https://doi.org/10.2352/ISSN.2470-1173.2020.6.IRIACV-049

Rare-Class Extraction Using Cascaded Pretrained Networks Applied to Crane Classification

Sander R. Klomp; Eindhoven University of Technology, SPS-VCA group of Electr. Eng.; Eindhoven, The Netherlands
Guido M.Y.E. Brouwers; ViNotion b.v.; Eindhoven, The Netherlands
Rob G.J. Wijnhoven; ViNotion b.v.; Eindhoven, The Netherlands
Peter H.N. de With; Eindhoven University of Technology, SPS-VCA group of Electr. Eng.; Eindhoven, The Netherlands

Abstract

Overweight vehicles are a common source of pavement and bridge damage. Especially mobile crane vehicles are often beyond legal per-axle weight limits, carrying their lifting blocks and ballast on the vehicle instead of on a separate trailer. To prevent road deterioration, the detection of overweight cranes is desirable for law enforcement. As the source of crane weight is visible, we propose a camera-based detection system based on convolutional neural networks. We iteratively label our dataset to vastly reduce labeling and extensively investigate the impact of image resolution, network depth and dataset size to choose optimal parameters during iterative labeling. We show that iterative labeling with intelligently chosen image resolutions and network depths can vastly improve (up to 70×) the speed at which data can be labeled, to train classification systems for practical surveillance applications. The experiments provide an estimate of the optimal amount of data required to train an effective classification system, which is valuable for classification problems in general. The proposed system achieves an AUC score of 0.985 for distinguishing cranes from other vehicles and an AUC of 0.92 and 0.77 on lifting block and ballast classification, respectively. The proposed classification system enables effective road monitoring for semi-automatic law enforcement and is attractive for rare-class extraction in general surveillance classification problems.

Introduction

Many practical real-world surveillance problems consist of detecting rare classes of objects, for which it is difficult to obtain labeled training data. This data is required to train supervised detection and classification systems. One way to obtain labeled training data in these cases, is to watch hundreds of hours of surveillance videos and then to manually label the rarely occurring class, which is infeasible in practice. This paper focuses on avoiding this laborious effort through an iterative labeling method applied to the practical case of classifying mobile construction cranes on public roads.

The weight of mobile crane vehicles is often beyond the legal per-axle limit, because they carry their lifting blocks and ballast on the vehicle instead of on a separate trailer. Enforcement is desirable, since this causes pavement and bridge damage, as pavement degradation is significantly higher when vehicles are above legal weight limits [1]. Monitoring of these overweight vehicles can be achieved using Weigh-In-Motion systems [2], although these are expensive to install and maintain, because they require altering of the pavement. Since these sources of weight are visible, a camera-based detection system can provide a more cost-effective alternative. Performing this detection is a straightforward detection/classification problem, for which supervised learning using Convolutional Neural Networks (CNNs) is well suited. Because mobile cranes occur rarely on public roads and labeling all video data is infeasible, an iterative learning system is desirable to gather sufficient data and train an effective crane classifier successfully.

Iterative labeling methods have been explored extensively in the past, especially in the field of active learning [3–5]. The basic method for active learning consists of a repeated two-step process: (1) train a classifier on a labeled subset of data, and (2) use some query function to retrieve unlabeled images to be labeled and adding them to the labeled set. However, most existing methods focus on labeling generic datasets and do not exploit the ‘rare new subclass’ case within the dataset. Specific active learning for rare classes, including discovering rare classes in data, is performed in [6]. From this, we adopt the same basic paradigm, but use a different variant of uncertainty sampling [7] to rapidly gather samples for labeling.

Our contributions are threefold. First, we describe a system to solve a specific fine-grained classification problem with a rare subclass: distinguishing mobile cranes from other types of trucks. The system is set up to be easily trainable with an iteratively labeled dataset. Second, we show a practical method to iteratively obtain more data of rare subclasses of object classes in surveillance problems. This method is based on transfer learning and is inspired by active learning and hard negative mining. It vastly reduces the manual effort required to label sufficient data to train a classifier on the rare subclass. Third, we extensively evaluate the impact of the amount of labeled data on the performance of the system for different image resolutions and CNN network backbones, to determine when sufficient rare-class samples have been collected.

