Advanced Engineering Informatics 39 (2019) 170–177

Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Full length article

A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network

Weili Fang[a,b,c], Botao Zhong[a,b,*], Neng Zhao[a,b], Peter E.D. Love[c], Hanbin Luo[a,b], Jiayue Xue[d], Shuangjie Xu[d]

[a] Dept. of Construction Management, School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, China
[b] Hubei Engineering Research Centre for Virtual, Safe and Automated Construction, China
[c] Dept. of Civil Engineering, Curtin University, Perth, Western Australia, Australia
[d] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei, China

ARTICLE INFO

Keywords:
Deep learning
Falls from height
Computer vision
Unsafe behavior
Mask R-CNN

ABSTRACT

Structural supports (e.g., concrete and steel) provide engineering structures with stability by transferring loads. During the construction of an engineering structure, individuals are often prone to taking short take-cuts by traversing supports to perform their daily activities and save time. Thus, the likelihood of an individual being subjected to an injury or even killing themselves significantly increases when performing such unsafe behavior. To address this problem, we have developed an automatic computer-vision approach that utilizes a Mask Region Based Convolutional Neural Network (R-CNN) to detect individuals traversing structural supports during the construction of a project. The algorithms developed are used to: (1) automatically identify the presence of people; and (2) recognize the relationship between people and concrete/steel supports to determine their presence of them. To validate our approach, we created an extensive database of photographs of people who had traversed structural supports from a number of different constructions project to train and test the developed Mask R-CNN. The recall and precision rates for overlapping detection were found to be 90% and 75%. The results demonstrate that the developed Mask R-CNN can accurately detect people that traverse concrete/steel supports during construction. We suggest that proposed computer-vision approach that we have developed can be used by site management to automatically identify unsafe behavior and provide feedback to individuals about their likelihood of falls from heights. By recognizing unsafe behavior in real-time, appropriate actions (e.g. education) can be instantly put in place to prevent their re-occurrence.

1. Introduction

Falls from heights (FFH) are a major contributor to workforce fatalities in construction [1]. It has been revealed that about 48% of accidents that occur in construction arise from FFH [2]. The construction of deep foundation pits for an underground metro-system, for example, invariably requires concrete and steel supports to be deployed to stabilize soil and transfer loads. In China, for example, individuals traversing structural supports over deep foundation-pits during the construction of underground metro-systems has been identified as a problem that site management have to regularly police, particularly as they a prone to not wearing a safety harness even when required to work at heights [3–5].

Numerous safety policies and procedures have been established to prevent people from FFH. For example, the Occupational Safety and Health Administration requires that a person working on a surface (horizontal and vertical) with an unprotected side or edge that is six feet or more above a lower level must be protected from falling by the use of guardrail systems, safety net systems, or a personal fall arrest system (PFAS) [6]. Yet, despite the considerable amount of research that has been undertaken and the implementation of policies, procedures and the development of protection measures, FFH remain a pervasive problem worldwide [5,7].

The effectiveness of the link between safety climate and behavior plays a role in mitigating unsafe behavior. Traditionally, interventions have centered on the physical work environment and procedures to prevent errors and accidents [8]. Examples include the documentation of detailed procedures designed to provide the safest way to complete

* Corresponding author at: Dept. of Construction Management, School of Civil Engineering and Mechanics, Huazhong University of Science and Technology, Wuhan, Hubei, China.
E-mail address: dadizhong@hust.edu.cn (B. Zhong).

https://doi.org/10.1016/j.aei.2018.12.005
Received 12 October 2018; Received in revised form 21 December 2018; Accepted 22 December 2018
Available online 09 January 2019
1474-0346/

work, and the requirement for Personal Protective Equipment (PPE) to be worn [8]. We have seen a subtle shift away from strict procedural guidelines that aim to mechanize and standardize behavior in construction, to a position that acknowledges individual differences and focuses on the psychological issues that influence behavior [9–11]. When an unsafe act occurs a failure in an individual's cognition is produced. Such failure may arise at one of the following stages of the cognition process [12]: (1) detecting hazards; (2) recognizing hazards; (3) perceiving hazards; (4) deciding a response; and (5) executing the decided response. We can see from the research undertaken by Zhang and Fang [5] that there is a proclivity for people working at height in China (e.g. scaffolders) for their cognition to fail at stage four as they purposefully choose not to wear a safety harness even though they are legally required to do so. Such actions are referred to as violations, which are influenced by a person's risk perception [13].

