International Journal of High-Rise Buildings
December 2020, Vol 9, No 4, 335-342
https://doi.org/10.21022/IJHRB.2020.9.4.335

International Journal of
High-Rise Buildings
www.ctbuh-korea.org/ijhrb/index.php

Equipment and Worker Recognition of Construction Site with Vision Feature Detection

Shaowen Qi[1], Jiazeng Shan[2], and Lei Xu[3]

[1]*Department of Civil Engineering, Tongji University, China*
[2]*Department of Civil Engineering, Tongji University, China*
[3]*Shanghai Construction No.1 (Group) Co., Ltd.*

Abstract

This article comes up with a new method which is based on the visual characteristic of the objects and machine learning technology to achieve semi-automated recognition of the personnel, machine & materials of the construction sites. Balancing the real-time performance and accuracy, using Faster RCNN (Faster Region-based Convolutional Neural Networks) with transfer learning method appears to be a rational choice. After fine-tuning an ImageNet pre-trained Faster RCNN and testing with it, the result shows that the precision ratio (mAP) has so far reached 67.62%, while the recall ratio (AR) has reached 56.23%. In other word, this recognizing method has achieved rational performance. Further inference with the video of the construction of Huoshenshan Hospital also indicates preliminary success.

Keywords: Object detection, Construction Site management, Transfer learning, CNN

1. Introduction

1.1. The Fusion between Civil Engineering & Computer Science

Since the end of the last century, the development of computers has evolved rapidly according to Moore's Law, and the trend has been towards miniaturization and more computing power. In this context, computers began to be integrated into various industries. In the civil engineering industry, the most notable results have been the advent of CAD and structural analysis software, which has greatly freed up the productivity of civil engineers, allowing them to focus on designing complex structures more efficiently rather than on drawing and performing complex calculations.

In addition to the widespread application of CAD in civil engineering drafting, the combination of computer and construction industry has been blossoming in many fields, and a large number of scholars have been discussing how to bring the great potential of computer into the civil engineering industry. For example, scholars such as Zhang, Yuyan et al. began to explore the use of real-time video surveillance for intelligent reservoir management and real-time river monitoring as early as 2009. Although the application of cellular network technology was still at the 3G level at that time, scholars have seen the future trend and introduced the concept of using the network for real-time video management, and believe that the unified nation-wide management using network will be the future trend, and in the era when 5G technology is gradually being applied, such a vision has obviously become a reality.

Besides the combined application of network technology, scholars such as Wang, Qiankun et al. discuss the potential of applying the best examples of combining the fields of surveying and geography with computers, the GIS (Geographic Information System), to the construction of crushed rockfill dam projects. In the traditional methods, controlling the crushing parameters and taking samples for inspection are manual, which are not only time-consuming and labor-intensive, but also cannot be monitored in real time. The researcher proposes to use GPS to obtain real-time information on the number of passes, the distribution of elevation and layer thickness after crushing, and to manage the data with a GIS system to facilitate data processing and result accessing.(Wang & Chen, 2009) The application of GPS to the overall automation of the construction process, starting with every detail of the construction, not only saves manpower and material resources and construction time, but also eliminates construction quality problems caused by human manage-ment.

When the academy explores the potential of combining computers and the civil engineering industry, the computer industry is also working hand-in-hand with companies in the civil engineering industry. On December 13, 2017, NVIDIA announced that Komatsu Group, one of the world's largest manufacturers of construction and mining equipment, has entered into a partnership with NVIDIA to deploy artificial intelligence for its construction sites to realize better safety and efficiency. NVIDIA's idea of using

†Corresponding author: Shaowen Qi
Tel: +86-176-1219-7114
E-mail: shaowen.qi@foxmail.com

GPUs in construction site cameras and drones to create 3D visualizations of construction site management fits well with the research philosophy that be put into practice in this article.

1.2. Contradictions between BIM Concepts and Manual Data Collection

The combination of today's computer and construction industry has given rise to a new hotspot, BIM, as an extension of CAD, which integrates all the important data of a project into a single model for unified management, the advancement of the concept is self-evident. The data about personnel, machine & materials in the BIM model needs to be constantly updated during the construction process in order to monitor the whole project. In today's situation where there is no reliable automated solution, using human statistics will consume a lot of manpower. In view of this, a new, reliable, automated personnel, machine & materials detection is one of the necessary prerequisites for the realization of BIM.