Method

This section describes the entire processing chain of creating a crane classification system. First, the system design is introduced, consisting of a cascade of fine-tuned CNN classification systems. Second, the proposed method for iteratively labeling the large training set of unlabeled images is explained. Finally, all network parameters are listed for reproducibility.

System architecture

The system architecture consists of a pretrained vehicle detector (based on SSD [8]), recognizing vehicles in the video stream from the camera. Each detected vehicle is processed by

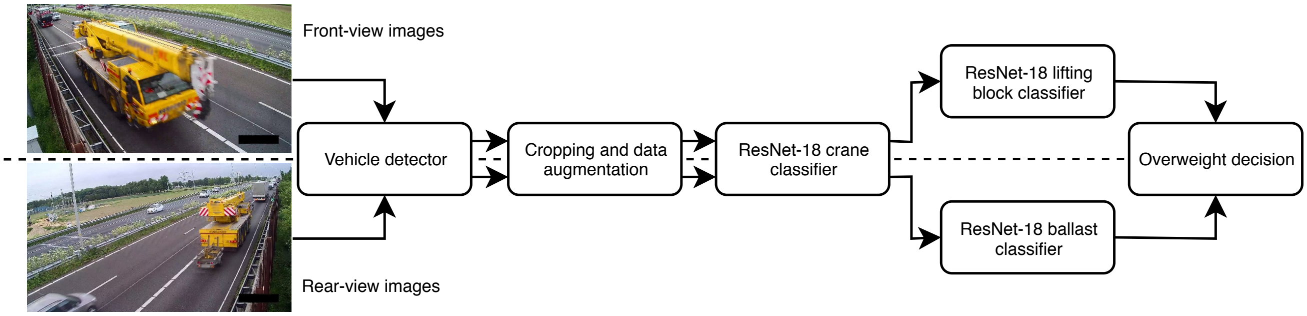

Front-view images
Vehicle detector
Cropping and data augmentation
ResNet-18 crane classifier
ResNet-18 lifting block classifier
Overweight decision
ResNet-18 ballast classifier
Rear-view images

Figure 1: Schematic representation of the system. The dashed line represents a split of data flows. The front- and rear-view camera images are processed by a single vehicle detector and crane classifier. Separate classifiers are applied for detecting lifting blocks on front-view images and ballast on rear-view images.

three separate image classifiers (ResNet-[9]). A schematic representation is shown in Fig. 1.

The vehicle detector has been trained with a large dataset of surveillance videos containing broad categories of traffic (cars, trucks, motorcycles). We opt for adding a separate classifier after the detector, as compared to retraining the detector. Repeatedly retraining with the large traffic dataset and only a few novel crane images is time-consuming, especially in combination with the iterative labeling method explained later. To construct the initial dataset from which the cranes will be extracted, the ‘truck’ detections from the vehicle detector are employed. The detection boxes are extended in all directions by 10%, after which the images are cropped to this extended box and used to train the classifier.

The additional classifier consists of a binary ResNet classifier, which is trained to distinguish between ’truck’ and ’crane’. In our specific case of overweight crane classification, we further subdivide this class into cranes with and without a lifting block, and cranes with and without ballast. For each independent problem a supplementary binary classifier is added, resulting in a total of three binary ResNet classifiers. Under the assumption that there are only a few samples of the rare subclass in the total training set, all three classification networks will train quickly, especially in early labeling iterations. This removes the need for incremental learning strategies to rapidly update the networks after every period of data acquisition. Finally, if either a lifting block or ballast is detected, the crane is marked as potentially overweight, which is to be manually verified by a crane legislation expert.

Dataset acquisition and labeling

To gather sufficient crane images for training, video data was recorded on two highway locations for a period of five weeks. Two cameras were installed at each location, capturing both rear and front views of traffic. A total of 3,360 hours of high-definition video was gathered. The vehicle detector described in the previous section is used to extract crops of all ’truck’ class vehicles, resulting in 1.2 million truck image crops. Since manual interpretation of these images to extract cranes is infeasible, we propose to exploit an iterative labeling procedure with the following steps:

1. Obtain small set of labeled ’crane’ vs. ’not crane’ images;
2. Apply data augmentation, train classification network;
3. Run inference on the unlabeled images;
4. Extract *N* images with the highest ’crane’ probability;
5. Label these *N* images manually;
6. Repeat from Step 2.

