Evaluating the Performance of Convolutional Neural Network for Classifying Equipment on Construction Sites

Mohammad M. Soltani[a], Seyedeh-Forough Karandish[b], Walid Ahmed[c], Zhenhua Zhu[a] and Amin Hammad[b]

[a]Department of Building, Civil and Environmental Engineering, Concordia University, Canada
[b]Concordia Institute for Information Systems Engineering, Concordia University, Canada
[c]indus.ai, Inc., Canada
E-mail: mo_solta@encs.concordia.ca, s_karand@encs.concordia.ca, walid.aly@indus.ai, zhzhu@bcee.concordia.ca, hammad@ciise.concordia.ca

Abstract –

Estimating the productivity of construction operations is one of the most challenging tasks for project managers. Therefore, the construction industry always looks toward new advancements for automating this process. New automated methods for productivity estimation aim to detect the types, locations, and activities of construction equipment based on sensory data. Computer Vision (CV) is one of the most promising automated methods and it provides an affordable opportunity for estimating the productivity since it only requires regular surveillance cameras for data collection, which are available on many construction sites. One of the widely used CV methods for classifying equipment is Histogram of Oriented Gradient (HOG). Additionally, Bag of Words (BoWs) and Local Binary Pattern (LBP) are other types of descriptors widely used for the object classification. However, these methods reduce the dimensions of the image features to train the classifiers for object detection, which may reduce the reliability of the results. Convolutional Neural Networks (CNN), which are a special type of Artificial Neural Networks (ANN) with deeper layer structure, provide a better approach for object detection compared to the conventional methods due to their deeper understanding of the object features. Furthermore, the advancements in Graphical Processing Units (GPU) made this computationally heavy method more applicable in practice. This paper aims to evaluate the performance of CNN for detecting equipment on construction sites. Several configurations of CNN are trained for detecting multiple equipment (i.e. dump trucks, excavators and loaders). The results of these configurations are compared with those of conventional detectors.

Keywords –

Computer Vision, Equipment detection, Convolutional Neural Networks, Sensory data

1 Introduction

Estimating the productivity of construction operations is one of the main concerns of the contractors since it is tied to the schedule and cost of the projects. The construction industry, with the annual revenues of more than $110 billion in Canada (Statistics Canada, 2016), is always eager to apply new methods and technologies for making the projects more cost-effective while increasing the safety on the sites. Currently, using the Global Positioning System (GPS) is the common practice in monitoring the location of the equipment (e.g. dump trucks, loaders, and excavators.). However, it is difficult to install a GPS receiver on each piece of equipment. On the other side, the availability of the surveillance cameras on many construction sites opens the opportunity for applying Computer Vision (CV) based methods to monitor the productivity of the equipment in addition to monitoring the safety and security of the sites.

Many CV-based methods have been proposed for monitoring the construction equipment, and each method has its benefits and drawbacks. Convolutional Neural Network (CNN) based methods are an emerging practice within the domain of CV. Thanks to the fast-growing advancement of Graphical Processing Units (GPUs), the applications of CNN are increasing within different engineering domains.

This paper aims to apply two comparative analyses. The first analysis compares several conventional CV-based classification methods with CNN-based methods for the classification of dump trucks, loaders, and excavators. In the second analysis, two CNN-based methods are compared using different training and

testing datasets. These comparisons aim to provide a better understanding of the opportunities of using the CNN-based methods in the context of construction operations.

2 Literature Review

The applications of CV are growing fast for monitoring construction projects including construction equipment classification, detection and tracking. Among these applications, equipment classification is an important task for monitoring the resources on construction sites since tracking without recognizing the type of the tracked object does not provide enough information for decision making. Moreover, the main step in CV-based object detectors is the classification. Whether using background subtraction methods or sliding windows, it is mandatory to classify the foreground segments of the images or to classify each sliding window passing over an image to localize the target object within the image or video frame. The classifier can be either binary for each object or a multi-class classifier.