1.3. The Development and Application of Machine Learning Concepts

Although it is only in the last decade that the concept of machine learning has been widely known and used in various industries, the original concept dates back to 1959, when IBM scientist Arthur Samuel proposed a way for computers to learn how to play checkers on their own, even to a level beyond its programmer, given only basic rules and parameters.(Samuel, 1959) In 1986, David Rumelhart et al. systematically described the back propagation of errors and its application to neural networks in their article, which adjusts the weights of connections within the network on its own, so that the output is always close to the expected result.(Rumelhart, Hinton, & Williams, 1986) The proposed error back propagation algorithm, the BP algorithm, is still one of the core concepts of machine learning.

Since the development of machine learning, benefiting from the leaps of computer computing power, many different applications have been performed, among which object detection is a widely performed application.(Lv, Zhong, & Zhang, 2018) Although civil engineering, as a traditional industry, is less sensitive to new technologies, it is foreseeable that machine learning has the potential to have a great impact on all aspects of civil engineering, and many scholars have done a lot of research on the application of machine learning in the civil engineering industry. For example, the use of deep learning and reinforcement learning in the conceptual design phase of the structure for topology optimization, to determine the optimal distribution of materials in the design space, (Sigmund & Maute, 2013) as well as machine learning and computer vision-based structural crack detection and monitoring and many other applications.(Lim, La, & Sheng, 2014) The concepts of machine learning and intelligent construction will be studied more deeply and applied more widely in the future.

2. Terminology in Object Detection Task

2.1. The Indicators for Evaluating the Model

For general machine learning tasks, we often consider accuracy and loss, which together indicate how well a machine learning model performs in the task. Loss is a value calculated by the loss function of specific machine learning algorithm. Typically, loss function is combined with a series of functions, such as cross entropy, normalization, etc. The implementations of loss function vary in different algorithms, but they all target at improving the accuracy and recall of the model, while avoiding overfitting. In general, higher accuracy and lower loss hint a better performance. although in some cases very high accuracy and very low loss is only a fake sign of poor results, in fact, there are problems of overfitting or underfitting, but overall the increase or decrease of these two indicators will reflect the model's situation.

However, for the object detection task, due to its unique complexity, which makes it meaningless to discuss the accuracy rate simply without any constraints. Hence another scalar called mAP (mean Average Precision), which depends on a critical threshold, called IoU (Intersection of Union), was made to evaluate the performance. When evaluating an object detection model, it is not enough to evaluate only its accuracy. The model should be able to identify all targets in a given image. Hence, AR (Average Precision) was introduced to describe if all the objects are fully identified. At the same time, the loss value is also one of the important parameters to describe the status of model, which is more abstract than the above-mentioned scalars, but can also reflect the deeper status of the model.

2.2. The Algorithm for Training the Model

2.2.1. Comparison between Different Algorithms

In the recent years when machine learning has flourished, many algorithms with different ideas have been developed for the task goal of object detection. It can basically be divided into two major categories: one-stage and two-stage. Algorithms like YOLO and SSD belong to one-stage algorithm, which directly localize and classify the target on the input image, the advantage is that the training and detecting speed are faster, but the problem is that the algorithm is not sufficiently precise. On the contrary, networks such as RFCN and Faster RCNN belong to the two-stage algorithm, which first extracts the candidate regions in the image, and then carries out the object localization and classification in the candidate regions. Its training and detecting speed are slower than the one-stage algorithm, but the advantage is that the accuracy is higher.(Fan, Zhao, Zhao, Hu, & Wang, 2020) The two-stage Faster RCNN algorithm is used in this study, because

the detection of personnel, machine and materials on the construction site does not require a high real-time performance.

2.2.2. The Path of the Development of Faster RCNN

The development of Faster RCNN has gone through successive iterations, from the first RCNN with high accuracy but slow operation, which has been continuously improved, to the current Faster RCNN, which combines both accuracy and speed.

