting potentially dynamic obstacles including vehicles and particularly people from vision based data. To this end, the more relevant prior work is that of [18] which exploits the standardised requirement for personnel on mine sites to wear high-visibility clothing equipped with retro-reflector strips. This enables a single IR camera with active flash to highlight personnel in view which can then be used for tracking [19].

Recent popularity of big data and deep learning have dominated the object recognition problem. Among these data driven approaches, deep convolutional neural networks (ConvNets) with recognition performance quickly approaching human levels [13, 5, 21, 23] are selected for use in this work. ConvNets themselves have been used for over 20 years [14] for tasks such as character recognition. Over recent years ConvNets have made an astonishing impact on the computer vision community [13, 6, 21, 10, 5] thanks to the availability of huge labelled image sets such as ImageNet [3].

Recognising what objects are in an image is only half of the object detection problem. The other half is locating the objects within the image. Sermanet et al. [23] sample over multiple scales and exploit the inherently spatially dense nature of the convolutions within ConvNets to identify regions with high responses. Similarly, [6] also perform convolutions over multiple scales and combine the responses over superpixel segmentation [9]. Another popular approach and the one that we base this work off is the region convolutional neural network (RCNN) of [10]. The RCNN framework efficiently combines the ConvNet of [13] with an object proposal method: selective search [27]. Generic object proposal methods aim to efficiently scan the entire image at different scales and aspect ratios to reduce potentially millions of search windows down to hundreds [11] of the most likely candidates. In this work we use edge box object proposals [30] as the accuracy is higher while also running at an order of magnitude faster [11].

3 Methodology

In this section we outline our detection pipeline and how it differs from [10]. Our method consists of three key phases: 1) Region proposals with non-maximum suppression (NMS), 2) ConvNet recognition and finally, 3) Detections are validated by checking for novelty against the background model. See Fig. 1 for a high-level overview of this pipeline. We bypass the problem of over-fitting on a small dataset

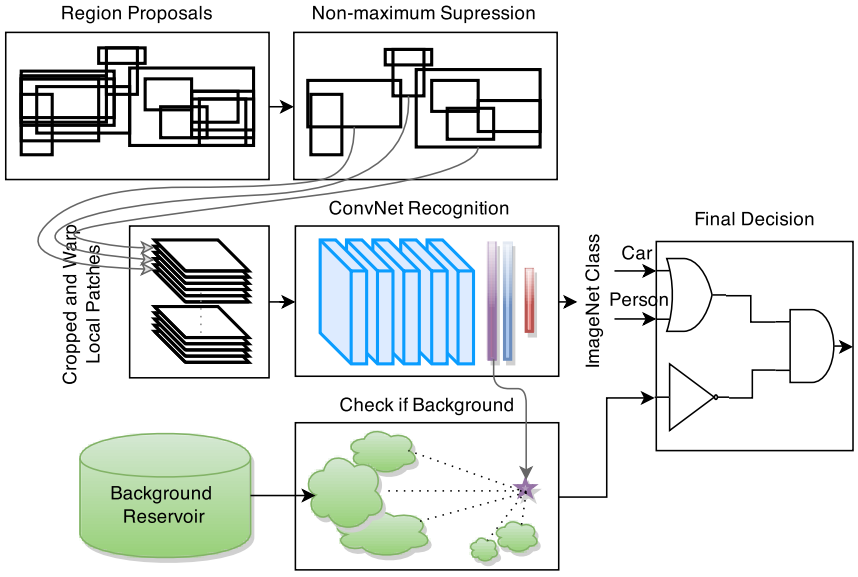

Fig. 1 An illustration of the detection pipeline used in this work. The system parameters are highlighted in blue and green which are learnt offline from an off-the-shelf network and background only images respectively. Note the red output layer of the ConvNet outputs are ImageNet classes (200 different). Any car or person is suppressed if it also matches the background model to minimise the number of false positives.

by using a pre-training ConvNet and map its output to mining relevant classes. This method is then extended with our proposed background modelling technique to significantly reduce the number of false positives generated by the system.