For very rare subclasses, performing Step 1 from the full unlabeled dataset may already be infeasible. An alternative is to gather images from the Internet. Although these are generally obtained from highly different viewpoints, as compared to the used surveillance camera footage, only mild discriminative power is required to start the iterative labeling procedure. In our specific case, we have obtained access to a small government database of 300 low-quality front-view crane images, which serve as a similar starting point to web-crawled images.

Data augmentation in Step 2 is applied to reduce the likelihood of the network, discovering only highly similar images to the small initial dataset. The augmentations are randomized between iterations of the labeling procedure, to improve the chances of discovering crane images with a lower similarity to the initial labeled images.

Step 4 is similar to uncertainty sampling. However, instead of searching for images with a probability close to 0.5, which will almost exclusively be regular trucks due to the dataset imbalance, we choose images with crane probability close to unity. The concept of selecting high-probability samples is based on hard negative mining, which is commonly used to improve object detection models. Selecting high-probability crane images for labeling will yield easy crane examples and the most crane-like truck examples. By labeling these and retraining, the number of false positive detections of the network reduces rapidly, since the network is forced to shape the decision boundary around the hardest negative samples. Note that this step is the most computationally intensive step of the algorithm, as it requires running inference on the entire unlabeled set of images (1.2 million). We process the entire dataset on every iteration because performing inference on the set only takes minutes. As a possible speed up, one may choose to order the dataset on rare-class probability after the first iteration and use a fraction of the unlabeled dataset with the highest rare-class probabilities for subsequent iterations. Note that although this avoids performing inference on many images, it risks missing a small number of rare-class images. Step 5 involves the most manual labeling effort, although significantly less effort than labeling the entire dataset. After 9 iterations of manually labeling about 2,000 images (consuming a few hours of labeling effort), over 90% of cranes are labeled. This has been verified with data from a nearby government-owned weigh-in-motion system that measures every crane observed in the camera view. Thus, the total manual labeling effort is reduced from 1.2 million to $\sim$18,000 images, which represents a reduction factor of 67. These labeled

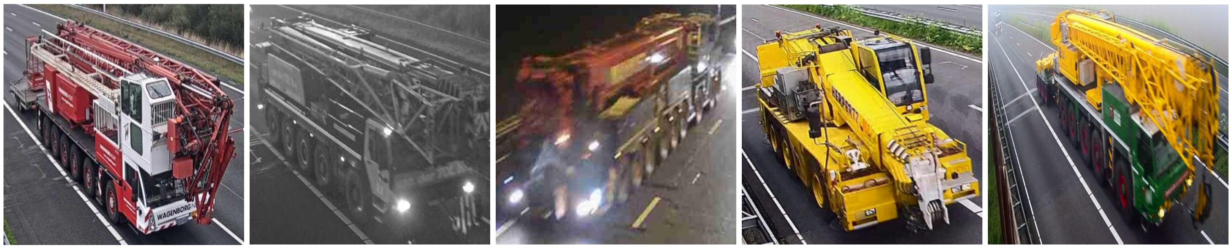


Figure 2: Examples of intra-class variation of the ‘crane’ class.

images contain 2,116 images of mobile cranes and 15,763 images of trucks. Some examples of crane images are shown in Fig. 2 to show the intra-class variation, which is larger than expected.

Network parameters

All ResNet classification networks (pre-trained on ImageNet [10]) are trained for 12 epochs using the PyTorch and FastAi frameworks, with batch size 32. We follow the convention of FastAi for transfer learning. This consists of cutting-off the ImageNet pretrained network at the last convolutional layer, freezing all convolutional layers and appending the following custom head to be trained on the new dataset. The head consists of a concatenation of the outputs of an adaptive average pooling layer and an adaptive max pooling layer, which is flattened and processed by several fully connected layers with dropout, batch normalization and ReLU activations. These modifications are according to the convention of FastAi. Finally, the chosen optimizer is stochastic gradient descent with cyclical learning rates for increased convergence speed [11].