The RCNN algorithm was originally developed by Girshick et al. in 2014 to combine the Region Proposal Algorithm with CNN for object detection, which greatly improved the efficiency of the RCNN compared to the previous sliding-window approach, which was similar to exhaustive approach,(Girshick, Donahue, Darrell, & Malik, 2014) and then improved its pooling layer again in 2015 to reduce redundant feature extraction operations. Its efficiency was improved to form a new network structure, Fast RCNN,(Girshick, 2015) and then Shaoqing, Ren et al. completely discarded the traditional algorithm for extracting candidate regions in Fast RCNN, Selective Search, which uses a CPU to run, and is clearly a very unreasonable bottleneck compared to the neural network part that runs on a GPU. By using the region extraction network (RPN), the algorithm of extracting candidate regions was integrated into the neural network, which again greatly improved the efficiency and finally resulted in the present Faster RCNN network.(Ren, He, Girshick, & Sun, 2015)

2.2.3. The Architecture of Faster RCNN

The reason why Faster RCNN can achieve the balance between efficiency and accuracy is closely related to its well-designed network architecture. A clear understanding of what happens in the black box of the algorithm during the machine learning process is important for understanding the results of the algorithm and adjusting it accordingly to achieve better results.

As described earlier, Faster RCNN belongs to the two-stage algorithm, which first extracts the candidate regions from the image. This extraction process uses the Region Extraction Network (RPN), which is a fully convolutional neural network, and used to provide candidate regions for subsequent classifier, and which is much faster due to the use of GPU operations. Since both RPN in the first step and the classifier in the second step need to be further computed by the feature maps obtained after the convolutional layer computation, the new algorithm uses a shared convolutional layer, which has two steps of computation, but the first step of the convolutional computation only needs to be done once, thus saving time and performance overhead. This concept is shown in Figure 1. In order to achieve convolutional layer sharing, Faster RCNN also uses methods such as alternate training and approximate joint training, the implementation of which

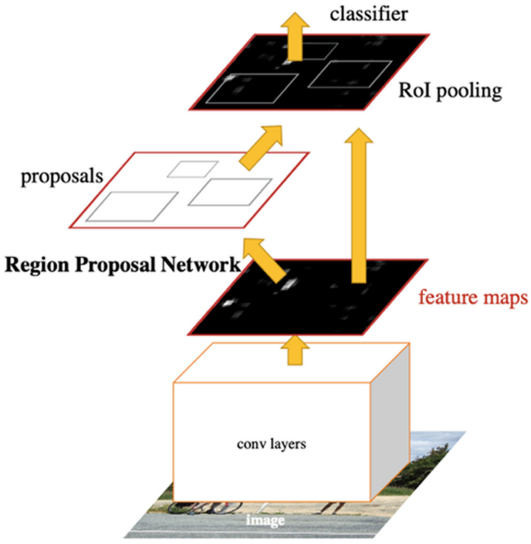


Figure 1. The Layers of Faster RCNN.

is described in detail in their own article.

The architecture of the Faster RCNN can be divided into four parts: the convolutional layers, the Region Proposal Network layer, the RoI pooling layer, and the classifier layer.

ResNet-101, a deep residual network proposed by Kaiming, He et al., is used as the feature extraction layer, which contains 5 groups of convolutional layers to calculate feature map.(He, Zhang, Ren, & Sun, 2016) The detailed architecture of ResNet-101 is presented in corresponding article, which will not be repeated here.

The Region Proposal Network layer uses a set of anchors to traverse the image first, and then uses the SOFTMAX classifier to determine whether the formed anchor frames contain object information. The next step is to correct the position of the positive anchors using regression to obtain relatively accurate position infor-mation for the next step.

The RoI pooling layer is the final processing step for the feature map and the relatively accurate object location information sourced from RPN layer. Since traditional CNNs require fixed size matrices for both input and output. This means that images with different input sizes would be cropped or compressed, which would obviously result in a loss of information. Therefore, a RoI pooling layer has been introduced, which allows the network to input matrices of varying sizes without losing information. (He, Zhang, Ren, & Sun, 2015)

The classifier layer is the last step in Faster RCNN. In this layer, the category of previously detected objects is firstly determined by the full connection layer and the SOFTMAX classifier, with a confidence score attached. Meanwhile, the position of the bounding box continues to be tweaked by regression to determine the most accurate object location.

3. Transfer Learning Driven Model Construction

For the general application of object detection, there is no need to build the network from scratch, the effort and time required to build a neural network from scratch is huge and, in most cases, unnecessary.(Chakraborty, Kovvali, Chakraborty, Papandreou-Suppappola, & Chattopadhyay, 2011) Hence, the approach of transfer learning is applied, using neural networks trained by other researchers and retrain the classifier after its full connectivity layer. This operation is called fine-tune. The object detection models trained by other researchers for usage in transfer learning are usually trained using datasets such as COCO, ImageNet, etc., with very broad category coverage and images in the millions. Therefore, the network is fine-tuned using the dataset contains the objects on the construction sites in this article to specifically detect the personnel, machine and materials on the construction site.

