A CV-Based Automatic Method of Acquiring and Processing Operation Data on Construction Site

Hui Li, Hongling Guo, and Zhihui Zhang

Abstract Image data of construction site is often of large volume and difficult to handle. This paper introduces a computer-vision-based automatic method of acquiring and processing this kind of data. A deep convolutional neural network along with region proposal network is used for on-site object detection including workers, materials and machines, followed by a light-weighed network to determine the real-time interaction between workers and working objects. A practical implication of the two network models and their experimental results is a scenario-based security and productivity management system and its basic structure is also introduced in this paper.

Keywords Construction site · Image processing · Convolutional neural network · Deep learning · Safety and productivity management

1 Introduction

The acquirement and process of operation data is a significant stage of construction project management. Based on reasonably collected and processed on-site data, researchers can obtain useful information related to all aspects of the project, and provide accurate basis for project safety, cost and schedule management. Conventionally, managers have adopted manual methods to acquire and process on-site data. For example, the number of on-site workers is determined by means of manual observation and counting, and information about the progress of the project is obtained by preparing a manual weekly report. The benefits of these methods are simple and intuitive, and it is relatively easy to obtain some field work data required for project management. However, the limitation is that it relies on manpower to carry out work, and its efficiency is difficult to be greatly improved. At the same time, the redundant and repetitive work carried out by manpower, whether subjective or objective, has the possibility of bias, and its accuracy is difficult to be guaranteed. Finally, when

H. Li (✉) · H. Guo · Z. Zhang
Department of Construction Management, Tsinghua University, Beijing, China
e-mail: hli17@mails.tsinghua.edu.cn

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

G. Ye et al. (eds.), *Proceedings of the 24th International Symposium on Advancement of Construction Management and Real Estate*,
https://doi.org/10.1007/978-981-15-8892-1_90

the scale of the project is large, a large amount of manpower is required for data acquisition, which restricts the economics of the project.

Fully automatic acquisition is an ideal way to obtain field work data. Through the automatic acquisition and analysis of data through the information system, it can achieve all-weather data acquisition, avoiding subjective bias, ensuring the efficiency and accuracy of data acquisition, and liberating human resources from the complicated and complicated data extraction process, reducing the cost of labor costs. Since the development and popularization of information technology, researchers have conducted a lot of exploration in this field, and proposed a series of methods for automatic acquisition of field operation data. These methods improve the limitations mentioned above to a certain extent, but are often limited by technical and hardware conditions, and their application range is narrow (for example, only supports the identification and location of a specific device), with the limitation of efficiency and real-time access.

With the further development of information technology, more and more researchers have begun to explore the automatic method of on-site image data processing, and gradually formed a three-step processing framework of image preprocessing-feature extraction-classification detection.

This study intends to build a field-based data acquisition system based on machine vision, obtain on-site images through monitoring equipment, use image algorithms in the depth learning field to process images, detect target objects (workers, materials, equipment) present at the scene, identify workers and interactions between materials and equipment, thus enables automatic extraction of field work data.

2 Literature Review

In the relatively early (before 2015) study, due to technical methods and hardware performance limitations, deep learning has not been widely used in the field, and researchers still use traditional image recognition methods. The main research idea is to use a series of methods to preprocess the image, then use the algorithm to extract the object features that you want to identify, and finally use the classifier to classify. In the process of pre-processing, the methods adopted are roughly divided into three categories: manual, semi-automatic, and automatic; the tools used include hardware tools such as radio frequency identification (RFID) and satellite positioning (GPS), as well as algorithms such as sliding window and background subtraction. The process of extracting features and classifications mostly uses algorithms popular in the technical field, such as support vector machine (SVM) for classification. Chi S (2011) proposed a construction site object automatic detection system [1], which divides the detection work into three steps: object detection, object response and object classification. It corresponds to the preprocessing, feature extraction and classification in the traditional target detection algorithm. The module becomes the basic mode for processing live image data. Due to the limitations of the traditional algorithms mentioned above, these studies tend to focus on one or several specific objects, such

as testing workers, helmets, or engineering equipment such as excavators and dump trucks.

The early identification of on-site objects was mostly based on the pedestrian detection framework that was widely studied and applied in the field of automatic driving at that time. That is, the background subtraction was removed according to the upper and lower frame analysis, and then the edge features of the object were extracted by Haar, HOG and other methods. Then, the SVM classifier is used to classify the target object in the image.

