Kim H., Kim H., Hong Y. W., and Byun H. (2017). "Detection of construction entities using deep convolutional networks." In: *LC3 2017: Volume I – Proceedings of the Joint Conference on Computing in Construction (JC3)*, July 4-7, 2017, Heraklion, Greece, pp. 517-524. DOI: https://doi.org/10.24928/JC3-2017/0335.

DETECTION OF CONSTRUCTION EQUIPMENT USING DEEP CONVOLUTIONAL NETWORKS

Hongjo Kim[1], Hyoungkwan Kim[2], Yong Won Hong[3], and Hyeran Byun[4]

Abstract: Vision-based monitoring methods have been investigated for understanding construction site contexts. However, detection capabilities of such methods are still insufficient to be utilized in general construction sites due to dynamic outdoor conditions and appearance variances of construction entities. To improve performance of a construction entity detector, we propose a detection method using a region-based fully convolutional network (R-FCN). R-FCN consists of two main parts, which are a fully convolutional network and a region proposal network. The fully convolutional network extracts hierarchical object features through a supervised learning process, while a region proposal network generates a set of object candidate regions in an image to localize target objects. To evaluate the generalization performance of the detection method, a benchmark dataset is collected from ImageNet for five classes (dump truck, excavator, loader, concrete mixer truck, and road roller), having various object appearances within a class in different backgrounds. A state-of-the-art performance, mean average precision of 95.61%, was recorded from the experiment. The proposed method shows a potential for the universal detector that can detect construction equipment on every construction site.

Keywords: construction site monitoring, object detection, convolutional networks, benchmark dataset.

1 INTRODUCTION

For successful execution of construction projects, it is necessary to observe the progress of the work performed in the field and to measure the quality of the results (Fathi et al. 2015; Son et al. 2015; Yang et al. 2015). Vision-based monitoring systems in the previous literatures have been developed for the field observations and quality measurements, but still lacks performance in the most basic requirement of object detection capabilities. In computer vision technology, object detection refers to locating objects in an image and classifying them simultaneously. If the reliability of the detection result is low, it is impossible to recognize the exact situation. Thus, significant improvement of detection capability is crucial for any vision-based monitoring system to be truly applicable to construction projects.

Research efforts have been made for vision-based object detection (Chi and Caldas 2012; Gong and Caldas 2011; Kim et al. 2016a; b; Memarzadeh et al. 2013; Park et al. 2015;

[1] PhD candidate, School of Civil and Environmental Engineering, Research Assistant of the Advanced Infrastructure Management Group, Yonsei University, Seoul, Korea, hongjo@yonsei.ac.kr

[2] Professor, School of Civil and Environmental Engineering, Leader of the Advanced Infrastructure Management Group, Yonsei University, Seoul, Korea, hyoungkwan@yonsei.ac.kr (corresponding author)

[3] PhD student, Department of Computer Science, Research Assistant of the Computer Vision and Pattern Recognition Laboratory, Yonsei University, Seoul, Korea, yhong@yonsei.ac.kr

[4] Professor, Department of Computer Science, Leader of the Computer Vision and Pattern Recognition Laboratory, Yonsei University, Seoul, Korea, hrbyun@yonsei.ac.kr

Rezazadeh Azar and McCabe 2012). Most successful detection algorithms among those are based on supervised learning. Supervised learning refers to a method of learning a model in which algorithms such as artificial neural networks predict appropriate output values for input values through training data (data having true output values for input values). In the object detection task, the input value is an image and the output value is the location and type of the object in the image. However, at present, these methods in the previous studies have not widely been used in actual construction sites, and the reason can be the followings. First object detection algorithms are optimized for specific datasets and environments, resulting in significant performance degradation for other construction sites. That is, the training data contains only the limited appearance of specific objects, or the number of the object samples is insufficient. Second, the algorithm does not extract good patterns from the data. In other words, the model must predict the appropriate output value through the input value. However, since the model obtained through the algorithm does not learn the important characteristics of the input value, it makes an incorrect prediction about the new input value.