3.1. The Data used for Transfer Learning

For the transfer learning approach taken in this article, the algorithm only needs to be understood without further elaboration or modification, so the more critical part is the training data used in this study. The quality of the data plays a crucial role in the performance of the model. High resolution, no loss of focus, good lighting conditions, and sufficient positive and negative samples in the image all contribute significantly to the quality of the data. Unfortunately, due to time and effort constraints, as well as the unexpected impact of the pandemic at the beginning of the year in which this article was written, collecting data in outdoor areas was unfortunately not possible. Instead, site images collected by researchers at Yonsei University are used in this paper. This dataset contains 1431 images with a resolution of 960 × 540, collected by DJI Phantom 3 Professional drones from six different construction sites, which contain a large amount of personnel, machine and materials objects.(Bang & Kim, 2020) This part of the image data is of great value from both a qualitative and quantitative point of view for the content of this study.

According to the categories of personnel machines and materials objects contained in the image data described above, the objects in the construction site images are roughly divided into 10 categories in this paper, as shown in Table 1.

In order to annotate the image quickly, the VIA annotation tool developed by the VGG group of Oxford University was used in this article.(Dutta & Zisserman, 2019) After weighing the accuracy and speed of annotation, a rectangular bounding box was used to annotate the ground truth, and a JSON file was finally exported for further processing. Finally, 7286 ground truth annotations were obtained, and their distribution by category is shown in Figure 2.

Table 1. The categories of objects

Categories	Personnel	Machine	Materials
Detailed Categories	Personnel	Truck	Precast Concrete
		Excavator	Steel
		Crane	Aggregates
		Other Machines	Timber
			Other Materials

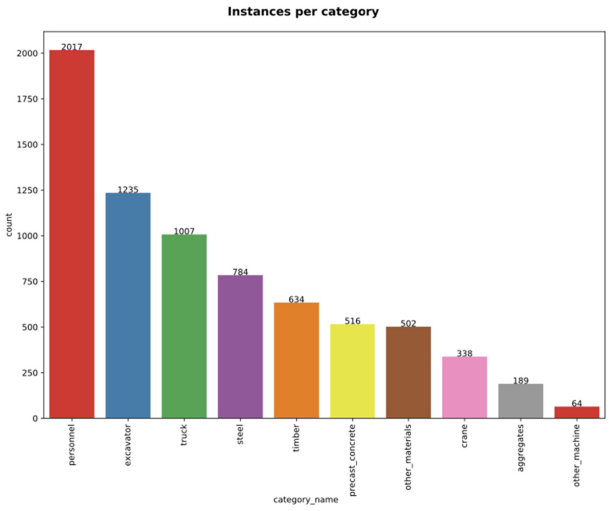


Figure 2. The distribution of annotations

For the machine learning framework and object detection API to be used in this article, the JSON files and image cannot be input directly, but have to be packaged as tfrecord files and then input into the model, which can not only ensure the integrity of the dataset, but also improve the performance of the program in reading data. Therefore, in this paper, we package the annotations and images into a training set, a validation set and a test set, which account for 60%, 20% and 20% of the total dataset, respectively.

3.2. The Framework used to Program

In this article, Tensorflow developed by Google is used. This framework is used mainly in Python and works well with many Python modules, such as NumPy, matplotlib, and so on. In addition, due to Google's massive investment in the field of machine learning, Tensorflow is extremely rich in documentation and help, and is available by default in the Colab platform that will be mentioned later, which can greatly facilitate the research of this paper. Meanwhile, Google has developed an advanced API for object detection based on Tensorflow, which can greatly reduce the code workload in the research.

3.3. The hyperparameters set before training

By using Python scripts, it is convenient to retrain a Faster RCNN model. However, before the retraining process, a few necessary adjustments to the hyperparameters of the

Faster RCNN model have to be made to ensure that the model converges quickly and effectively.

The batch size controls speed of the I/O operation in training process. Weighing the balance between speed and capacity of memory, this hyperparameter was set to 12.

The learning rate is a critical hyperparameter which involves the result of convergence. Hence this hyperparameter has to be decided with discretion. After dozens of trials, the best solution so far is 3×10^{-4} at start, 3×10^{-5} after 90, 000 iterations and 3×10^{-6} after 120,000 iterations.