Experimental results

Iterative labeling vastly reduces labeling effort. To investigate the impact of several hyperparameters on the effectiveness of the iterative labeling procedure, four experiments are performed. Selecting optimal parameters based on the current dataset size can reduce training time and thereby reduce the time for additional iterations, or reduce the number of iterations required by training a better classifier with less data. The investigated problem has been addressed from the viewpoints of:

1. Relation between image size and dataset size;
2. Network depth compared to dataset size;
3. The impact of data augmentation;
4. Optimal image resolution per classification sub-problem (crane vs. truck, lifting block, ballast classification).

The experiments consist of training classification networks for 360 combinations of the corresponding parameters (see Table 1). The images from one highway camera setup are used for training and the images from the second for testing. We perform all experiments with three different random seeds to measure variation during the training process. As the classification networks are binary, Area Under the receiver operating characteristic Curve (AUC) is employed to measure network performance. Similar experiments have been performed by Kolesnikov *et al.* [12], who provided insight into the impact of the amount of training data for very large general classification datasets (1.3M to 300M images) and deeper networks. However, they investigate neither much smaller amounts of data (relevant to our iterative labeling problem), nor the impact of image resolution compared to dataset size. In the following subsections we present the separate experiments.

Table 1: Network parameter sweep values.

Parameter	**Values**
Class problem	Crane vs. truck, lifting block, ballast
Classifier model	ResNet-18, ResNet-34, ResNet-50
Image width & height	32, 64, 128, 256, 512
Subsampling factor	1, 4, 16, 64
Data augmentation	Yes, no

Relation between image size and dataset size

When training on a small number of training samples, training on too high resolution may cause networks to overfit on noisy high-frequency features, instead of focusing on stable low-frequency features in the images. To verify this and determine at which point overfitting becomes relevant, we train the "crane vs. truck" classifier with various amounts of data by subsampling the dataset, while varying the input image resolution. Sub-sampling is done with factors 4, 16 and 64, and resolution is varied from 32×32 to 512×512 pixels. Resolution changes are applied both during training and testing. For reference: source image resolutions vary between 400-600 pixels in both width and height. The results are shown in Fig. 3. Very low resolutions always result in worse performance when comparing to using higher resolutions regardless of dataset size (the three lowest resolution lines are exclusively below the top-two resolutions). However, there appears to be a saturation point around 256×256 pixels, where the performance is very close to 512×512 pixels for large datasets and can even outperform the higher resolution on the smallest dataset size (overlapping standard deviation areas at the right-side of the figure).

Considering that training with lower-resolution images is significantly faster than when using high-resolution images, training at 256×256 pixels is optimal for initially gathering samples from the new rare class. When a significant number of samples (>10,000) have been gathered, one may consider changing to a higher resolution, as it can yield a small performance improvement.

Optimal resolution per classification problem

While using higher image resolutions is generally superior for the crane vs. truck classification problem, this does not necessarily generalize to problems of ballast detection and lifting block detection, as features for these problems can be informative on different scales. Fig. 4 shows the result of performing a sweep over image resolution when training over the entire

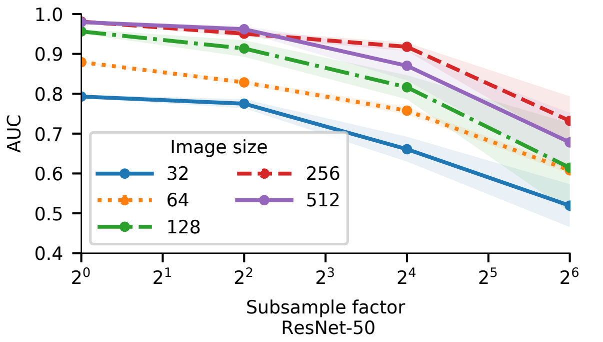


Figure 3: Impact of dataset size and image size on crane vs. truck AUC scores for different ResNet architectures. Shaded areas indicate standard deviations. Sub-sample factor 1 equals ∼9000 training images and sub-sample factor 64 equals 140 training images.

dataset. The saturation point of performance remains consistently at 256×256 pixels. Specifically for the ballast detection problem, training on 512×512 pixels may even deteriorate performance, although the large standard deviation in the results means this conclusion is uncertain.