3.1 Region Proposals

The aim of region proposals is to efficiently scan the image to eliminate millions of potential windows, keeping only the regions that are likely to contain an object of interest. We use the `EdgeBoxes` region proposal method [30] over the `selective search` [27] used in the original RCNN work as this method is orders of magnitude faster with comparable accuracy. For a detailed comparison of region proposal methods we refer the reader to [11].

The default parameters for `EdgeBoxes` were adjusted to return a fixed 1000 proposals. These region proposals are then further reduced to approximately 100 regions through a process of non-maximum suppression (NMS). The NMS process considers the score produced by the `EdgeBoxes` method and the overlap with other bounding boxes. As the name suggests it then greedily suppresses all but the

maximum scoring proposal for all adjacent regions overlapping by 30% or more. In contrast to applying NMS after the ConvNet [10], this way we can speed up the detection pipeline by reducing the number of proposals going into the ConvNet while maintaining comparable coverage over the image.

3.2 Region Classification

Having selected regions of the image that have the general characteristics of an object, we now perform object recognition to distinguish the object category. For this we apply the ConvNet from RCNN [10] which is based on the winning architecture [13] for the ImageNet Large Scale Recognition challenge in 2012. For this work, we used the RCNN implementation provided with the Convolutional Architecture for Fast Feature Embedded (`caffe`) [12] framework out-of-the-box.

The original detection task for RCNN was to predict one of 200 classes that represent common objects found in images taken from the internet. For this application we are only interested in distinguishing between three high level categories, namely: `background`, `person` and `light vehicles(LV)`. Using this model in a mining context raises several issues that need addressing:

1. Most of the 200 classes are irrelevant, e.g. jellyfish, miniskirt, unicycle etc.
2. How to associate mining classes with ImageNet classes?
3. Semantically the `background` is significantly different from many of the existing object specific classes.

To gain some insight, we use a small validation set of 200 images to investigate the output of the ConvNet out-of-the-box. This set is made up of cropped mine-site images containing the classes `person` and `LV` along with 90 interesting region proposals extracted from background only images. We also included a few `heavy vehicles (HV)` images in this set but keep them as a separate class to identify any correlations. In Fig. 2 we show the results of naively applying the pre-trained RCNN model to this image set. To better visualise the output we applied a soft-max transform to approximate the output class prediction as a probabilistic estimate. [1]

Not surprisingly, the `person` and `LV` classes are well represented and can be directly mapped from the `person` and `car` ImageNet classes used to train the original ConvNet. On the other hand, the `background` closely resembles uniform random sampling of classes as there are no relevant classes in the existing model such as trees, buildings, or road signs etc. Similarly, the `HV` class prediction also mostly resembles a uniformly random distribution with a slight bias towards the ImageNet classes `snowplow`, `cart` and `bus`. As for this application, we are only

[1] It is important to note that this is for visualisation purposes only and that the *y*-axis does not represent the true probability since the final SVM layer of RCNN was not calibrated for probabilistic outputs.