2.1 CV-based Methods in Construction

There are many methods developed for the recognition of workers and construction equipment. Zou and Kim (2007) proposed using Hue, Saturation, and Value (HSV) color space to track the equipment. Detecting the workers through the color of their hardhats was studied by Weerasinghe and Ruwanpura (2009). Azar and McCabe (2012b) investigated the applicability of fusing HOG and Haar-Like features for detecting the equipment. Moreover, they proposed applying a part-based model for detecting excavators (Azar & McCabe, 2012a). The idea of part-based models originally came from the research of Felzenszwalb et al. (2010). Memarzadeh et al. (2013) showed that integrating HOG features with Histogram of Color (HOC) improved the detection accuracy compared to relying only on HOG features. Soltani et al. (2016) proposed using a large dataset of the synthetic images for training HOG detectors.

On the other hand, there are a number of studies that focused more on the equipment classification. Basically, the moving objects on the videos are separated in each frame using background subtraction methods. Then the blobs or foregrounds are fed to the classifiers to determine the class that the foreground belongs to. Azar and McCabe (2011) trained eight classifiers from different angles around a dump truck using HOG features. The moving objects in the video were segmented after subtracting the background. They fed each segment to their classifiers to find whether it is a dump truck or not.

Chi and Caldas (2011) selected four types of feature for classification, such as aspect ratio, height-normalized area size, percentage of occupancy of the bounding box, and average gray-scaled color of the area. They compared the normal Bayes and neural network classifiers after training the aforementioned features for each classifier.

Park and Brilakis (2012a) proposed a two-stage classification approach after subtracting the background. In the first stage, they applied the classifiers based on Haar-Like features trained with Adaboosting. In the second step, they applied a Support Vector Machine (SVM) based classifier trained by Eigen-images on the candidate identified in the previous classification step. In another study, Park and Brilakis (2012b) trained HOG features using a SVM classifier to find moving people on the site. Then they applied *k*-NN classifier trained by color histogram to differentiate workers from other people.

Unfortunately, there is no benchmark dataset similar to CIFAR 10 or 100 (Krizhevsky, 2009), ImageNet (Deng, et al., 2009), or Caltech 101 or 256 (Fei-Fei.et al., 2004) within the construction domain to evaluate and compare the CV-based methods for the construction applications in a fair manner.

2.2 Recent Progress of CV-based Methods

While many industries are adopting CV-based methods for their specialized applications, these methods are advancing rapidly. Therefore, it is necessary for the researchers in the construction domain to follow the recent progress in these methods and to take advantage of them.

2.2.1 Deep Learners for CV

According to the review done by Schmidhuber (2014), excellent results were achieved in image classification using deep learning methods. Guo et al. (2015) represents these methods under four categories: CNN-based, Restricted Boltzmann Machines (RBM) based, Autoencoder-based, and Sparse Coding-based Methods. CNN have a hierarchy of convolutional layers, pooling layers, and fully connected layers. Filters or kernels play the main role by convolving the whole image to generate the feature maps. Pooling maps reduce the dimensions of the feature maps, and fully-connected layers convert the 2D feature maps to 1D feature vector. RBMs are generative stochastic neural networks (Hinton & Sejnowski, 1986) that apply restrictions to make the training algorithms more efficient. Autoencoders are another category within the deep learning domain, and they learn the efficient encodings for the imput data. These encodings can be

further used for reducing the dimentionality of the data (Hinton & Salakhutdinov, 2006). Finally, Sparse Coding is the learning process of an over-complete set of basic functions for describing the input data (Olshausen & Field, 1997).

Moving to the applications of the deep learning in the CV domain, the aforementioned methods, especially CNN methods, can be used under five categories defined by Guo et al. (2015) including: (1) Image Classification, (2) Object Detection, (3) Image Retrieval, (4), Semantic Segmentation, and (5) Human Pose Estimation. Since this research aims to explore the capabilities of deep learning for classifying construction equipment, the literature regarding the CNN is discussed.

Among enormous and fast growing number of deep neural network models, there are couple of models that outperformed the others over the last five years (Canziani et al., 2016). These widely used models include, but are not limited to, AlexNet (Krizhevsky et al., 2012), Network In Network (NIN) proposed by Lin et al. (2013), GoogLeNet (Szegedy et al., 2015), ResNet-18, -34, -50, and -101 (He et al., 2016), VGG-16 and -19 (Simonyan & Zisserman, 2014), and Inception-v3 (Szegedy et al., 2016).