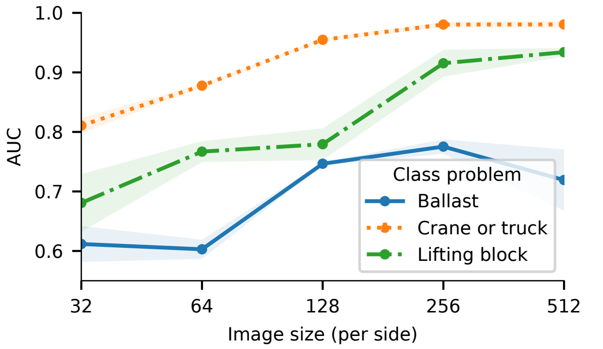


Figure 4: Impact of resolution on performance for the three different classification problems. Shaded areas indicate one standard deviation.

Relation between dataset size and network depth

The relation between dataset size and network depth for the crane vs. truck problem at fixed size 512×512 pixels is shown in Fig. 5. As expected, deeper networks outperform more shallow networks when sufficient data is available. Additionally, when very little data is available (1/64th of the total train dataset, which corresponds to only 12 crane images), shallow networks outperform larger ones. However, the deeper networks rapidly outperform the smaller networks when more data is gathered.

Considering that training a pretrained network to convergence with very little data is extremely fast, even for the deeper networks, it makes sense to use the deeper networks as soon as their performance is expected to surpass the shallower networks. From Fig. 5 it can be seen that ResNet-50 is optimal from around sub-sample factor 4 (2^2) or lower, which corresponds to ∼2,000 labeled images or more.

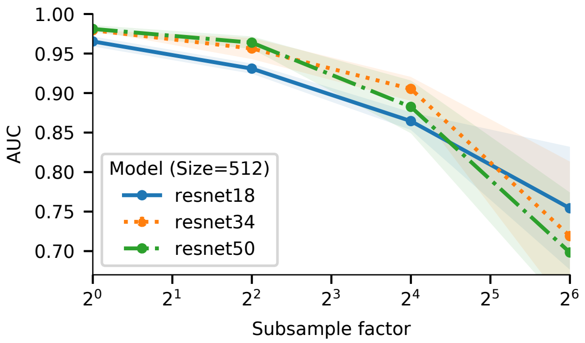


Figure 5: Impact of network depth on crane vs. truck classification performance for different dataset sizes.

Impact of data augmentation

Data augmentation is commonly employed to improve performance of classification networks, although it may be much less valuable in constrained situations, such as video surveillance scenarios where the viewpoint hardly varies and similar cameras with identical (and limited) resolution are employed. To verify whether data augmentation can still lead to performance improvements, we apply the following augmentation techniques: (a) Random horizontal flipping; (b) Small random rotations (up to 10 degrees); (c) Random brightness and contrast changes; (d) Random minor perspective warping. All augmentations are sufficiently mild that a human can still easily recognize the images, and they are chosen to be representative changes for surveillance scenes. Rotations and perspective warping represent the small differences in camera viewpoints between train and test set. Likewise, lighting changes are covered by variations in brightness and contrast, and horizontal flipping represents placing a camera on the other side of the road (vehicles are typically symmetrical). Because data augmentation is known to be more valuable for smaller datasets, we measure the influence of data augmentation for different dataset sizes. In addition, different resolutions are explored, since reducing image resolution also corresponds to a reduction in the amount of data. The latter aspects are illustrated by Fig. 6. It appears that in our constrained setup, much of the advantage of data augmentation disappears, regardless of both the dataset or the image size. This means that in a constrained surveillance setting, as long as imaging setups are sufficiently similar, dataset augmentations have limited to no impact on performance.