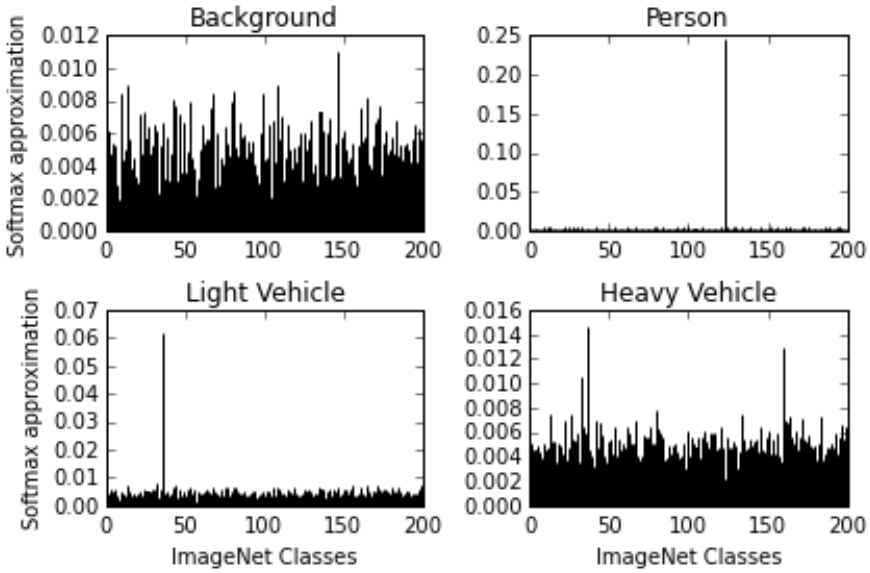

Fig. 2 The average class estimate for a set of mining related images. Notice that person (class 123) and light vehicle/car (class 36) are existing classes for the pre-trained network and can be used directly. The background and the heavy vehicle classes are novel and show a wider spread as they are not modelled with the pre-train ConvNet.

concerned with distinguishing `person` and `LV` from the background, we simply assign all 198 non `person` or `car` outputs as background.

With this simple class mapping approach and assuming that falsely picking one of the positive classes is in fact uniformly random, we expect to eliminate 99% of all the proposed background regions. However, when processing around 100 proposals per frame, the expected false positive rate is once per frame. Next we propose a simple background model that reuses the ConvNet computation to provide a background likelihood estimate for reducing this false positive rate.

3.3 Background Modelling

While on a mine-site the landscape is constantly changing from a geometric perspective, the bleak visual appearance of the background is generally constant. For this, we model the background regions as belonging to one of an arbitrary set of categories, such as the semantic categories of rock, sky, tree etc. If a sample differs significantly from any of these background classes then we can assume it is an object of interest.

Rather than using supervised techniques that require a set of manually annotated images, we instead partition the background data without explicit semantic labels.