2.2.2 Architecture and Parameters Analysis of CNN for CV

Canziani et al. (2016) analysed and compared the aforementioned models not only from the accuracy point of view, but also considering the computation cost and efficiency of the models. As shown in Figure 1, Inception-v3 has the highest top-1 accuracy. Selecting the the best architecture only considering the accuracy may ignore the applicability of a model by not considering its computation load. For instance, although VGG-19 provides an accuracy of higher than 70%, its operations required for a single forward pass are dramatically higher than ResNet-18 or -34, which have similar accuracies to VGG. Therefore, Canziani et al. (2016) proposed comparing the models based on their information density (accuracy per parameters). This metric provides the capacity of a specific architecture to better utilise its parametric space.

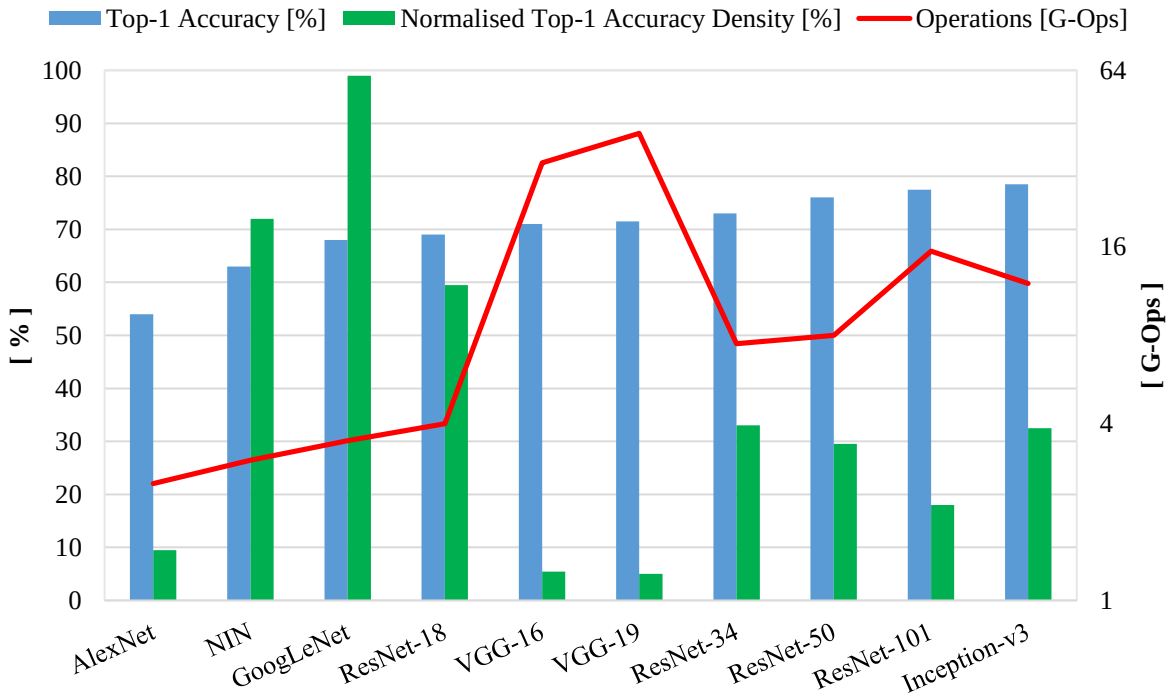

Figure 1 Current deep neural network models comparison, adapted from (Canziani et al., 2016)

Finally, they came up with the conclusion that GoogLeNet has the best architecture in terms of parameters' space utilisation.

Reviewing the recommendtions regarding the design of CNN for CV applications shows that when the network is trained using grayscale images, it can focus more on the shape (i.e. corners and edges) rather than the color. Also, it can be interpreted that if the testing

images include variations of the color, the training images should also include the corresponding color variations (Zheng et al., 2015).

Mishkin et al. (2016) studied the impact of a range of choices for activation function, pooling, learning rate policy, image pre-processing (i.e. different color spaces and grayscale images), batch normalization, classifier design, batch size and learning rate, network width (e.g. changing the number of filters), dataset size and noisy labels, and bias in convolution layers. The bottom line of their research was the suggestion to use Red, Green, Blue (RGB) color space, the linear learning rate decay policy, and mini-batch size around 128 or 256. Also, they stated that having a larger dataset increases the accuracy but it also increases the computation load dramatically. Furthermore, the cleanliness of the data seems to be more important than the size of the dataset.