Final classification results

After training, the crane vs. truck classification network achieves an AUC score of 0.985 score on the test set, by employing ResNet-50 and all gathered data at image resolution 512×512 pixels. The high score is as expected, because cranes and trucks are relatively easily distinguishable, even by non-expert humans. It is interesting to see in which cases the classification network fails. There are five primary causes for mistakes, examples of which are shown in Fig. 7: First, in some crops created by the vehicle detector, both a crane and a truck are visible, which generally causes the network to label the image as "truck", because the imbalance in the dataset gives "truck" an implicitly higher prior probability. In contrast, humans tend to label these images as cranes instead, as that is the rare class of interest

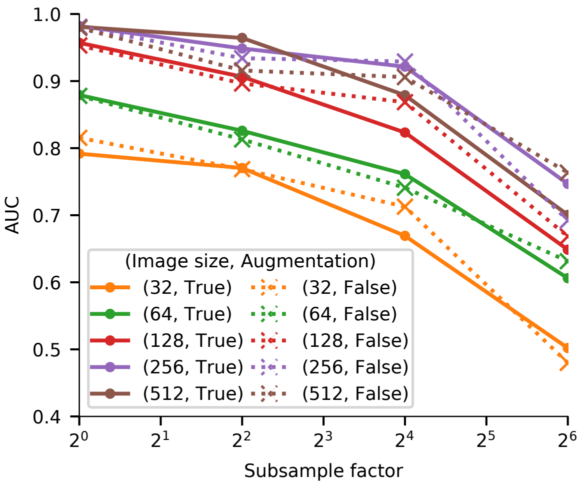


Figure 6: Impact of data augmentation for different image sizes using ResNet-50. Standard deviation is not plotted to avoid visual clutter, but generally the dotted and straight lines are within one standard deviation of each other.

for this problem. This is likely solvable by class balancing, either by re-sampling the dataset or by weighting the loss per class. Second, a few images are simply manually mislabeled while the network was correct. Third, some images suffer from extremely poor lighting conditions, which also makes it hard for a human to distinguish the classes. Fourth, the crops are sometimes imperfect. If a large part of the vehicle falls outside the cropping area, important features to distinguish the classes fall outside the image. This results from inaccurate box estimation by the vehicle detector (which was not explicitly trained using the cranes). Fifth, some classes of trucks are extremely similar to mobile cranes, but are technically speaking not mobile cranes. The most common case of such mistakes are cherry pickers (shown in Fig.7e). The aforementioned five cases cover approximately half of the incorrect classifications. The other incorrect classifications are mainly "obvious mistakes", which are likely solvable through additional data, since Fig. 5 does not appear to be completely saturated yet when using the full dataset.

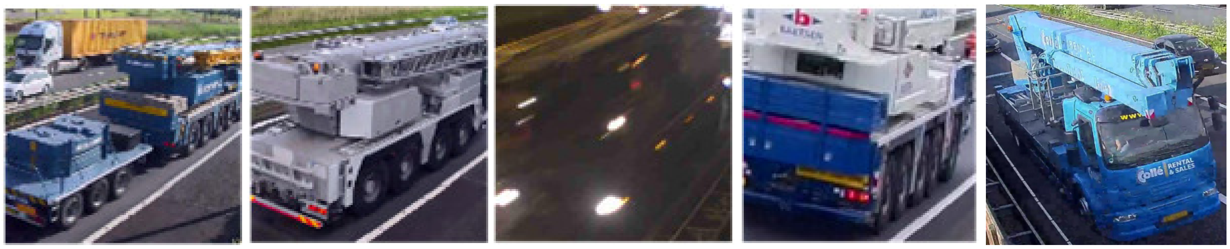
The two classification networks for the ballast and lifting block sub-problems achieve lower AUC scores: 0.77 and 0.92, respectively (using the entire train set, ResNet-50, 512×512 pixels). The first reason for the lower score is the smaller training set size, as these networks train on subsets of the total set of crane images. For ballast the performance is low, because ballast is often painted in the same color as the crane, and may be carried at several different places on the crane. This makes distinguishing ballast a challenging problem for both neural networks and (non-expert) humans. Despite being trained with expert-annotated labels, the ballast detection network achieves similar performance to non-expert humans. For lifting block classification, the primary problem is dataset bias. A certain type of mobile cranes generally carries lifting blocks in the train set, while another type of cranes does not, causing the network to become biased to the crane type. Therefore, the lifting block classification result depends in several cases on the crane type rather than the actual presence/absence of lifting blocks. When using a detection network instead, the score may improve, because the bounding box label would partially prevent the network from overtraining on other crane properties unrelated to the lifting block.