To address the issues, we propose a detection method using a region-based fully convolutional network (R-FCN; Dai et al. 2016). The R-FCN consists of a feature extraction part and a region proposal network that extracts candidate regions that are likely to have objects in the image, as shown in Fig. 1. A new benchmark data collected from ImageNet (Russakovsky et al. 2015) for five classes was used to validate and verify the method. The contributions of this paper are threefold:

- We proposed a deep CNN-based method to detect construction equipment with a high level of performance.
- Despite a small amount of publicly available image data in the construction industry, we have successfully trained the deep convolutional network by using transfer learning.
- A benchmark dataset that can verify the generalization capability of a model was made to judge the performance of the algorithm on the change of the appearance of the construction objects and the background change.

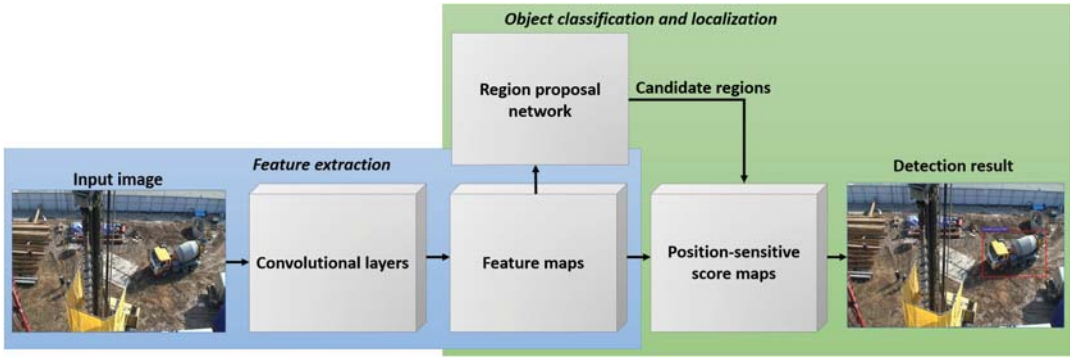

Figure 1: Proposed construction object detection pipeline using R-FCN

2 Related Works

2.1 Construction Site Monitoring

Many studies have been conducted to recognize objects (worker, equipment, materials, etc.) in construction sites (Chi and Caldas 2012; Dimitrov and Golparvar-Fard 2014; Gong and Caldas 2011; Memarzadeh et al. 2013; Park et al. 2015; Rezazadeh Azar and McCabe 2012;

Soltani et al. 2016). Conventional recognition methods were intended to extract hand-engineered features such as histograms of oriented gradients (HOG) and scale invariant feature transform (SIFT). These features are known to have limitations in representing objects. In recent years, CNN-based methods have been used to automatically extract features during the training process, which show a-state-of-the-art performances in many visual recognition tasks (LeCun et al. 2015). However, those have not actively been investigated in the construction industry. This paper intends to address the potential of deep CNNs in their construction applications.

3 R-FCN for Construction Object Detection

R-FCN is a kind of convolutional networks (CNNs) for detecting objects in images (Dai et al. 2016). In a layer of CNN, input image pixels are convolved (element-wise multiplication) by a certain size of a convolutional filter, and then the output is passed through a non-linear function such as a rectified linear unit to obtain the elements constituting a feature map of that layer. Many convolution filters exist for each layer, and the hierarchical features of objects are learned through weight updating processes such as the gradient descent. Convolution filters composed of weights represent the characteristics of objects after training.

The more an image passes through the layers of CNN, the larger scale feature can be extracted. However, as the number of layers increases, the accuracy becomes saturated. To prevent this, residual learning can be implemented to keep the identity of the input image at each layer (He et al. 2016). With this method, it is possible to prevent the distortion of the original image information from becoming worse as the layer becomes deeper.

R-FCN is a fully convolutional network that has no fully-connected layer, and convolution filters of all layers can learn by simple end-to-end training. After creating the feature maps through the final convolutional layer, the region proposal network finds candidate regions that are likely to have target objects (Dai et al. 2016; Ren et al. 2015). A position-sensitive score map is generated on the candidate regions to determine whether there is a class that the model wants to find.