While collecting a large dataset and labeling the images are time-consuming, Peng et al. (2015) proposed generating the synthetic data from 3D models for training the deep learners. They had a prior study on the applicability of synthetic images for classification based on HOG features using SVM (Sun & Saenko, 2014). They explored the minimum Average Precision (mAP) of the CNN-based classifiers while using real images, virtual images with uniform gray texture (V-GRAY), and virtual images with real texture (V-TX). The results show that V-TX had the best performance while V-GRAY was the worst one.

3 Scope Definition of Analyses

One of the goals of this paper is to compare the performance of the conventional handmade features such as Bags of Words (BoW), Local Binary Pattern (LBP), and HOG with CNN-based features. For classifying the conventional features, SVM is used while for CNN-based features both SVM and ANN are applied. Three widely used equipment on construction sites are excavators, loaders, and dump trucks, which are the focus of this study.

The first descriptor is used for the comparative analysis is BoW or bag of keypoints that use the vector quantization of affine invariant descriptors of image patches. In the original study, Naïve Bayes and SVM classifiers were used because of their simplicity (Csurka et al., 2004); however, this paper only implements SVM for BoW and all other conventional descriptors. The second descriptor, LBP, recognizes certain local binary patterns, known as uniform. The occurrence histogram of these patterns is supposed to be very strong in terms of texture feature (Ojala et al., 2002). HOG is the next descriptor that uses locally normalized histogram of gradient orientations features in a dense overlapping grid (Dalal & Triggs, 2005).

Two CNNs, AlexNet and VGG-f (Chatfield et al., 2014) are considered in the first analysis, because of their relatively simple architucture compared to the very deep and complex networks such as Inception-v3 or ResNet-101. AlexNet is used in this study folowing two approaches. In the first approach, one of the layers of the pretrained AlexNet is used for extracting the features of the training dataset and then the features are fed to SVM for training the classifiers and it is shown as AlexNet-SVM in Table 2. In the second approach, the architecture of AlexNet is used for creating a new network but for classifying three classes instead of its original 1000 classes. The second approach is also repeated for VGG-f to train a network with the similar architecture but for three classes.

Moreover, the performance of the construction equipment classification is evaluated within the domain of CNN. Various configurations and datasets are used to reach to a conclusion for classifying the construction equipment. Multiple training and testing datasets of real, synthetic, and mixed real-synthetic images are used with different numbers of images in each dataset. AlexNet is used independently and in collaboration with SVM as described above.

4 Implementation and Comparative Results

The implementation was done in Matlab 9.1 (Mathworks, 2016) on a mobile station with Intel i7 Quad-Core processor, 32 gigabytes Random-Access Memory (RAM), and NVIDIA Quadro K2000M graphics card. Five datasets were used for both training and testing phases as shown in Table 1.

4.1 Conventional versus CNN-based Methods

In this test, five main types of classifiers were investigated: BoW, LBP, and HOG integrated with SVM as the conventional methods and AlexNet integrated with SVM, independent AlexNet, and independent VGG-f as CNN-based methods. As hown in Table 2, one configuration with 500 clusters was used for extracting BoW features while four cell sizes for LBP features and six cell sizes for HOG features were tested. Three layers were selected from AlexNet considering the available computation resources in this research. Two layers were close to the end of the network and one layer was near the starting border of the network. Selecting the features from both end of the network helps to compare the effectiveness of the feature at different level of the CNN architecture. The features obtained from three different layers of the original AlexNet network were used separately for training the SVM classifiers. Due to the limitation of the

computation capacity, only the first convolutional layer was tested while the other two layers were from the fully connected layers at the end of the network structure. The last two classifiers had similar structure to AlexNet and VGG-f, respectively, except the last layer (classification layer), which has three classes instead of 1000 classes in the original model. Also, in another scenario, AlexNet and VGG-f were fed and tested by gray scale images. For training each of the aforementioned classifier, the datasets of synthetic images with 6,000 images (first row) shown in Table 1 were used while for testing the dataset of real images was used with 1,169 images was used (second row).