When a behavior is perceived to be a lower risk than others, then an individual's motivation for safety is more positive [9]. But, Bohm and Jonathan [14] observed during construction when individuals perceive that there is a high risk of an unsafe behavior, they still engage in this action. Essentially, workers break-rules to make their work more efficient [15–17]. In the case of structural supports, people tend to use the structural supports as a short take-cut even though there is a likelihood that they can fall and injure themselves or even be killed [15,16,18]. People also judge managers' attitudes to safety through their: (1) behaviors, interactions with them; (2) willingness to listen and learn; and (3) provide feedback [9–11]. However, the recognition of unsafe behavior often only arises during site inspections, or accidently or even when a co-worker informs site management. Thus, the opportunity for providing feedback to individuals about their unsafe behavior is limited.

To provide site management with the ability to recognize in real-time unsafe behavior and provide feedback to mitigate FFH, we develop a computer vision-based approach that can be deployed to automatically identify and capture people that traverse structural supports without PFAS in place. The approach that we have developed provides a cost-effective alternative to having a person physically observe workers and can be used to engender a climate of safety [19,20]. Our research aligns with the benefits identified by several computer-vision based studies that have been undertaken in construction as we are able to automatically identify objects and recognize the activities undertaken by workers [19–24]. While there has been a plethora of computer-vision studies undertaken in construction, there has been, to the authors knowledge, no research that has examined the issue of people traversing structural supports. Therefore, the research presented in this paper seeks to address this knowledge gap. In doing so we utilize a deep learning Mask Region Based Convolution Neural Network (R-CNN) to develop a computer vision approach that can accurately detect the presence of people traversing structural supports. We commence our research by commencing with a review of the extant computer vision literature that surrounds behavior recognition, then we introduce our novel computer vision approach, which is subsequently tested and validated using an experiment.

2. Literature review

It has been observed that approximately 88% of all accidents that occur during construction materialize as a consequence of unsafe behavior [25]. With this in mind, a number of studies have been undertaken that have focused on monitoring and capturing peoples' activities during construction, specifically their unsafe behavior using a variety of technologies [26–34]. For example, non-visual sensors (e.g., radio frequency identification (RFID) [35], ultra-wide band (UWB) [36–38] and global positing system (GPS) [39,40]) have been utilized. Teizer et al. [41] developed an automatic system to a capture information for equipment operators to help identify individuals who were working in their blind spots to mitigate hazards. Likewise, Yang et al. [31] used a wearable inertial measurement unit (WIMUs) to collect people's kinematic data, and applied a semi-supervised approach to detect near miss falls. Similarly, Akhavian et al. [33] used a smartphone to determine a person's body movement, and then developed a machine learning approach to classify their activity.

Existing methods based on non-visual sensors generally track an individual's location. The use of such technology requires the installation of sensors or markers to be attached to a person's body to track their motion, but can restrict a person's movement [26]. Moreover, non-visual sensors have not been designed to capture contextual information, which renders them obsolete for determining when a person traverses a structural support.

With the advent of high-resolution video cameras, the augmented storage capacity of databases and increasing accessibility of the Internet, the ability to document operations in construction has been transformative. As a result, applications of computer vision have become increasingly popular to monitor progress, individual activity and recognize their unsafe behavior, as a rich set of information (e.g., location and pose) from images and videos that can be acquired. The recognition of unsafe behavior has been a popular application of computer vision in construction, particularly the of detection of PPE. For example, Fang et al. [19] developed a hybrid deep learning approach to determine if a worker was wearing their safety harness while working at heights. Several studies have applied computer vision to detect people who enter hazardous work areas. Kim et al. [42], for example, integrated computer vision with a fuzzy inference to monitor and assess the safety of people performing their tasks in the vicinity of plant.