4 Experiment

4.1 Experimental setting

The R-FCN code was used in this work (Dai et al. 2016). As a feature extractor, a deep residual network with 50 layers (ResNet-50) proposed by He et al. (2016) was used. The region proposal network generated 300 proposals at maximum per image. For training the R-FCN model, stochastic gradient descent was used to update the weights of the model. During 100,000 iterations of training, we used a learning rate 0.001, a weight decay 0.0005, a momentum 0.9 in the loss function. The learning rate controls the magnitude of the weight change by the gradients, and the weight decay determines the impact of the regularization term that limits the weights to a certain range. To prevent the oscillation of the weight values, the momentum term controls the magnitude of the weight update by considering the previous weight update. After 80,000 iterations, the learning rate dropped to 0.0001.

A benchmark dataset (Table 1) for construction objects was collected from ImageNet (Russakovsky et al. 2015); the dataset was categorized into five classes, which are a dump

truck, excavator, loader, concrete mixer truck, and road roller. Each class sample was divided into training, validation, and test sets, by the portion of 60%, 20%, and 20%, respectively. The training and validation sets were used in training the model, and the test set was used only for evaluation.

Table 1: A benchmark dataset statistic

Class	Count	Source
dump truck	760	ImageNet
excavator	361	ImageNet
loader	787	ImageNet
concrete mixer truck	659	ImageNet
road roller	353	ImageNet

4.2 Experimental Results

We can evaluate the performance of the detection model by whether the model captures the region with a bounding box and correctly classifies it as the actual class in the test image. The model determines the class of the candidate region when the score value of a particular class exceeds a threshold value. Mean average precision (mAP) was used as an evaluation criterion for object detection following the PASCAL Visual Object Class Challenge (Everingham et al. 2015). Mean average precision is a mean of average precision (AP) of each class. The AP summarizes the precision/recall curve by averaging precision values over equally spaced recall values. This performance index, mAP, enables the overall performance evaluation of the model rather than presenting a single precision and recall values. This is because the precision and recall values are in a trade-off relationship when changing a threshold value. For example, if a precision value is increased, a recall value is decreased. For bounding box evaluation, if the intersection between the predicted and the ground truth bounding boxes exceeds 50%, then it was regarded as a correct prediction.

Table 2 shows the result of the experiment - the mAP of 95.61%. Compared to the performances of other studies, the result is a success of unprecedented level. Samples of the detection results are shown in Fig. 2.

Table 2: Summary of the experimental results

Category	%
mAP	95.61
AP for dump truck	92.04
AP for excavator	93.94
AP for loader	95.64
AP for concrete mixer truck	98.17
AP for road roller	98.27

4.3 Discussion

The high level of mAP (95.61%) obtained in this study signifies the strong potential of deep CNN for the construction industry. So far, detectors of construction objects have not received strong welcome from the industry because they were not regarded as universal detectors. In other words, those detectors only work for the specific settings. However, the result of our study indicates that the concept of universal construction detectors are likely to be materialized. This argument is more valid considering the fact that the dataset used in this study was from ImageNet; ImageNet is known to be a challenging dataset, which is hard to be understood by an artificial intelligence-based detector.

The samples of the detection failures are shown in Fig. 3. The first row examples of Fig. 3 show that the model misclassified an object as an irrelevant class, and the second row examples show that the model missed a construction object. It is interpreted that the detection failure occurred when the size of the object was small, when an object is truncated at the edge of the image, or when dust noise exists.

Future studies are required to complete a truly universal detector of construction objects. This study is, to our knowledge, the first attempt to propose a deep CNN in the construction object detection. The deep CNN used is in our study is even the state-of-the-art in the computer vision community. However, the dataset we used is still limited in the sense that the total number of images used is 2,920. Increase of the dataset and the number of construction object classes would result in more validation and verification of the proposed method for the construction industry.

Figure 2: Samples of the detection result

Figure 3: Samples of the detection failures

5 Conclusions

We proposed a construction object detector based on deep CNNs. Strengthened by convolutional layers and region proposal network, the deep CNN (R-FCN)-based detector could achieve the high performance of mAP (95.61%). This result is more meaningful considering that the 2,920 images used in this study came from ImageNet that is known to have challenging data.