Based on this framework, some researchers conducted an empirical application: Giovanni Gualdi (2010) proposed a method to improve the detection speed and accuracy by adding context information to the model in a complex field environment map [2]. The sliding window and covariance features are detected and classified, and the classification problem of whether workers wear helmets is explored, which is of great significance for on-site safety management. Man-Woo Park (2012) extended the framework by adding a color detection module (HSV color histogram) to distinguish whether the detected person object belongs to a worker by the color of the clothing on the character [3]. Arsalan Heydarian (2012) used the spatial–temporal feature extraction-HOG-SVM framework to study the excavator and dump truck, and added the spatial–temporal feature points to the excavation arm and dump truck during the feature extraction process [4]. At the same time, it uses the classifier to automatically detect and analyze the fragments containing the device actions from the long video. Compared with some traditional sensor-based field data extraction methods, this method based on live image data recognition and detection avoids additional operations such as adding sensors, and does not require manual segmentation of the surveillance video. It offers the possibility of low cost, high efficiency, fully automatic monitoring and management in the field. ER Azar (2012) explored the feature extraction algorithm and compared the efficiency and accuracy of the Haar-HOG feature extraction method and the Blob-HOG method in detecting the dump truck [5]. W Yang (2014) detected the excavator object in the live image by using the "inverted V" shape of the excavator arm as a salient image feature [6]. Ren X (2015) discussed the effectiveness of the framework in low-light environments, and proposed a histogram equalization conversion on the grayscale image to eliminate the effect of luminance on the image [7]. In addition to the application level, researchers also made adaptations to the identification and detection framework at the technical level due to the characteristics of many identification objects and complex scenes on the construction site image data. Memarzadeh M (2013) introduced sliding windows of different scales to extract the candidate frames in addition to the traditional HOG + SVM framework, which avoids the tedious work of manually labeling the candidate frames and realizes the automation of the extraction work [8]. The sliding window remained the mainstream extraction algorithm for target detection until it was replaced by Faster RCNN's RPN network in 2015.

The above research mainly follows the three-step processing framework of image preprocessing-feature extraction-classification detection in field data extraction. The limitations of this framework are mainly reflected in: (1) Image preprocessing methods are various, and are often affected by changes in image characteristics

and detection objects, which makes it difficult to form a unified standard; (2) Manual design of image features is required, and algorithm engineers are needed for designing features. On the basis of this, it is also necessary to design a suitable classifier algorithm. Simultaneous design of features and classifiers, and the synergy between the two to achieve optimal results, requires a lot of experimental experience and a full understanding of the field. (3) The model is poorly versatile, can only be used for a few individual objects.

In recent years, the field of deep learning has received extensive attention, and the emergence of convolutional neural networks (CNN) has made far-reaching developments in the field of computer vision. CNN uses local sensation, feature parameter sharing, down-sampling and other characteristics to make up for the limitations of traditional machine learning algorithms based on statistical learning, and realizes the rapid and accurate image data processing. The generalization ability of this kind of models provides theoretical basis and algorithm basis, and has broad development prospects in the field of image data processing at the construction site.

Since the traditional statistical learning method relies on the effective extraction of features, the use in the scene environment with complex scenes and variable objects has been limited. W Fang (2018) noticed this problem and proposed applying the deep learning network for image feature extraction [9]. They improved the network structure of Faster RCNN, first performed convolution processing of the image, and obtained the feature map of the input data. Then, the FPN network was used to extract the candidate frame on the feature map, and then the ROI pooling and classification operations were performed. The researchers also debugged the parameters of the FPN network to meet the needs of on-site object detection. Although the target detection targets are concentrated in a few categories such as workers and excavators, the superior performance of the deep learning model goes beyond the traditional methods, bringing new ideas for on-site and safety management.

Based on the research results in the technical field, H Luo (2018) used the CNN network combined with optical flow diagram to extend the application of image data processing to the field of worker action and behavior recognition, and realized the classification of workers' actions in specific scenarios [10]. The inspection provides new research ideas for the safety and productivity management of the construction site. Based on the on-site target detection, X Luo (2018) put forward the evaluation index of the coexistence relationship of different target objects in specific scenarios, and explored the productivity management, which has certain theoretical significance and practical application value [11]. The limitation is that only the problem of the existence level of the object is considered, and there is no exploration research on the interaction relationship between the objects.

3 Research Method

The traditional statistical learning method relies on the effective extraction of features, and its use in the scene environment with complex scenes and variable

objects has been limited. In recent years, image recognition algorithms based on deep learning have solved the above limitations to some extent, but the application scenarios are narrow. This research hopes to further expand on the basis of the traditional video surveillance equipment, to detect the workers, materials, equipment and other targets on the construction site, and to identify and judge the possible human-object interaction relationship, and to realize the automatic extraction of the field data. On the basis of extracting the on-site information, it is possible to define the work scene, the judgment of the worker's work status, the identification and early warning of the dangerous scene, and provide the basis for on-site management. The main content of this research is divided into three parts: target detection model based on convolutional neural network, field interaction relationship discrimination model based on human-object interaction model, and visual system for field operation data acquisition and monitoring constructed by combining the two.