3.4. The Process of Fine-tuning

After the process of fine-tuning starts, once the program retrains the model using all data in the training set, it means that one iteration is performed. The program will continue to retrain until it reaches a predetermined number of iterations. Every 10 minutes, a checkpoint is saved, and a validation process is started to validate the current performance of the model using images in the validation set and calculate mAP, AR and loss that indicate the current status of model.

The model was trained using the Google Colab platform, with NVIDIA Tesla P100 GPU. The total training time was 11 hours and 13 minutes, and 2×10^5 iterations were conducted. The overall mAP reaches 0.4305 as shown in Figure 3. The mAP value still achieves rational result under strict requirement of IoU = 0.75.

When under lax requirement of IoU = 0.5, the mAP reaches 0.6762, which is an excellent performance.

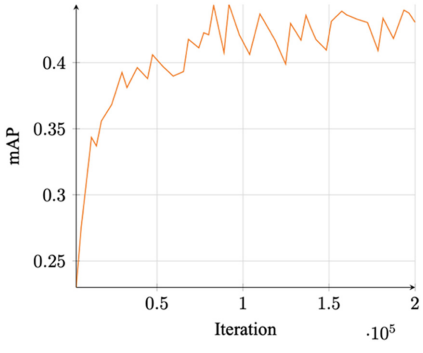


Figure 3. mAP under IoU = 0.75

However, the performance varies between different object sizes. Larger the object is, better the performance can achieve. Considering the cost of time and space to run the model, the images of the dataset used in this paper have been compressed, which accelerates the efficiency of the operation, but causes the information contained in the small bounding box to be greatly reduced, thus resulting in the unsatisfactory performance under the small objects. As shown in Figure 4, the mAP under large and medium objects indicates better performance, while the mAP under small objects is volatile and low.

In addition to evaluating how accurately the model can detect objects, the model's ability to detect all objects in the image is also an important consideration, and this is done by applying the average recall rate (AR) mentioned above. In this article, AR@1 reaches 0.3624, AR@10 reaches 0.5482, and AR@100 reaches 0.5623 at the end of retraining. it can be seen that the values of AR@10 and AR@100 are almost identical, which hint that the value of AR@100 can already well reflect the current detection performance of the model. As shown in Figure 5, the trend of the AR values under the three conditions is shown.

According to the results shown in Figure 6, the small amount of information in the small size bounding box also causes similar problems as in mAP, where the AR fluctuates greatly with the number of iterations and does not converge significantly, resulting in AR@100(small) only reaches 0.3169 at the end of the training, which is barely half the final AR value for the other size bounding boxes. The reason for this problem is the same as that in mAP, where the lower resolution speeds up the model computation but also causes heavy reduce of information in the small size bounding box.

The implementation of the back propagation of errors (BP) algorithm, which is a key concept in neural network, depends on the existence of a loss function in the model. Therefore, the loss function plays a guiding role in the model. The loss function of the model is designed to find a balance between fitting the training data set and improving the generalization ability of the model by considering various aspects. The changes of the loss

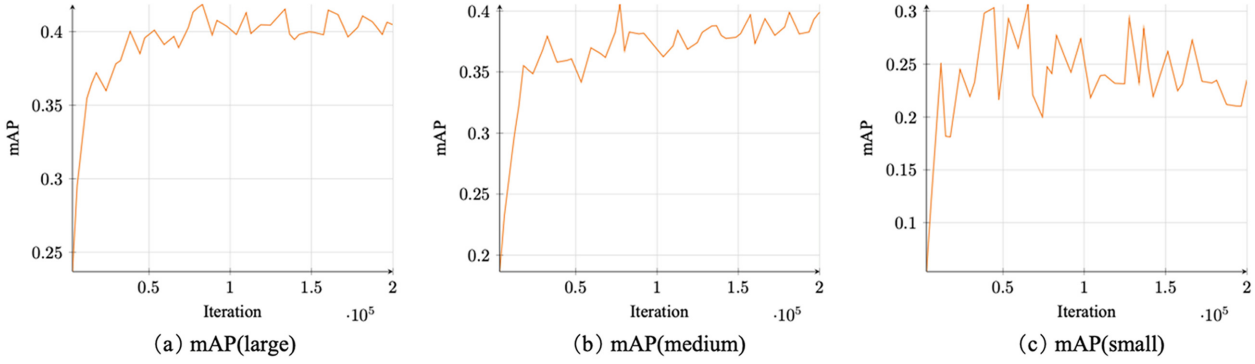


Figure 4. mAP under three sizes.