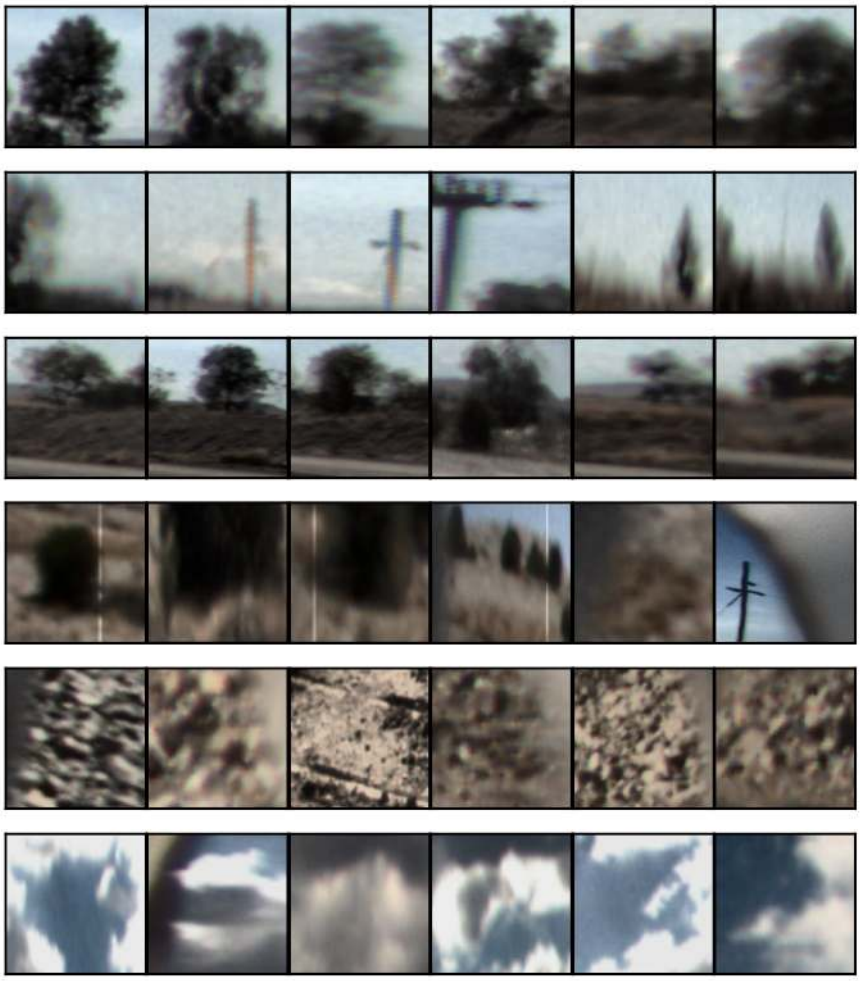

Fig. 3 An illustration showing six of the most common types of background region proposals. The rows represent different clusters while the columns show a random background region which is a member of the associated cluster. Each cluster gathers samples with similar visual appearance such as centred on a tree (top row) or centred on sky with an adjacent vertical structure (second row).

To do this, we exploit the assumption that intra-category samples generally appear visually similar to each other, yet may be distinctively different to other background categories. Put another way, the background regions form natural clusters enabling us to employ unsupervised techniques to model their visual appearance. See Fig. 3 for an illustration of the natural background clusters found by applying this method to a mining dataset.

To describe the visual appearance of each region, the intermediate layers of the ConvNet provide a free and compact representation suitable for this task. Addition-

ally, these features have been shown to be robust against lighting and viewpoint changes without any re-training [25]. We refer the interested reader to [13] for an illustration of the ConvNet's inner workings. In general, the first layer of a ConvNet extracts simple colour and texture features in the first layer, and through subsequent layers, these features eventually transition to the learnt specific task [29] such as classifying the 200 ImageNet classes. Along the way irrelevant visual information for the original task (e.g. features describing sky) are lost once it reaches the final layer. With this intuition we reuse the transformed data from one of the ConvNet's intermediate layers as an input to our background model.

To learn this cluster based model, a reservoir of negative samples is required. Gathering background data is a relatively simple task since only inspection for the presence of target objects is necessary. Specifically any image sequence not containing any of the target objects can be used to build an extremely large reservoir by extracting proposals from each frame. Furthermore, we only focus on difficult regions by perform hard-negative-mining [8] of background samples by running the ConvNet detection pipeline over these sequences. By lowering the confidence threshold, near false positive background regions can also be added to build a sufficiently large reservoir.

After extracting an intermediate layer of the ConvNet for each background patch, we then cluster these samples using k-means clustering. At test time, each `person` or `LV` predicted patch is verified by measuring the Euclidean distance between its intermediate feature and each cluster centre. If the nearest background cluster is close in this feature space, i.e. is visually similar, then we suppress the detection and regard it as background.

In building this background model the following choices are to be made: Which layer from the ConvNet? How many clusters? At what distance should a sample be considered background? In the following section we address these design choices through experimental validation.

4 Experiments

4.1 Mining Dataset

The dataset we use for evaluating this work was collected from a light vehicle mounted camera operating in an active mine-site, see Fig. 4. While the motivation is to put vision based sensing on a heavy vehicle, a light vehicle is more practical for gathering a diverse set of visual sequences. The dataset contains both static and dynamic instances of a `person`, `LV` or `HV`.

Continuous video was gathered with and without the camera in motion and on various haul roads and a few light vehicle only zones to capture variation in the environment. This video data was captured at 10 fps and partitioned into various sequences. In this work we use 5 sequences where no people or vehicles are visible

Fig. 4 The experimental dataset gathering vehicle with cameras mounted to the bullbar. Note: all images used in this paper were captured from the camera on the left hand side of the vehicle.

to build our background model. Collectively these background sequences make up 8952 frames in total (approximately 14km).