3.1 Field Object Detection Model Based on Mask RCNN

Computer vision refers to the use of sensors such as cameras and cameras, and the corresponding program algorithms to process and analyze the image information collected by the sensors, thereby performing functions such as identification, detection and measurement of visual objects. In short, it gives the machine a function similar to human vision. Computer vision has developed from the beginning to the present, and has a history of development for about forty years. As an application system, it has been continuously improved with the development of industrial automation.

In the mid-to-late 1990s, with the development of computer hardware performance and computing power, and the establishment of large-scale image databases, deep learning algorithms began to receive renewed attention. The advent of convolutional neural networks (CNN) has made it possible to use self-learning image features and classifiers, and machine learning based on deep learning has developed rapidly. Y LeCun (1998) proposed the LeNet network model and created the original prototype of CNN [12]. LeNet uses the three characteristics of local sensing, feature parameter sharing and down-sampling (pooling) to make up for the defects of the traditional fully connected neural network in the field of image recognition, greatly improving the recognition effect and performance of complex images. It has been used well and has received extensive attention. On the basis of LeNet, various CNN models began to appear. In 2012, Hinnet's CNN network AlexNet won the ImageNet Large-Scale Image Recognition Competition (ILSVRC 2012), which showed superior performance over other models [13]; afterwards, GoogLeNet, ResNet and other networks won the ILSVRC competition. The deep learning-based CNN network surpasses other types of methods and becomes the mainstream algorithm in the field of machine vision.

Object detection refers to finding the location and category of an object in a given picture. This work needs to solve two problems, namely, where the object is (where

the object is positioned on the image) and what the object is (image recognition). Traditional target detection algorithm uses the sliding window to perform target positioning [14], that is, select candidate frames of different sizes and positions, and judge the candidate frame by detecting the evaluation function (IOU, that is, the overlap ratio between the target window generated by the model and the original mark window) to determine whether the location is accurate, followed by CNN process to determine the classification of the object in the candidate frame. The limitations of this method are obvious: when the image contains multiple candidate objects, it is necessary to enumerate a large number of different candidate boxes, and classify and return each frame. This "violent exhaustive" regional selection strategy does not have a clear target. A lot of meaningless calculations of redundant windows are performed, resulting in very low detection efficiency.

In order to solve this problem, the researchers tried to and uniformly extract the convolution feature of the whole picture to obtain the convolution feature map before the candidate frame selection, and then map the candidate frame to the convolutional layer feature map on the original map to perform subsequent classification work. Based on SPP Net proposed by He et al. [15], Girshick improved the RCNN model in 2015 and proposed Fast RCNN, which solves the problem of repeated convolution on the basis of RCNN [16], and uses Multi-task Loss to connect the original series so the classification and positioning work is combined into a loss function. These two changes have greatly increased the speed of training and monitoring. Based on the Fast RCNN, Girshick introduced the Region Proposal Network to hand over the task of searching for candidate boxes to the neural network. Well-trained RPN can greatly reduce the time to produce candidate frames, break through the performance bottleneck of Fast RCNN, and take a big step toward end-to-end real-time detection. So the network is named Faster RCNN [17]. The above mentioned W Fang, H Luo and others used the Faster RCNN for on-site object detection.

The network has good performance and recognition in the experimental environment but not in the actual on-site environment. The reason is that the situation on the site is complex and varied, and the target objects to be detected have different shapes. Besides, the pixel-level segmentation of the detection target is required for subsequent work, and the faster RCNN cannot meet the needs of researchers. Based on this, we have adopted the latest Mask RCNN model for our on-site object detection work. He et al. made three changes based on the Faster RCNN and proposed the Mask RCNN [18]. The first is to replace the original VGG network with a residual network with stronger feature representation net (ResNet, ResNext). The second is to replace the original ROI Pooling layer with the ROI Align layer, which reduces the loss of information at the pixel level. In the section of Multi-task Loss, a branch of Mask loss is added to locate the Mask. After the method is proposed, the highest recognition accuracy in the RCNN class method is realized. Due to the fine-to-pixel-level object segmentation method, Mask RCNN not only achieves high-precision target detection, but also has strong expansion potential in machine vision segmentation such as human pose estimation.

We chose the Mask RCNN model project written by Matterport to build our own target detection model. The model has been modified based on the paper model of

He et al., including zoom fill processing for image input and a smaller learning rate. Along with the model Matterport also provides a weight file for the model to be pre-trained based on MS COCO data. The MS COCO dataset is a dataset provided by Microsoft for common object detection and segmentation. The 2014 version of the dataset includes 82,783 training images, 40,504 validation images, and 40,775 test images, 270 k segmented people and 886 k segmented Object. The entire dataset size is about 20 GB. Based on the pre-training weight file provided by Matterport, the model can better achieve the target detection work in 91 categories in daily life scenes. Since the purpose of our model is to detect the target object in the field, which are somehow similar to common objects in the original COCO dataset, we do not need to relearn the low-level features of the image (points, lines, corners) and some advanced features (texture, color, body block, etc.), and most of the content in the pre-trained model is reusable. We only need to use the transfer learning method to retrain the classifier part on the live image dataset, so that the model can achieve the target detection based on the dataset we defined. Based on the transfer learning method, we need to do the following works: data collection and screening, image annotation and data set construction, model training and verification.