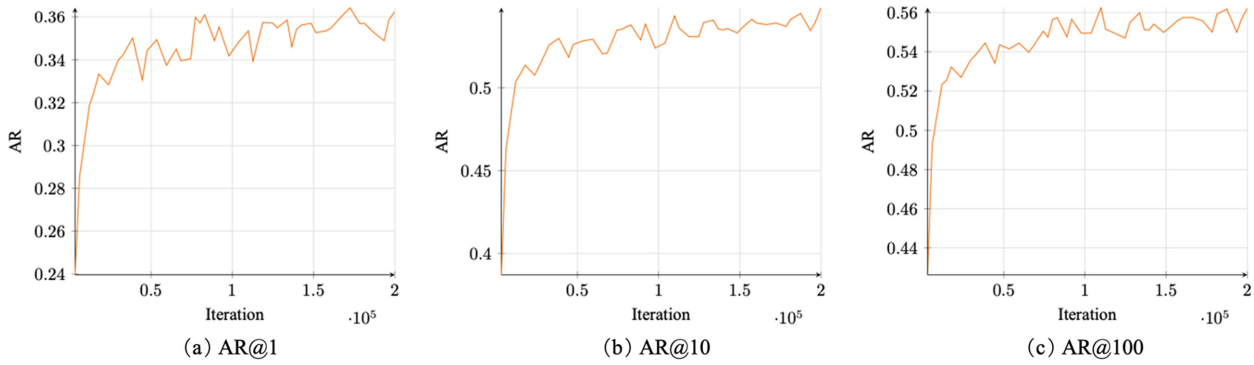


Figure 5. AR under three conditions

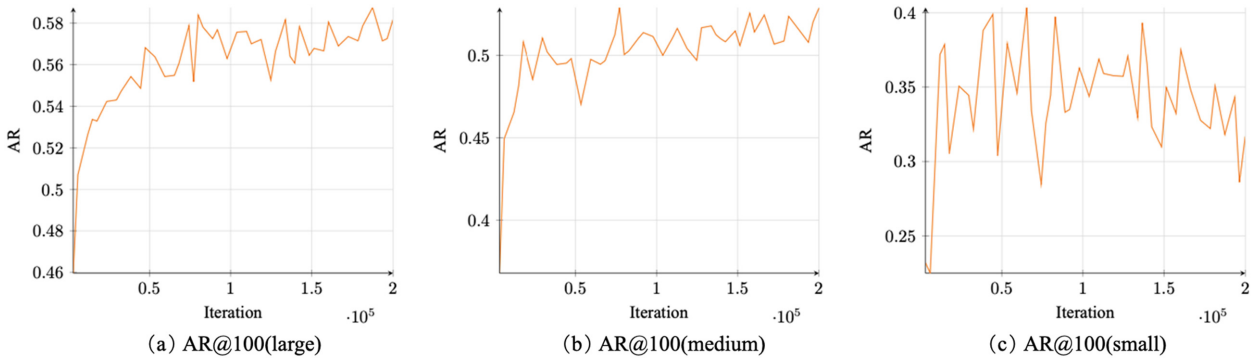


Figure 6. AR under three sizes.

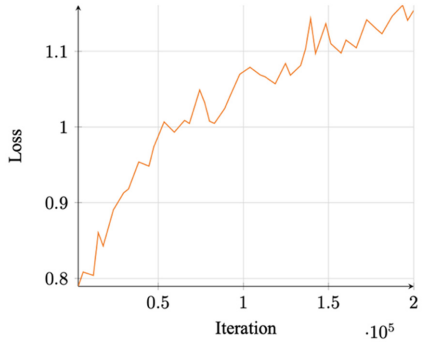


Figure 7. Total loss of the model

function of the network in the training process are obtained through the calculation during each validation, as shown in Figure 7.

It can be seen that the loss function is ascending, which is an anomalous phenomenon. The image data of the construction site is relatively messy and noisy in nature, while the source of the image data and the shooting angle are relatively single and the quantity of the images is insufficient. Given that the scalars such as mAP and AR increase while the loss function is ascending, it is very likely that the overfitting phenomenon is caused by the insufficient data set. Hence the anomalous ascend happens.