To evaluate the performance we use another 5 sequences with several instances of `person`, `LV` or `HV`, that we personally annotated using the tool developed by Vondrick et al [28]. These annotated sequences contain 9405 frames in total (approximately 10km). In addition to these sequences we made a small validation set of 200 using other images collected on a mine site from various sources including a few captured at night. This set was used to generate Fig. 2.

4.2 Background Model Validation

Here we describe the experiments performed to design our background modelling system explained in the previous section. From the 5 background sequences, we applied the region proposal and ConvNet detection framework to find challenging region proposals from every tenth frame. While some of the false objects may be observed in multiple frames, the time difference is sufficient to capture a variety of view points for these distracting objects. We lowered the detection threshold to collect region proposals if the ConvNet predicted either a `person` or `car` in the top 5 out of 200 class responses. With this configuration we collect around 8000

hard negatives for our background reservoir. We held out 90 of the most interesting background regions and added them to the validation set.

To address the design decisions for this model, we perform an empirical study using the reservoir containing only negatives and the validation set with both negative and positives. We jointly test different combinations of ConvNet layer features and number of clusters by evaluating their performance on the validation set. For the distance threshold we set this to the distance corresponding to a 95% recall on the positive set. With the recall fixed, the overall performance of the background model is measured by the precision at which it can identify a true negative.

Fig. 5 shows the relative performance of sweeping the number of clusters for different ConvNet layers. While `fc6` layer with 2048 clusters achieved the highest precision of 90% we instead opted to use only 128 clusters with a precision of 89% which is significantly faster to compute. A detailed view of the distances between the validation samples and the cluster centres can be seen in Fig. 6.

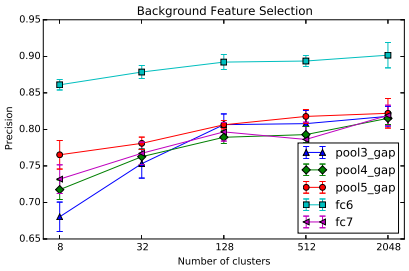

Fig. 5 Cross-validation precision at 95% recall for different ConvNet layers and the number of clusters used to represent the background. Each point shows average of 5 trials.

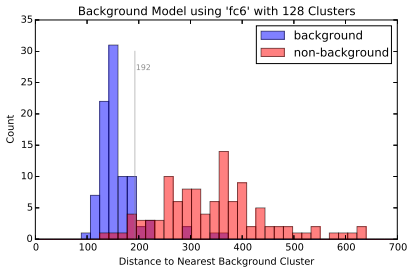

Fig. 6 Detailed view of the distribution of the validation images distance to their nearest background cluster centre. The grey line marks the 95% recall distance threshold.

Implementation Detail

The first 5 ConvNet layers produce dense tensor representations which gradually reduce in size. Then there are two `fully connected` layers `fc6` and `fc7` before the final prediction layer. Again we refer the interested reader to [13] for details of the ConvNet structure. Due to the density of data and the computational complexity of computing distances in such high dimensional feature spaces we only evaluate the ConvNet layers 3-7 and compress convolutional layers 3-5 by pooling all filter responses across the feature map for each tensor in [15] this is referred to global average pooling. In Fig. 5 these are marked as pool{3-5}_gap.

Fig. 7 Validation samples where the background model failed. Images are shown in their warped form, representing the ConvNet input. The four right false negatives were collected at night.

The false negatives and some of the false positives are also shown in Fig. 7. The false negatives are mostly night images which can be put down to the fact that similar images are rare if not non-existent in the ImageNet samples used to train the ConvNet. For the false positives, these are mostly signs which make up a minority of the scene. From these samples we can describe our background model as a form of novelty detection where interesting parts of the scene such as signs are distinguished from the general background. This finding along with the unsupervised clustering shown in Fig. 3 are a testament to the ConvNet's expressive capabilities in representing visual similarity.