3.2 On-Field Interaction Detection Model Based on HOCNN

In recent years, more and more researchers have begun to use the emerging deep learning methods to carry out target detection work on the construction site. However, most of their work stays at the level of detection and classification of different targets, without further thinking about other explorations that can be carried out based on object detection. Based on above mentioned object detection model based on Mask RCNN, this study hopes to judge the possible interaction between the pixel information of different objects by spatial shape and relationship, and realize a series of management objectives such as on-site safety management and productivity management.

Human-Object Interaction (HOI) detection is another sub-area in the field of machine vision. HOI is mainly concerned with the interaction between people and objects, which can be spatial representations (people sitting on chairs) or logical abstractions (people are repairing bicycles). Unlike the action/activity recognition through video, HOI still focuses on the content of static images, focusing on the state relationship between people and things rather than continuous action relationships. Due to static images rather than dynamic video, the HOI model is small and easy to migrate. At the same time, due to the rapid development of recognition and detection algorithms based on static pictures in recent years, the development prospects of HOI are relatively broad. Yu-Wei Chao et al. proposed the HICO (Human Interacts with Common Objects) benchmark [19] in 2015 and the HOCNN model [20] in 2018, and established a data system for HOI detection. Humans and objects were used by the popular RCNN at that time. The identification and classification, and the concept of Semantic Knowledge, including Composition and Co-occurrence, are proposed in

the process of interaction classification. By adding the V-O-Composition and the Co-occurrence to the model for training, HOCNN has achieved better results in the HOI detection field than the traditional algorithm. In 2018, YW Chao proposed HICO-DET and improved HO-CNN, adding a branch that judges the relative positional relationship between human and object space, by overlapping the bounding boxes of people and objects. The feature learning of spatial relative position is an important basis for the classification of interaction relationships. At the same time, this article also gives the format and database construction method of HICO-DET data set. Most of the research in the HOI field is based on the object detection work, and the classification and content of the interaction relationship are obtained through the judgment of the spatial relationship. Extending it to the engineering field, it is possible to classify and detect the spatial interaction between workers and materials and mechanical equipment in the image, and in the field of productivity management, it can make up for the problem of weak logic relationship brought by only the target detection and coexistence evaluation indicators; In the field of security management, judge and warn of abnormal interactions.

The interaction between field workers and materials and equipment can be roughly divided into two categories: spatial location and logical interaction. The relationship between the spatial locations is relatively straightforward: it can be considered that they are irrelevant if the distance is far away; if the workers are close to the material equipment, we may think that the two are about to interact or are interacting. The specific logical interaction is judged based on the spatial interaction and the type of object. For example, if the outline of the worker on the image is inside that of a scaffolding, the worker can be considered to be inside of it; if the outline of the worker is at the edge of the image and far away from the scaffold, it can be considered that there is no interaction relationship between the two. Based on the above-mentioned on-site target detection model, we have been able to extract the type of worker and material/equipment and the target pixel location from the live image. Using this as an input condition for the model, the interaction between workers and equipment/materials can be determined.

In the stage of processing the image data of the scene, the work of judging the type of the object from the shape, color, texture and other features of the image has been completed by the target detection network, and the interaction relationship detection work is only based on the relative spatial relationship between the person and the object. So the input layer of our model is a two-channel image, and each pixel point can only have a value of either 0 or 1, with 0 meaning background and 1 meaning pixel of the person/object.

4 On-Site Object Detection Model

In order to construct a small-scale data set for empirical research, we collected and screened 4.8 k on-site images through various channels, and divided them into two parts: training set and test set according to the ratio of 3:1.

Before making the image annotation, we defined the classification of the on-site targets. According to the common on-site work scenes and expert guidance, as well as several principles such as frequency of occurrence, universality, importance to safety and on-site management, the site targets are divided into three categories: workers, equipment, and materials, and 18 sub-categories (see Table 1).

Subsequently, we used the Labelme software to annotate the workers, equipment, and materials on the image in a polygonal manner according to the established labeling rules. A total of 4.8 k images were completed. These annotations are mainly done to the students in the lab. Since the data input interface of the Matterport version is in the MS COCO dataset format, we also collated the finished images and created the MS COCO format dataset. After making the necessary modifications to the model, starting with the modified weight file as a starting point, the training results obtained by training 40 Epoch on our own data set are shown in Fig. 1.

We randomly select a few images from the test set and use the trained model to perform target detection. The results are shown in Fig. 2.