4. Test with the Fine-tuned Model

In order to check the performance of the model under data captured from totally different source. A video captured from the monitoring camera on the construction site of Huoshenshan Hospital is used as input to implement this test. The video data format is an H.264 encoded MP4 video file with a resolution of 864 × 486 and a frame rate of 25fps. It is a compressed video format with a resolution of 480P, which has some advantages in terms of speed, but again, there is a reduce of information in the video.

Leaving aside the concepts of keyframes and interframes in video coding, a video is simply a still image that is played continuously, and therefore can be processed by directly intercepting each frame of the video. According to the information of the pre-trained model provided by Google, it is pointed out that the average time for Faster RCNN to perform a single detection is 106 ms, while in the implementation in this paper, it can reach 126 ms, which means that the frame rate can reach 7.9fps, which indicates the excellent performance of Faster RCNN in the aspect of inference speed and can basically meet the requirement of real-time detection. The detection of this video is feasible.

After the original 8-hour video is intercepted, six 5-minute long clips are selected and tested for detection. The detection results are inferior compared to the result



Figure 8. Detection Result in Huoshenshan Hospital.

under validation set, but still some progress has been developed, as shown in Figure 8.

As for the excavator, crane and truck, the performance while detecting these objects are relatively the best among all objects due to the larger amount of information contained in larger bounding boxes. Besides, these objects are the most frequently annotated objects in the training data set, as a result, they are detected more accurately and robustly.

Although the personnel objects are also frequently annotated in the training data set, the detecting performance is not truly satisfactory. The resolution of the video is quite low, thus the small bounding boxes of personnel objects can only contain a small amount of information, limited by the resolution. Hence, the performance of personnel detection still needs to be enhanced.

Compared to the objects described above, which have developed initial performance, the detection of the material objects has not formed credible result. On the one hand, the appearance of these materials is relatively fixed in the training data set, which makes the model biased towards detecting materials with these specific characteristics, which is the result of an insufficient dataset. On the other hand, the resolution of the video is only 480P, making it is difficult to distinguish between many of these materials even by human, thus it is irrational to expect a better performance while using the model.

5. Conclusions

The research in this paper finally achieves the best mAP of 67.62% and AR of 56.23% on the validation dataset through transfer learning, and achieves rational detection of personnel, machine and materials objects in the construction site images taken by the drone at low altitude. By performing detection on the monitoring video captured from Huoshenshan Hospital construction site, though some defects in the model is exposed, the preliminary success is realized, and the potential of the model is revealed. With further improvements, the model can definitely become a powerful tool for intelligent site management.

References

Bang, S., & Kim, H. (2020). Context-based information generation for managing UAV-acquired data using image captioning. *Automation in Construction*, 112, 103116.

Fan, L., Zhao, H., Zhao, H., Hu, H., & Wang, Z. (2020). Survey of target detection based on deep convolutional neural networks. *Optics and Precision Engineering*, 28(05), 1152-1164.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *37*(9), 1904-1916.

Lim, R. S., La, H. M., & Sheng, W. (2014). A robotic crack inspection and mapping system for bridge deck maintenance. *IEEE Transactions on Automation Science and Engineering*, *11*(2), 367-378.

Lv, W., Zhong, Z., & Zhang, W. (2018). Overview of AI Technology. *Journal of Shanghai Electric Technology*, 11(01), 62-64.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533-536.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3), 210-229.

Sigmund, O., & Maute, K. (2013). Topology optimization approaches. *Structural and Multidisciplinary Optimization*, 48(6), 1031-1055.

Wang, Q., & Chen, Q. (2009). Method of Real Time Supervisory Control Filling and Rolling Quality Construction in the Dam by GPS. *Journal of Wuhan University of Technology*,

31(08), 79-82+108.

Chakraborty, D., Kovvali, N., Chakraborty, B., Papandreou-Suppappola, A., & Chattopadhyay, A. (2011). *Structural damage detection with insufficient data using transfer learning techniques.* Paper presented at the Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2011.

Dutta, A., & Zisserman, A. (2019). *The VIA annotation software for images, audio and video.* Paper presented at the Proceedings of the 27th ACM International Conference on Multimedia.

Girshick, R. (2015). *Fast r-cnn.* Paper presented at the Proceedings of the IEEE international conference on computer vision.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). *Faster r-cnn: Towards real-time object detection with region proposal networks.* Paper presented at the Advances in neural information processing systems.