4.3 Detection Evaluation

We now evaluate the system on the set of 5 sequences with `person` or `LV` where the task is to locate objects of interest. In this evaluation we consider a true detection if at least 50% of the detection region is covered by a single ground truth object. This differs from the intersection-over-union (IOU) definition of overlap, as we accept detecting a `person`'s head and shoulders without their whole body while IOU would count this as both a miss detection and a false positive. It should be also noted that any detection or miss detection of a `person` or `LV` labelled as partially occluded in the ground truth is ignored in this evaluation. While the system is not designed to detect `HV` we consider any detections which overlap with `HV` objects as neither true or false and are excluded from the evaluation. Additionally, if multiple detections overlap a single ground truth instance, we count this as a single true positive and neither of the overlapping detections are false. An example would be if a person's head is covered by a single detection and their body another.

Table 1 shows the performance of the system before and after applying background suppression. From these results we can see that while there is a slight drop in recall our method for suppressing background regions reduces the false positive rate by an order of magnitude.

Table 1 System Comparison before and after Background Suppression (BGS) on Mining Sequences

Sequence (frames)	F1 Score[a] baseline	(Precision, Recall) with BGS[3]	Mostly Hit[b] -	BGS	Mostly Missed[b] -	BGS	False Positives -	BGS
1 (1462)	0.38 (0.57,0.29)	**0.40** (0.77,0.27)	2	2	16	16	242	**87**
2 (2950)	**0.94** (0.96,0.91)	0.93 (0.97,0.89)	3	3	6	6	73	**47**
3 (599)	0.02 (0.01,0.09)	**0.06** (1.00,0.03)	0	0	2	2	349	**0**
4 (2826)	0.64 (0.56,0.74)	**0.80** (0.95,0.69)	**2**	1	**4**	5	186	**9**
5 (1568)	**0.68** (0.78,0.61)	0.43 (0.92,0.28)	**4**	1	**3**	6	177	**24**
Total			11	7	31	35	1027	**167**

[a] F1, Precision and Recall is computed treating each frame as independent.
[b] Mostly indicates where a single object instance was detected or missed 50% of the time.
[c] The proposed background suppression (BGS) is applied to the baseline `EdgeBox` and ConvNet detector.

5 Conclusions and Future Work

In this paper we presented a vision only system that takes advantage of recent developments in computer vision and machine learning to detect both personnel and light vehicles. We circumvented the problem of ConvNet over-fitting on small datasets by reusing a pretrained model directly and mapping its output to mining classes. We further presented a method for exploiting the abundance of background only images to learn a background cluster model leading to a significant reduction in false positives. This sensing approach was evaluated in an active open-pit mine site environment. The experiments show that the in-pit environment is suitable for object proposals along with background modelling techniques such as the one presented here.

While this work is only concerned with single camera based sensor data we see many opportunities to combine techniques incorporating stereo [2] or range-based sensors [20] for improved robustness. As an initial investigation of vision as a possible sensor on a mine we see many opportunities to further improve on the results. As

more labelled mining image data becomes available we expect to be able to design and fine-tune a ConvNet that performs better in this domain than the existing network. We also plan to extend this work to fuse information from multiple frames by combining the ConvNet appearance model with recent motion segmentation techniques [1].

Acknowledgements This research was funded by the Australian Coal Association Research Program (ACARP). The authors would also like to acknowledge AngloAmerican for allowing data collection at the Dawson operation. Acknowledgement also goes to the high performance computing group at Queensland University of Technology for both support and use of their services when conducting the experiments in this paper.