This study result strongly demonstrated the potential of a universal detector of construction equipment. Once the intelligent detector is developed, it would not require any more training to be applicable to different construction sites. This elevated capability of the detector is expected to facilitate the active introduction of computer-vision based monitoring systems for many applications of construction management.

6 Acknowledgments

We would like to thank Inhae Ha and Seongdeok Bang for helping with image data annotation. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT and Future Planning) (NRF-2014R1A2A1A11052499 and No. 2011-0030040).

7 References

Chi, S. and Caldas, C. H. (2012). Image-Based Safety Assessment: Automated Spatial Safety Risk Identification of Earthmoving and Surface Mining Activities. *Journal of*

Construction Engineering and Management, 138(3), 341-351. doi: 10.1061/(ASCE)CO.1943-7862.0000438

Dai, J., Li, Y., He, K. and Sun, J. (2016). R-FCN: Object detection via region-based fully convolutional networks, *Advances in Information Processing Systems*.

Dimitrov, A. and Golparvar-Fard, M. (2014). Vision-based material recognition for automated monitoring of construction progress and generating building information modeling from unordered site image collections. *Advanced Engineering Informatics*, 28(1), 37-49. doi: 10.1016/j.aei.2013.11.002

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J. and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1), 98-136. doi: 10.1007/s11263-014-0733-5

Fathi, H., Dai, F. and Lourakis, M. (2015). Automated as-built 3D reconstruction of civil infrastructure using computer vision: Achievements, opportunities, and challenges. *Advanced Engineering Informatics*, 29(2), 149-161. doi: 10.1016/j.aei.2015.01.012

Gong, J. and Caldas, C. H. (2011). An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automation in Construction*, 20(8), 1211-1226. doi: 10.1016/j.autcon.2011.05.005

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, The IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778 http://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

Kim, H., Kim, K. and Kim, H. (2016a). Data-driven scene parsing method for recognizing construction site objects in the whole image. *Automation in Construction*, 71, 271-282. doi: 10.1016/j.autcon.2016.08.018

Kim, H., Kim, K. and Kim, H. (2016b). Vision-Based Object-Centric Safety Assessment Using Fuzzy Inference: Monitoring Struck-By Accidents with Moving Objects. *Journal of Computing in Civil Engineering*, 30(4), 04015075. doi: 10.1061/(ASCE)CP.1943-5487.0000562

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444. doi: 10.1038/nature14539

Memarzadeh, M., Golparvar-Fard, M. and Niebles, J. C. (2013) Automated 2D detection of construction equipment and workers from site video streams using histograms of oriented gradients and colors. *Automation in Construction*, 32, 24-37. doi: 10.1016/j.autcon.2012.12.002

Park, M. W., Elsafty, N. and Zhu, Z. (2015). Hardhat-wearing detection for enhancing on-site safety of construction workers. *Journal of Construction Engineering and Management*, 141(9), 04015024. doi: 10.1061/(ASCE)CO.1943-7862.0000974

Ren, S., He, K., Girshick, R. and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in information processing systems*. http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks

Rezazadeh Azar, E. and McCabe, B. (2012). Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos. *Automation in construction*, 24, 194-202. doi: 10.1016/j.autcon.2012.03.003

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L. (2015). ImageNet Large Scale

Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3), 211-252. doi: 10.1007/s11263-015-0816-y

Soltani, M. M., Zhu, Z. and Hammad, A. (2016). Automated annotation for visual recognition of construction resources using synthetic images. *Automation in Construction*, 62, 14-23. doi: 10.1016/j.autcon.2015.10.002

Son, H., Bosché, F. and Kim, C. (2015). As-built data acquisition and its use in production monitoring and automated layout of civil infrastructure: A survey. *Advanced Engineering Informatics*, 29(2), 172-183. doi: 10.1016/j.aei.2015.01.009

Yang, J., Park, M.-W., Vela, P. A. and Golparvar-Fard, M. (2015). Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future. *Advanced Engineering Informatics*, 29(2), 211-224. doi: 10.1016/j.aei.2015.01.011