It can be seen that the model can properly detect the workers and various materials and equipment in the construction scene. We believe that it is feasible to use the deep convolutional neural network to detect the targets on the construction site. If the number and quality of datasets, as well as the structure of the model and algorithm can be improved, the speed and accuracy of the inspection work can be further improved.

Table 1 Definition and classification of on-site objects

Category	Subcategory
Worker	Worker-helmet
	Worker-nohelmet
Material	Rebar-working
	Rebar-material
	Steel
	Concrete-pouring
	Formwork-working
	Formwork-material
	Scaffolding
Machine	Excavator
	Bulldozer
	Dump-truck
	Concrete-bucket
	Concrete-mixer
	Concrete-pump
	Tower-crane
	Crane
	Machine-other

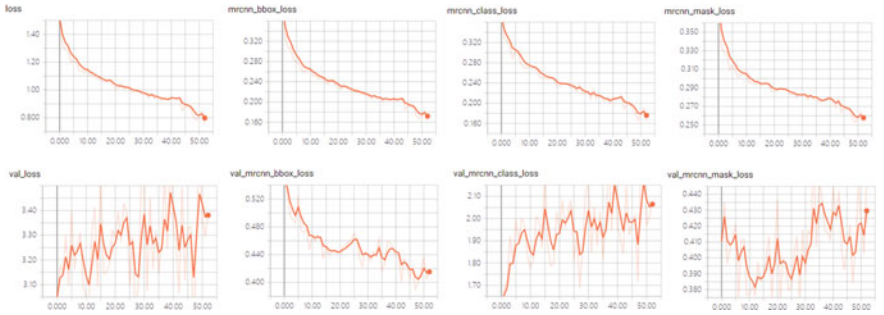

Fig. 1 Result of training/test loss

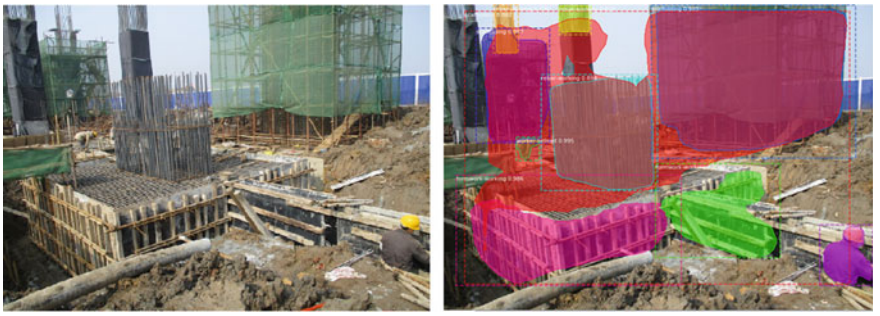

Fig. 2 Example of results of on-site object detection

5 Worker-Object Interaction Classification Model

Based on the target detection data set constructed above, we constructed an on-site interaction dataset. We divided the annotated site targets into two categories: "workers" and "objects". When the bounding boxes of "workers" and "objects" overlap, we think that there may be an interaction relationship between them, and extract them as one sample. Based on the original object detection dataset, we extracted 7 k interactive samples, including all 16 object types and 11 common verbs defined (including not-interacting-with). List of applicable verbs and counts of the underlying subcategory are shown in Table 2.

We also established an online annotation platform. These samples are annotated by college students and the labeling results are as shown in Fig. 3.

Rebar, formworks, scaffolding and steel structures are the four most common materials on the construction field, and are the objects with the most samples in our dataset. We observed the two-channel image of the scaffolding object in different interactions across the entire dataset and the sum-up images are as shown in Fig. 4, with yellow parts of the picture indicating high frequency of object appearance. We can intuitively see that in the sample of the worker-scaffold interaction, the

Table 2 Valid verbs for different subcategory of objects

Category	Subcategory	Counts	%	Verbs-valid
Material	Rebar-working	2350	33.11	Next-to, installing
Material	Scaffolding	1570	22.12	Next-to, standing-on, installing, standing-in
Material	Rebar-material	901	12.69	Next-to, carrying, welding
Material	Steel	697	9.82	Next-to, standing-on, installing
Material	Formwork-working	545	7.68	Next-to, standing-on, installing, uninstalling
Material	Formwork-material	386	5.44	Next-to, carrying, standing-on, installing
Material	Concrete-pouring	7	0.10	Leveling
Machine	Machine-other	250	3.52	Operating, next-to
Machine	Excavator	138	1.94	Next-to, operating, standing-under
Machine	Tower-crane	114	1.61	Standing-under, next-to
Machine	Crane	87	1.23	Operating, standing-under, next-to
Machine	Dump-truck	27	0.38	Operating, standing-on, next-to
Machine	Concrete-pump	13	0.18	Operating, next-to
Machine	Concrete-mixer	10	0.14	Operating, next-to
Machine	Bulldozer	2	0.03	Operating, next-to
Machine	Concrete-bucket	1	0.01	Operating, next-to