References

1. Alex Bewley, Vitor Guizilini, Fabio Ramos, and Ben Upcroft. Online Self-Supervised Multi-Instance Segmentation of Dynamic Objects. In *International Conference on Robotics and Automation*, Hong Kong, China, 2014. IEEE.
2. Alex Bewley and Ben Upcroft. Advantages of Exploiting Projection Structure for Segmenting Dense 3D Point Clouds. In *Australian Conference on Robotics and Automation*, 2013.
3. Huawu Deng and David a. Clausi. Unsupervised image segmentation using a simple MRF model with a new implementation scheme. *Pattern Recognition*, 37(12):2323–2335, December 2004.
4. Jai Deng, Wei Dong, Richard Socher, Li-Jai Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.
5. Jeff Donahue, Judy Hoffman, Erik Rodner, Kate Saenko, and Trevor Darrell. Semi-supervised Domain Adaptation with Instance Constraints. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 668–675, June 2013.
6. Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning Hierarchical Features for Scene Labeling. *Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, 2013.
7. Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, April 2007.
8. Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–45, September 2010.
9. Pedro F. Felzenszwalb and Daniel P. Huttenlocher. Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision*, 59(2):167–181, September 2004.
10. Ross B. Girshick, Jeff Donahue, Trevor Darrell, and J Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.
11. Jan Hosang, Rodrigo Benenson, and B Schiele. How good are detection proposals, really? In *British Machine Vision Conference (BMVC)*, 2014.
12. Yangqing Jia. {Caffe}: An Open Source Convolutional Architecture for Fast Feature Embedding. \url{http://caffe.berkeleyvision.org/}, 2013.
13. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, pages (Vol. 1, No. 2, p. 4), 2012.

14. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1:541–551, 1989.
15. Min Lin, Qiang Chen, and Shuicheng Yan. Network In Network. *arXiv preprint*, 2013.
16. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pages 740–755. Springer International Publishing, 2014.
17. Joshua A. Marshall and Timothy D. Barfoot. Design and field testing of an autonomous underground tramming system. *Springer Tracts in Advanced Robotics*, 42:521–530, 2008.
18. Rafael Mosberger and Henrik Andreasson. Estimating the 3d position of humans wearing a reflective vest using a single camera system. In *International Conference on Field and Service Robotics (FSR)*, 2012.
19. Rafael Mosberger, Henrik Andreasson, and Achim J. Lilienthal. Multi-human Tracking using High-visibility Clothing for Industrial Safety. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 638–644, 2013.
20. Tyson Phillips, Martin Hahn, and Ross McAree. An evaluation of ranging sensor performance for mining automation applications. In *2013 IEEE/ASME International Conference on Advanced Intelligent Mechatronics: Mechatronics for Human Wellbeing, AIM 2013*, pages 1284–1289, 2013.
21. Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014.
22. Jonathan M. Roberts and Peter I. Corke. Obstacle detection for a mining vehicle using a 2D laser. In *Proceedings of the Australian Conference on Robotics and Automation*, pages 185–190, 2000.
23. Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. *International Conference on Learning Representations (ICLR 2014)*, December 2014.
24. Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun. Pedestrian Detection with Unsupervised Multi-stage Feature Learning. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3626–3633. IEEE, June 2013.
25. Niko Sunderhauf, Feras Dayoub, Shirazi Sareh, Upcroft Ben, and Milford Michael. On the Performance of ConvNet Features for Place Recognition. In *arXiv*, 2015.
26. A Torralba, R Fergus, and W Freeman. 80 Millions Tiny Images: a Large Dataset for Non-Parametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1958–1970, 2008.
27. J. R. R. Uijlings, K. E. a. Sande, T. Gevers, and a. W. M. Smeulders. Selective Search for Object Recognition. *International Journal of Computer Vision*, 104(2):154–171, April 2013.
28. Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently Scaling up Crowdsourced Video Annotation. *International Journal of Computer Vision*, 101(1):184–204, 2012.
29. Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks ? In *Advances in Neural Information Processing Systems (NIPS)*, 2014.
30. CL Zitnick and P Dollár. Edge Boxes: Locating Object Proposals from Edges. In *European Conference on Computer Vision (ECCV)*, 2014.