Discussion

The system can effectively support law enforcement by automatically recognizing cranes and the presence of ballast and lifting blocks. Additionally, several interesting conclusions can be drawn from the performance curves (Figs. 4-6). However, this study has several notable limitations. First, the accuracy for lifting block and ballast detection is not yet high enough for fully automated enforcement, so that system still requires a small amount of manual verification. This can potentially be combined with likely desirable checking actions that prevent unjust fines, which would be desirable anyhow for systems with very high accuracy for legal justification. Second, our camera-based system is a viable replacement to weigh-in-motion systems in settings where the placement cost is important, such as temporary measurement locations, although perhaps not in permanent setups where sufficient budget is available to alter the pavement. Finally, the datasets on which the experiments have been performed are somewhat limited in scope, hence the conclusions may not generalize to situations beyond fixed-camera traffic-based surveillance problems. However, the experimental results are consistent over the three subproblems that we have investigated, which provides an indication that the results are at least likely to generalize to other problems. Experiments to confirm this on larger generic datasets are left for future work.

Conclusion

In this work we have shown that automated detection and classification of cranes from image data can aid law-enforcement in discovering overweight cranes using automated recognition from images of road-side cameras. The system detects cranes on public roads with high accuracy (0.985 AUC) and recognizes cranes with lifting blocks (0.92 AUC) and ballast (0.77 AUC) automatically. To solve this problem, we have proposed an iterative labeling approach, which vastly reduces (factor 67) the manual labor, involved with labeling the images. Several experiments have been performed to determine the optimal image resolution and convolutional neural network depth for various stages of the iterative labeling process. These experiments show that for low amounts of training data, shallow networks and medium image resolutions are optimal, while for larger amounts of training data, deeper networks and higher image resolutions achieve the best performance. This suggests that for optimal results, these parameters should be dynamically updated while the iterative labeling is in progress. Designing a method for automatically updating these parameters is an interesting direction for future work. A second direction is comparing the current cascaded detector + classifiers approach with a single-model hierarchical detector that has explicit sub-classes for the three classification problems, as the latter could more easily generalize to a broad range of surveillance problems. Overall, we conclude that a camera-based inspection system with deep learning is sufficiently accurate for semi-automatic law enforcement of overweight mobile cranes.



(a) Both classes present. (b) Incorrect label. (c) Extreme lighting. (d) Poor detection box. (e) Confusing class.

Figure 7: Examples of misclassified images in the test set.

References

[1] K. C. Dey, M. Chowdhury, W. Pang, B. J. Putman, and L. Chen, "Estimation of Pavement and Bridge Damage Costs Caused by Overweight Trucks," *Journal of the Transportation Research Board*, vol. 2411, no. 1, 2014.

[2] H. Wang, J. Zhao, and Z. Wang, "Impact of Overweight Traffic on Pavement Life Using Weigh-In-Motion Data and Mechanistic-Empirical Pavement Analysis," in *9th International Conference on Managing Pavement Assets*, no. 1500, 2015, pp. 1–13.

[3] Y. Baram, R. El-Yaniv, and K. Luz, "Online Choice of Active Learning Algorithms," *Journal of Machine Learning Research*, vol. 5, pp. 255–291, 2004.

[4] H. H. Aghdam, A. Gonzalez-garcia, J. V. D. Weijer, and M. L. Antonio, "Active Learning for Deep Detection Neural Networks," in *ICCV*, no. Cvc, 2019, pp. 3672–3680.

[5] Y. Geifman and R. El-Yaniv, "Deep Active Learning over the Long Tail," *ArXiv 1711.00941*, no. m, pp. 1–10, 2017. [Online]. Available: http://arxiv.org/abs/1711.00941

[6] T. M. Hospedales, S. Gong, and T. Xiang, "Finding rare classes: Active learning with generative and discriminative models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 2, pp. 374–386, 2013.