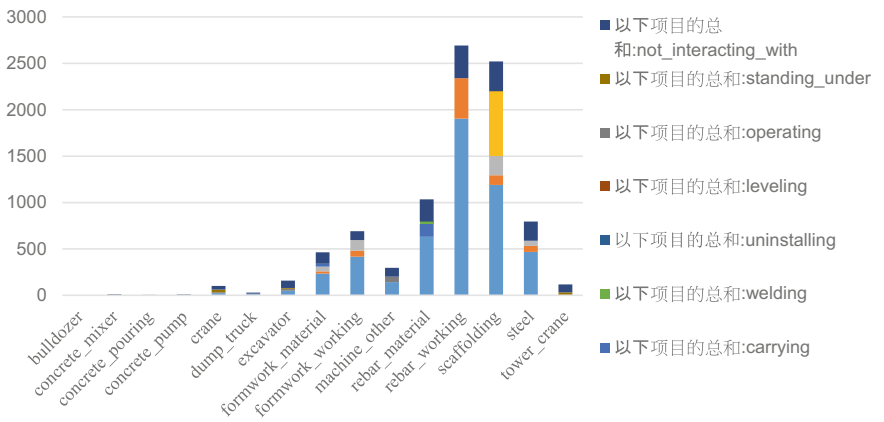

Fig. 3 Structure of worker-object interaction dataset

scaffolding is usually at the center of the sample screen; in the sample where there is no interaction, the scaffolding is usually located above the screen, and the worker is located at the bottom of the screen corner. This is well understood: such images are usually scaffolding in the distance and the workers are close to the camera. The distance between the two is far, but there is overlap on the picture.

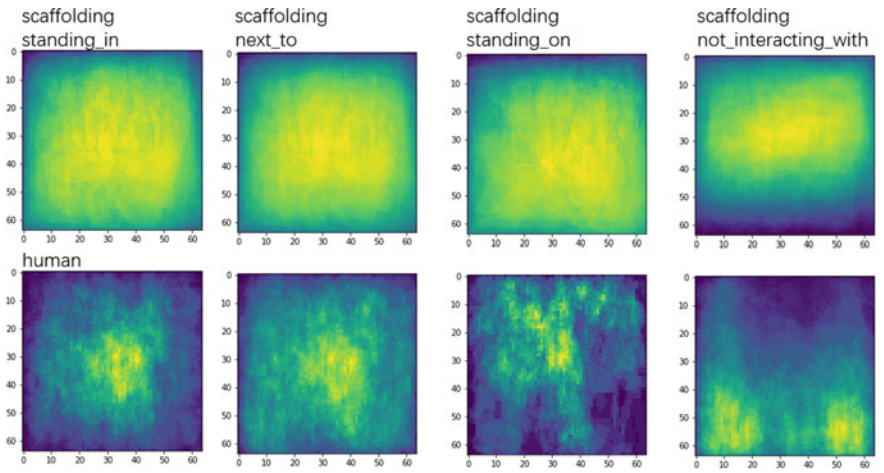

Fig. 4 Sum-up of different interactions with scaffolding

After completing the necessary data preparation work, we built the appropriate model for training. Since the input data is a two-channel contour image and the format is relatively simple, we have adopted a similar approach to Yu-Wei Chao: on the basis of the convolutional network commonly used for image processing, a fully connected network is added as an effect comparison. The image input of the model is 64 * 64 in size. For different sizes of samples, we have two solutions: stretch the image to the input scale (without padding), or scale it down and pad it with 0 on the short side (with padding). Figure 5 shows these two solutions of resizing samples.

At the same time, we constructed two kinds of logic as of the structure of the model: "interaction only" and "object to verb"; the former model directly judges the type of interaction (56 categories) from the input image, and the latter model input not only contains image data, but also includes the type of the object that interacts

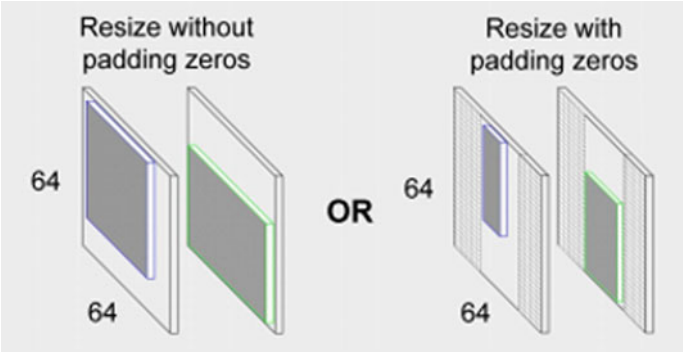

Fig. 5 Different solutions of resizing samples

with the worker. We randomly selected 5.9 k interactive samples as the training set and 1.1 k samples as the test set. Each model was trained for 50 epochs, and the accuracy rate was used as the evaluation index. The two model structures and their train/test results are shown respectively in Figs. 6 and 7.