Due to space limitation, only the accuracies of correct detections for each class from the confusion matrix are shown in Table 2 in addition to the average accuracy of the all classes. The averages accuracy of BoW was 38%, while the best achieved average accuracies for LBP and HOG were 41% and 47%, respectively. The results show that all the conventional classifiers had more difficulties for classifying the loaders. Investigating the poor accuracies for the loader shows that the 3D model of the loader used for creating the synthetic images looks similar to a brand new loader which has usually a yellow bucket. However, the bucket of the loader looks very dark close to black or dark brown color and the bucket of loader includes a large portion of its appearance. Opposite to HOG-based method, it is highly possible that CNN can be sensitive to the color of the object during training phase. On the other hand, the best average accuracies of AlexNet-SVM, AlexNet, and VGG-f were 83%, 78%, and 68%, respectively. It can be concluded from this analysis that CNN-based methods outperformed the conventional methods.

Table 1 Image Datasets Specifications

		Excavator	Loader	Truck	Total
Synthetic		2,000	2,000	2,000	6,000
Real		555	267	347	1,169
Real		100	100	100	300
Mixed	**Real**	555	267	347	2,338
	Synthetic	555	267	347	
Mixed	**Real**	100	100	100	600
	Synthetic	100	100	100	

4.2 Performance Evaluation of CNN-based Method on Different Datasets

At this stage, the best two groups of classfiers (trained for three classes of equipment) resulting from the previous test in Section 4.1 were used for a more detailed analysis and testing. The two groups are the AlexNet-SVM classifiers with ‘fc7’ layer and the independent AlexNet classifiers.

The purpose of this test is to find how the accuracy of each classifier can be be affected using the five datasets shown in Table 1. There are two different datasets of real images, one with 1,169 images and the other with a small number of images (300 in total) to find the impact of the number of the training images. Also, two more datasets are prepared by adding the synthetic images with the same number of the real images in each of two previous datasets. The advantage of the synthetic images is that they are taken from all views around the equipment while the real images are mostly captured from low heights and views. A sample of real and synthetic images are shown in Figure 2 One the other hand, the synthetic images may look artificial and that may have negative effects on the CNN-based classifiers compared to conventional methods that did not show such effects on the accuracy (Soltani et al., 2016). Therefore, using the mixed datasets can evaluate the generality of the classifiers on different brands, colors, and views of the equipment.

Twenty scenarios are created and tested, which are shown in Table 3. Starting with AlexNet-SVM classifiers, the results of the classifiers trained by synthetic images and tested on real images are 83% and 86%. The accuracy achieved by classifying the dataset with smaller number of images shows better performance as expected. In the next scenarios, the larger dataset of the real images was used for training and the smaller datasets of real and mixed images were classified. The results show that adding synthetic images from various views reduces the accuracy. By using the two larger datasets of mixed images for the training, the accuracies were improved for both previousely tested datasets. In the next comparison, the smaller dataset of real images was used to mimic the situations where there is a very limited number of available training images. This classifier is applied on two larger datasets of real and mixed images and the accuracies dropped significantly. In the last two scenarios, the smaller dataset of the mixed images was used for training and applied on the larger datasets of real and mixed images. It is clear from this test that we can improve the quality of training (and the resulting accuracy) by adding synthetic images to the training dataset when the number of real images is small.

In the next test, similar comparisons were repeated for evaluating the independent AlexNet CNN. The trend of the results is close to the previouse test except in the scenario where the larger dataset of real images is used for training and applied on the smaller datasets of real and mixed images. Opposite to the AlexNet-SVM classifier, the independent AlexNet outperfomed on the smaller dataset of the mixed images compared to the smaller dataset of the realimages. Additionally, using

the larger dataset of mixed images shows its maximum performance on the smaller datasets of the real and mixed images. However, achieving the accuracy of 100% does not mean that this classifier is able to reach the same accuracy on the larger datasets and it simply proves that adding synthetic images to real images during the training can improve the performance of the AlexNet classifier.