2.1. Convolutional neural network-based object detection

Recognition of objects (e.g. people, plant, materials, equipment, and PPE) is a core to evaluating the performance of tasks in construction. The information that is obtained can assist site management with their decision-making and help improve safety performance. Using a background subtraction algorithm Chi and Caldas [43] extracted features from video and then used a naïve Bayes classifier and neural network to identify people and plant (e.g., loaders, and backhoes). Contrastingly, Park and Brilakis [44] and Rezazadeh Azar and McCabe [45] utilized Histogram of Oriented Gradient (HOG) and Haar-like features to detect people and equipment. Similarly, using video Memarzadeh et al. [46] combined a HOG and color features with a new multiple binary Support Vector Machine (SVM) classifier to automatically detect and distinguish between people and equipment. Previous research has tended to reply on manually hand-crafted feature approaches to extract them from inputs, and then send them to a classifier (e.g., SVM, *k*-Nearest Neighbors (*k*-NN)) to enact the process of detection [47]. While such research has been able to provide good detection performance of objects, studies have been unable to distinguish and generalize from the background that surrounds them [47,48].

A CNN is considered to be an effective way to automatically extract and learn features from high-dimensional images contained within a database using end-to-end processing [48]. A number of object approaches have been employed in construction (e.g., Faster R-CNN, You Only Look Once (YOLO), and Single Shot Multi-box Detector (SSD)) [49]. The Faster R-CNN has been identified as the best approach for object detection due to its ability to provide high levels of accuracy [49,50]. But, the Mask R-CNN has been found to outperform all existing object detection approaches that have been developed [51]. Thus, to achieve a higher level of accuracy to detect workers that traverse structural supports, we apply the Mask R-CNN approach in this research.

3. Research methodology

In accordance with previous computer vision research that has

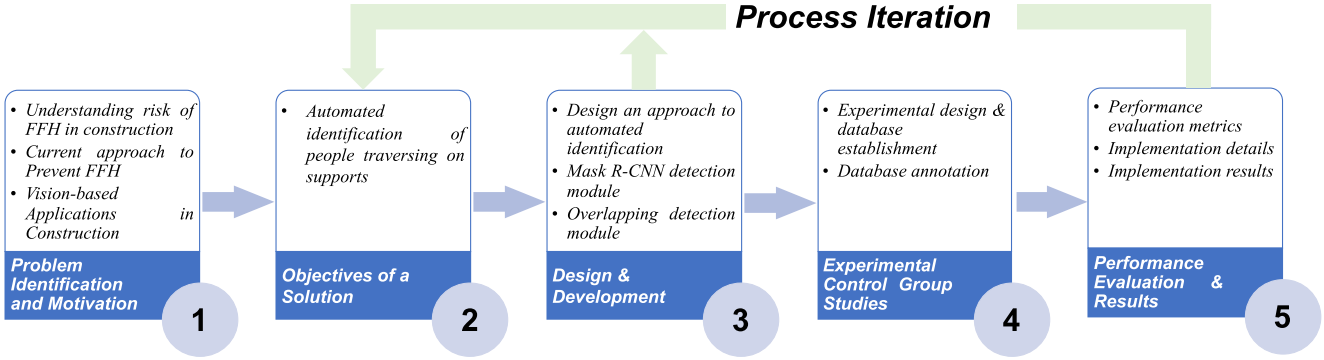

Fig. 1. Design science approach: Research process [54].

examined unsafe behavior, we have adopted a design science research methodology to develop a deep learning model that can be used to automatically recognize the unsafe behavior of workers that enacted to traverse structural supports [19,21,52]. Design science is widely accepted as a practical multi-discipline method that has been widely applied by the fields of engineering, computer science and business [53,54]. With the innovation and creation of artifacts, design science can be used to understand, explain and improve existing systems in a systematic manner [55]. The design science process typically incorporates five steps (Fig. 1): (1) problem identification and motivation, (2) definition of the objectives for a solution; (3) design and development; (4) demonstration; and (5) evaluation [56]. The research process that we have adopted to develop our model to identify unsafe behaviors of people that traverse on structural supports is presented in Fig. 1. We have identified the problem and explained above why we need to develop a computer vision solution to identify unsafe behavior using a new algorithm: The Mask R-CNN. The process used to design, develop, test and evaluate our approach is presented below.