It can be concluded that for an "interaction only" network, whether it is a fully connected structure or a convolutional structure, its accuracy on the training set is always around 10%; for the "object to verb" network, the accuracy can reach a level of about 45%. Since the performance of the deep learning method is very dependent on the size and quality of the data set used for training, and the industry lacks a complete image dataset. We used a relatively small-scale dataset in our research (4.8 k images and 7 k interaction samples, with COCO containing over 120 k images and HICO containing over 200 k samples) and the model performance will inevitably have a gap with the mainstream model performance of computer vision academia. It should also be pointed out that the performance of the model is also subject to the application scenario, so the absolute value of the accuracy cannot be used simply. The mainstream practice in the technology field is to compare model performance

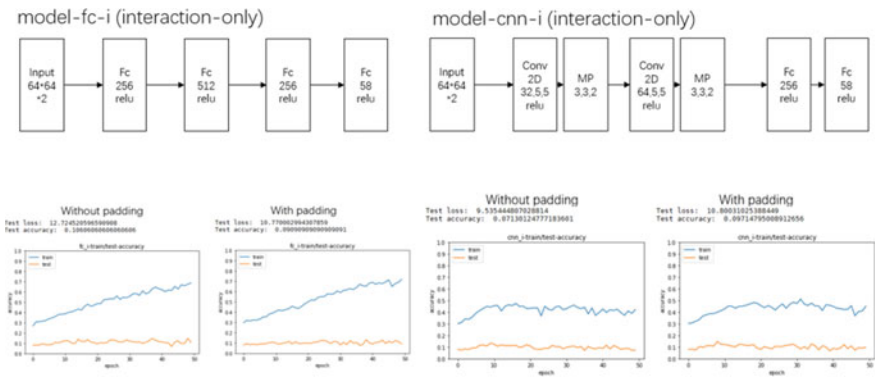

Fig. 6 Interaction-only model structures and train/test results

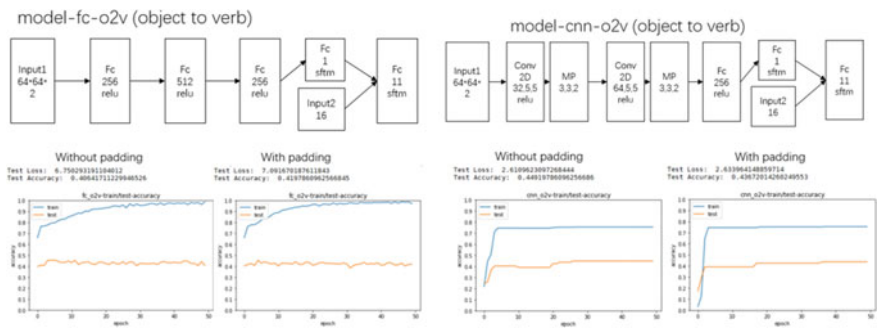

Fig. 7 Object-verb model structures and train/test results

with the accuracy of manual operations in the same scenario. For the scene of the interactive relationship judgment of the scene, since the definition of the "interaction relationship" itself has certain uncertainty, the accuracy rate is only 50%-60% when the artificial interaction relationship is judged, so it can be considered that our model is effective. We have reason to believe that the performance of the model has a lot of room for improvement in the future when industry standards are established and the scale and quality of the dataset are greatly improved.

6 Practical Implication: Scenario-Based Security and Productivity Management System

Based on experimental outcomes of the two models mentioned above, we explored a new scenario-based security and productivity management system. Its main structure is shown in Fig. 8.

Based on the future large-scale, high-quality industrial image dataset, the system is promised to achieve high-precision on-site object detection and worker-object interaction recognition. We can define the interactions in a particular scenario, thinking that whether the interaction is dangerous (safety management) or efficient (productivity management). Thanks to the fully automatic, all-weather data acquisition and real-time processing, we can quickly alert and deal with unsafe behavior on the site in a short time; for a long time period, we can work on the overall efficiency and