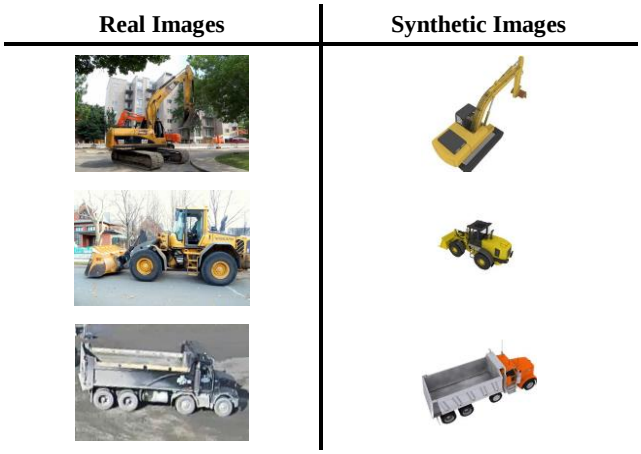

Figure 2 Sample real and synthetic images

5 Conclusion

Reviewing the literature showed that while the applications of CNN-based methods are growing very fast, the construction industry is lagging in taking advantage of these methods. Therefore, this paper investigated the applicability of CNN-based methods while comparing them with the conventional descriptors. In the first analysis, BoW, LBP, HOG, AlexNet independently and integrated with SVM, and VGG-f networks were investigated. CNN-based methods clearly outperformed the conventional methods. In the second test, AlexNet and AlexNet-SVM were compared using different image datasets including real, synthetic, and mixture of real and synthetic images with different number of images in the training and testing datasets. The results show that it is better to add synthetic images to the real images for the training of the classifiers, especially when there is limited number of available real images. Including the synthetic images to the real images helps the classifier consider various views of the target objects which are not available in the real image dataset.

On the other hand, training and applying CNN-based requires a very powerful computer configuration, especially when developing a very deep network is required. It is recommended to include more classes of the equipment in the future. Moreover, investigating the performance of CNN-based object detectors for localizing the equipment is the next step.

References

[1] J. Zou and H. Kim, "Using hue, saturation, and value color space for hydraulic excavator idle time analysis," *Journal of computing in civil engineering*, vol. 21, no. 4, pp. 238-246, 2007.

[2] I. T. Weerasinghe and J. Y. Ruwanpura, "Automated data acquisition system to assess construction worker performance," in *In Building a Sustainable Future, Construction Research Congress*, Seattle, WA, USA, 2009.

[3] E. R. Azar and B. McCabe, "Vision-based Recognition of Dirt Loading Cycles in Construction Sites," in *Construction Research Congress*, West Lafayette, IN, USA, 2012.

[4] E. R. Azar and B. McCabe, "Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos," *Automation in construction*, vol. 24, pp. 194-202, 2012.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, 2010.

[6] M. Memarzadeh, M. Golparvar-Fard and J. C. Niebles, "Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors," *Automation in Construction*, vol. 32, pp. 24-37, 2013.

[7] M. M. Soltani, Z. Zhu and A. Hammad, "Automated annotation for visual recognition of construction resources using synthetic images," *Automation in Construction*, vol. 62, pp. 14-23, 2016.

[8] E. R. Azar and B. McCabe, "Automated visual recognition of dump trucks in construction videos," *Journal of Computing in Civil Engineering*, vol. 26, no. 6, pp. 769-781, 2011.

[9] M. W. Park and I. Brilakis, "Construction Worker Detection in Video Frames for Initializing Vision Trackers," *Automation in Construction*, vol. 28, pp. 15-25, 2012b.

[10] M. W. Park and I. Brilakis, "Enhancement of construction equipment detection in video frames by combining with tracking," in *International Workshop on Computing in Civil Engineering*, 2012a.

[11] A. Krizhevsky, "The CIFAR-10 dataset," 2009.

[Online]. Available: https://www.cs.toronto.edu/~kriz/cifar.html. [Accessed 03 February 2017].

[12] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009*, 2009.

[13] L. Fei-Fei, R. Fergus and P. Perona., "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[14] G. E. Hinton and T. J. Sejnowski, "Learning and releaming in Boltzmann machines," in *Parallel Distrilmted Processing*, 1986, p. 4.2.

[15] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504-507, 2006.

[16] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision research*, vol. 37, no. 23, pp. 3311-3325, 1997.

[17] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu and M. S. Lew, "Deep learning for visual understanding: A review," *Neurocomputing*, vol. 187, pp. 27-48, 2015.

[18] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85-117, 2014.

[19] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012.

[20] M. Lin, Chen, Q. and S. Yan, "Network in network," *ArXiv e-prints*, 2013.

[21] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[22] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv e-prints*, 2014.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[25] A. Canziani, A. Paszke and E. Culurciello, "An Analysis of Deep Neural Network Models for Practical Applications," *ArXiv e-prints*, 2016.