[7] C. Campbell, N. Cristianni, and A. J. Smola, "Query Learning with Large Margin Classifiers," in *Proc. IntâĂŹl Conf. Machine Learning*, 2000.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9905 LNCS, pp. 21–37, 2016.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Microsoft Research Asia, Tech. Rep., 2015.

[10] Jia Deng, Wei Dong, R. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009.

[11] L. N. Smith, "Cyclical learning rates for training neural networks," *Proceedings - 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017*, no. April, pp. 464–472, 2017.

[12] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Large Scale Learning of General Visual Representations for Transfer," *ArXiv:1912.11370v1*, 2019. [Online]. Available: http://arxiv.org/abs/1912.11370

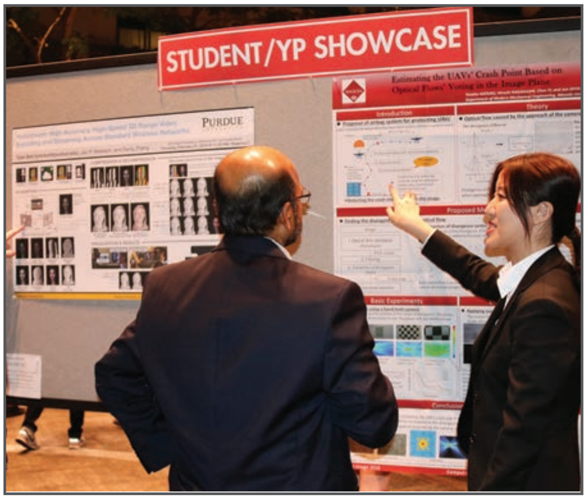
Author Biography

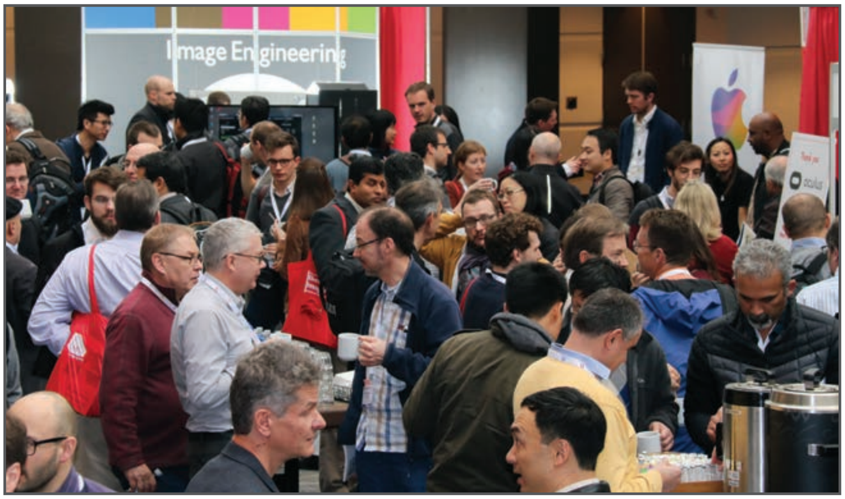
Sander Klomp received both his BSc and MSc degrees from the Eindhoven University of Technology (2016,2018) with the designation Cum Laude. He is now pursuing a PhD degree at TU/e in collaboration with ViNotion, with a focus on efficient deep learning algorithms.

Guido Brouwers received his MSc Electrical Engineering degree from the Eindhoven University of Technology (2015). He is now employed as a research and development engineer at ViNotion where he focuses on deep learning algorithms.

Rob Wijnhoven graduated in Electrical Engineering from the Eindhoven University of Technology in 2004. In 2013, he obtained his PhD in object categorization and detection. He is CTO at ViNotion and leading the research group in computer vision.

Peter H.N. de With is Full Professor of the Video Coding and Architectures group in the Department of Electrical Engineering at Eindhoven University of Technology. He is an IEEE Fellow, has (co-)authored over 400 papers on video coding, analysis, architectures, and 3D processing and has received multiple papers awards. He is a program committee member of the IEEE CES and ICIP and holds some 30 patents.





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!