3.1. Design and development

Our developed approach consists of: (1) a Mask R-CNN module, which is used to detect structural supports and people; and (2) an Overlapping Detecting Module, which is used to recognize the relationship between a person and the structural support. In this research, we use Mask R-CNN to train three object classes: (1) people; (2) steel support; and (3) concrete supports. Fig. 2 presents the workflow of our proposed unsafe behavior recognition approach. The procedure to implement our approach is described as follows:

Step 1: The Mask R-CNN network takes an entire image as an input. After extracting a feature map from the original image, a network referred to as Region Proposal Network (RPN) is introduced into the model to propose the candidate object bounding boxes, which forms the key element for Step 2.
Step 2: Input the candidate object bounding boxes obtained from Step 1 into the Region of Interests (RoI) Align layer, which is described in more detail below. Then, the new feature maps from each candidate box are extracted from RoIAlign layer. These feature maps are used to perform classification, bounding box regression and mask generation.
Step 3: The last module is referred to as the Overlapping Detection Module, which we use to automatically determine a person's safety. We input the bounding box and mask that are obtained from Step 2 into the Overlapping Detection Module, then people that traverse the structural supports are detected according to their 'masking' relationship. If the mask of these objects overlap, the unsafe behavior is detected from images.

3.1.1. Mask R-CNN Module

(1) Mask R-CNN Architecture and Training

The Mask R-CNN is similar in nature to the Faster R-CNN, which also adopts a two-stage procedure. The first stage, of the Mask R-CNN network takes an entire image as an input from the ResNet network [58] to extract feature maps. Then, a Region Proposal Network (RPN) is used to propose candidate object bounding boxes from the feature maps, which are extracted from original images. The second stage, requires a RoIAlign layer, described below, to be introduced to preserve and extract spatial locations from each candidate box and perform classification, bounding box regression and mask generation. The parameters used in this research are derived from Kaiming et al. [51].

RoIAlign: Traditional methods for classification tasks mainly output a short vector from fully connected layers [59]. To generate a high-quality segmentation mask from mask branch, each layer is required to maintain a spatial layout of objects. In other detection methods, such as the Faster R-CNN, the RoIPool acts as a standard operation for extracting a small feature map from each RoI. However, quantization can introduce misalignment between the RoI and the extracted features. To maintain the alignment of pixel-to-pixel, the Mask R-CNN adopts a new method, called the RoIAlign, which employs bilinear interpolation to compute the exact values of the input features and avoid any quantization of its boundaries or bins. The RoIAlign can preserve exact spatial

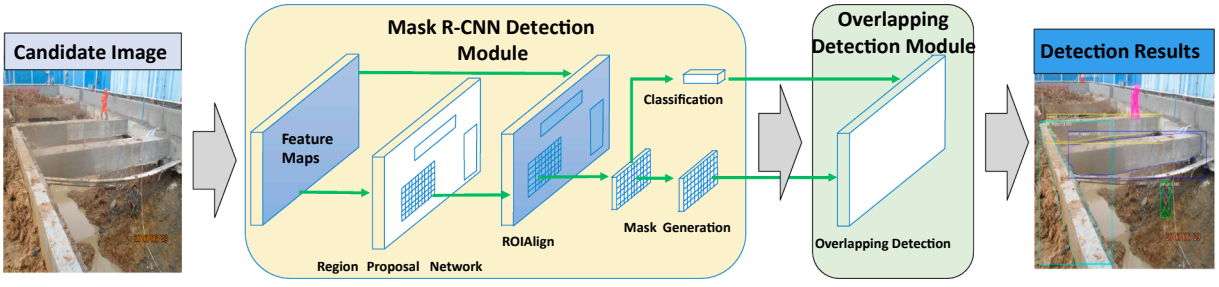

Fig. 2. Workflow for detecting people traversing structural supports.

locations and output reliable masks.

Decoupling: The Mask R-CNN splits the prediction of tasks into mask their generation and classification. So, the network of a Mask branch can separately predict a binary mask for each class. For example, if we want to a detect an individual from an image, their mask and background will be generated. At the same time, an object classification is also performed. Finally, the mask of objects and classification results are simultaneously outputted.

Multi-task Loss: A Mask R-CNN network has three sibling output layers. The first outputs a discrete probability distribution (per RoI), P $(P_0, \ldots, P_k)$ over $K+1$ categories. The second sibling layer outputs bounding-box regression offsets, $t^k = (t_x^k, t_y^k, t_w^k, t_h^k)$, for each of the K object classes, indexed by k. The third sibling layer is mask branch that has a $K \times m \times n$ dimensional output $S = (S_1, \ldots S_k)$ for each RoI encoding K binary masks of resolution $m \times n$, one for each of the K classes.