[26] Z. Zheng, Z. Li, A. Nagar and W. Kang, "Compact deep convolutional neural networks for image classification," in *International Conference on Multimedia and Expo (ICME 2015)*, Torino, Italy, 2015.

[27] D. Mishkin, N. Sergievskiy and J. Matas, "Systematic evaluation of CNN advances on the ImageNet," *ArXiv e-prints*, 2016.

[28] X. Peng, B. Sun, K. Ali and K. Saenko, "Learning deep object detectors from 3D models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[29] B. Sun and K. Saenko, "From Virtual to Reality: Fast Adaptation of Virtual Object Detectors to Real Domains," *BMVC*, vol. 1, no. 2, p. 3, 2014.

[30] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on statistical learning in computer vision, ECCV*, 2004.

[31] T. Ojala, M. Pietikainen and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971-987, 2002.

[32] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005.

[33] K. Chatfield, K. Simonyan, A. Vedaldi and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *ArXiv e-prints*, 2014.

[34] Mathworks, "MATLAB," 2016. [Online]. Available: http://www.mathworks.com/. [Accessed 21 December 2016].

[35] Statistics Canada, "Construction (NAICS 23) : Gross domestic product (GDP)," 2016. [Online]. Available: https://www.ic.gc.ca/app/scr/sbms/sbb/cis/gdp.html?code=23&lang=eng#gdp2a. [Accessed 05 January 2016].

[36] S. Chi and C. H. Caldas, "Automated object identification using optical video cameras on construction sites," *Computer‐Aided Civil and Infrastructure Engineering*, vol. 26, no. 5, pp. 368-380, 2011.

Table 2 Results of comparative analysis between conventional and CNN-based classifiers

Method	Size of Images	Additional Information	Accuracy (%)			
			Excavator	Loader	Truck	Average
Bag of Words	128×128	500 Clusters	78	3	33	38
LBP-SVM	128×128	4×4 Cells	72	3	39	38
	128×128	8×8 Cells	72	5	38	38
	128×128	16×16 Cells	76	5	41	41
	128×128	32×32 Cells	51	10	56	39
HOG-SVM	128×128	2×2 Cells	69	1	35	35
	128×128	4×4 Cells	70	1	39	37
	128×128	8×8 Cells	75	3	43	40
	128×128	16×16 Cells	73	5	61	46
	128×128	32×32 Cells	55	21	65	47
	128×128	64×64 Cells	41	45	35	40
AlexNet-SVM	227×227	conv 1 Layer	51	42	60	51
	227×227	fc7 Layer	76	78	93	83
	227×227	fc8 Layer	72	60	96	76
AlexNet	227×227	Colored	74	48	99	74
	227×227	Grayscale	95	60	78	78
VGG-f	224×224	Colored	70	40	94	68
	224×224	Grayscale	92	10	52	51

Table 3 Results of comparative analysis on AlexNet-SVM and AlexNet classifiers

Method	Training Images		Testing Images		Accuracy (%)			
	Type	Number	Type	Number	Excavator	Loader	Truck	Average
AlexNet-SVM	Synthetic	6,000	Real	1,169	76	78	93	83
	Synthetic	6,000	Real	300	95	83	79	86
	Real	1,169	Real	300	89	100	95	94
	Real	1,169	Mixed	600	94	88	93	91
	Mixed	2,338	Real	300	94	100	96	97
	Mixed	2,338	Mixed	600	97	100	98	98
	Real	300	Real	1,169	54	66	100	73
	Real	300	Mixed	2,338	78	69	97	81
	Mixed	600	Real	1,169	83	90	99	90
	Mixed	600	Mixed	2,338	91	95	99	95
AlexNet	Synthetic	6,000	Real	1,169	74	48	99	74
	Synthetic	6,000	Real	300	95	51	97	81
	Real	1,169	Real	300	88	100	95	94
	Real	1,169	Mixed	600	99	94	97	97
	Mixed	2,338	Real	300	100	100	100	100
	Mixed	2,338	Mixed	600	100	100	100	100
	Mixed	600	Real	1,169	77	88	99	88
	Mixed	600	Mixed	2,338	88	94	100	94
	Real	300	Real	1,169	81	67	100	83
	Real	300	Mixed	2,338	100	71	96	89