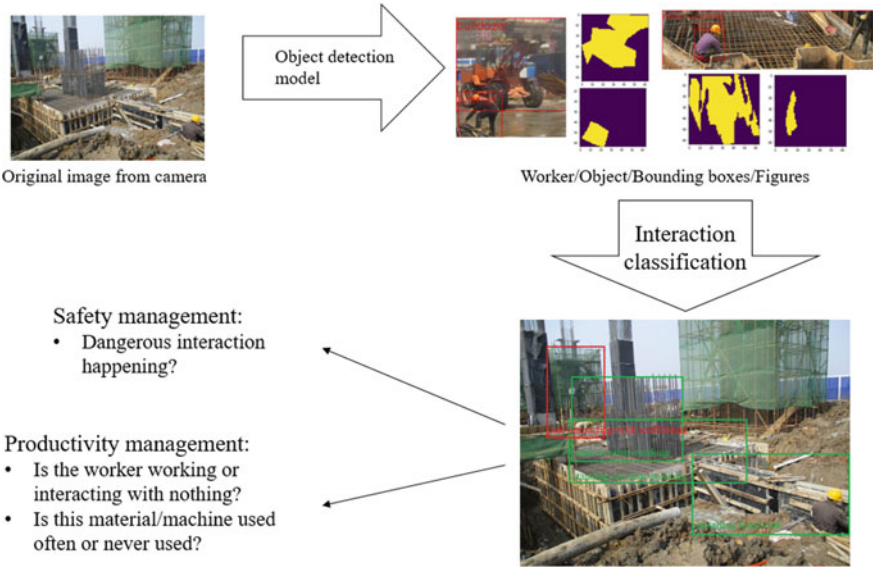

Fig. 8 Safety and productivity management system

productivity of the site, conduct assessments to provide a basis for coordination of workforce and materials on site.

7 Conclusion

Based on characteristics of on-site image data, this paper introduces a deep learning method to collect and process images, and explores the application of on-site safety and productivity management system based on this method.

For now, the biggest limitation of the study is the lack of scale and quality of the training dataset. Though our model shows considerable potential, we have relatively small amount of raw image data, and the collation and labeling work is done manually by the students, so there are deficiencies in both quantity and quality of the dataset. In addition, the structure and algorithms of the model itself can be further optimized for industry and data characteristics. The current object detection model is large in volume, and real-time detection is not possible on a single GPU platform. How to streamline parameters and improve the performance of the model is also a problem we need to consider.

In the future, we will promote data collection and labeling outsourcing work, and strive to establish our own image application datasets for the construction industry; on the other hand, we will continue to follow the computer vision field, especially the latest developments in deep learning related research, to improve the algorithms and structure of our model. Through these two aspects of work, improve the performance level of the model and prepare for future system applications.

Acknowledgements This work was supported by the National Key R&D Program of China (no. 2016YFC0802001); the National Natural Science Foundation of China (Grant No. 51578318, 51208282); and Tsinghua-Glodon BIM Research Center

References

1. Chi, S., & Caldas, C. H. (2011). Automated object identification using optical video cameras on construction sites. *Computer-Aided Civil & Infrastructure Engineering, 26*(5), 368–380.
2. Giovanni, G., Andrea, P., & Rita, C. (2011). Contextual information and covariance descriptors for people surveillance: An application for safety of construction workers. *EURASIP Journal on Image and Video Processing, 2011*(1), 1–16.
3. Park, M. W., & Brilakis, I. (2012). Construction worker detection in video frames for initializing vision trackers. *Automation in Construction, 28*(15), 15–25.
4. Heydarian, A., Golparvar-Fard, M., & Niebles, J. C. (2012). Automated visual recognition of construction equipment actions using spatio-temporal features and multiple binary support vector machines. In *Construction research congress 2012: Construction challenges in a flat world* (pp. 889–898).
5. Rezazadeh Azar, E., & McCabe, B. (2011). Automated visual recognition of dump trucks in construction videos. *Journal of Computing in Civil Engineering, 26*(6), 769–781.

6. Yang, W., Li, D., Sun, D., & Liao, Q. (2014, November). Hydraulic excavators recognition based on inverse" v" feature of mechanical arm. In *Chinese Conference on Pattern Recognition* (pp. 536–544). Berlin, Heidelberg: Springer.
7. Ren, X., Zhu, Z., Germain, C., Dean, B., & Chen, Z. (2015). A case study of construction equipment recognition from time-lapse site videos under low ambient illuminations. In *Computing in civil engineering 2015* (pp. 82–89).
8. Memarzadeh, M., Golparvar-Fard, M., & 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.
9. Fang, W., Ding, L., Zhong, B., Love, P. E., & Luo, H. (2018). Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach. *Advanced Engineering Informatics, 37,* 139–149.
10. Luo, H., Xiong, C., Fang, W., Love, P. E., Zhang, B., & Ouyang, X. (2018). Convolutional neural networks: Computer vision-based workforce activity assessment in construction. *Automation in Construction, 94,* 282–289.
11. Luo, X., Li, H., Cao, D., Dai, F., Seo, J., & Lee, S. (2018). Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *Journal of Computing in Civil Engineering, 32*(3).
12. LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324.
13. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).
14. Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 580–587).
15. 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.
16. Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1440–1448).
17. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems* (pp. 91–99).
18. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)* (pp. 2980–2988). IEEE.
19. Chao, Y. W., Wang, Z., He, Y., Wang, J., & Deng, J. (2015). Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1017–1025).
20. Gkioxari, G., Girshick, R., & Malik, J. (2015). Contextual action recognition with r*cnn. *International Journal of Cancer Journal International Du Cancer, 40*(1), 1080–1088.