Each training RoI is labeled with a ground-truth class u, a ground-truth bounding-box regression target v and a ground-truth mask w. We define a multi-task loss L on each labeled RoI to jointly train for classification, bounding-box regression and mask generation:

$$L_{cls}(p, u) + L_{box}(t^u, v) + L_{mask}(s, w) \tag{1}$$

In this approach, the first and second task loss are derived from Kaiming et al. [57]. Thus, the first task loss is expressed as:

$$L_{cls}(p, u) = -\log P_u \tag{2}$$

where log loss for the true class is u

The second task loss is expressed as:

$$L_{box}(t^u, v) = \sum_{i \in \{x,y,w,h\}} smooth_{l1}(t_i^u, v_i) \tag{3}$$

where

$$smooth_{l1}(x) = \begin{cases} 0.5x^2 & if\ |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{4}$$

is defined over a predicted tuple of true bounding-box regression target for class u, $v = (v_x, v_y, v_w, v_h)$, and a predicted tuple $t^u = (t_x^u, t_y^u, t_w^u, t_h^u)$. Here, the parameterization for t^u is derived from Krizhevsky et al. [60]. Notably, x is the value of pixels between the predicted and target position.

The third task loss is expressed as;

$$L_{mask} = -\frac{1}{m \times n} \sum_i \sum_j w^{(i,j)} \log S_u^{(i,j)} \tag{5}$$

is an overage binary cross-entropy loss, L_{mask} is only defined on the kth mask. That is, other mask outputs do not contribute to the loss.

3.1.2. Overlapping detection module

We introduce the Overlapping Detection Module to automatically determine the position of people. If a person is identified as committing an unsafe act, then a warming is delivered to site management. Initially, the Overlapping Detection Module is used to mask all objects in a given image as an input and then divide them into two groups: (1) 'person'; and (2) 'support'. The masks labeled as 'person' are classified into the 'person' list ϕ. The masks labeled as 'steel support' and 'concrete support' are classified in same 'support' list ψ. After these tasks are performed, we select a mask in ϕ and another in ψ. These two candidates are then sent to the Dual Mask Translation (DMT) for further processing, which we describe below. Notably, our scanning mechanism first selects the one element from 'person' (ϕ) to constitute pairs with each of 'support' (ψ) one by one. Then selecting second element from 'person' (ϕ) to each of 'support' (ψ). Therefore, every potential pairing of objects is guaranteed to be identified until all of 'person' (ϕ) are selected.

To identify the safe and unsafe behaviors, we propose an intuitive but effective approach that calculates the overlapping area σ of a couple of masks. Formally, the decision-making process y is formulated as:

$$\begin{cases} y = 1 & if \sigma \geqslant \delta \\ y = 0 & otherwise \end{cases} \tag{6}$$

Here δ is a hyper-parameter and serves as the threshold and σ is the overlapped pixels between 'person' and 'support'.We found, however, many dangerous scenarios were unexpectedly discarded during the selection process. For the purpose of abstaining from missing any possible pairings and refraining from detecting unsafe behaviors by error, we propose two mechanisms to ameliorate the performance of our model.

(1) Dual Mask Translation

In this section, we propose the use of a DMT operation. The DMT takes two masks as inputs and translate them from the 'person' list ϕ by η pixels (η is a hyper-parameter). The motivation of proposing DMT is as follows: The mask of a person and structural supports are often separated when the Mask R-CNN is used to detect these two objects, this is due to nearby pixels are unable to belong to both workers and supports. Therefore, the unsafe behavior of people that traverse the structural supports would be incorrectly recognized as their masks are unable to be to overlapped. To improve the accuracy, we propose the use of a DMT to translate peoples' mask down η pixels to structural support mask. Here, η is set to five pixels according to results of our experiments.

(2) IOU Elimination

The introduction of the DMT provides the ability to improve the performance of the overlapping process. The DMT has, however, a major limitation that we need to acknowledge. An inappropriate setting of η can significantly reduce the distance between a support and person who may be actually standing several meters away. In this instance an incorrect detection can be made. To harness the benefits of the DMT and address this limitation, we introduce the Intersection over Union (IoU) Elimination. The IoU is an evaluation metric that is widely used in object detection tasks. The IOU can be defined as:

$$IOU = Area\ \ of\ \ Overlap / Area\ \ of\ \ Union \tag{7}$$

The IoU enables all the candidate pairs to be separated into two sets by the threshold δ. Thus, the Overlapping Detection Module is used to eliminate unexpected pairs and then proceeds to use the correct sets.

4. Experiment control group studies

The developed Mask R-CNN framework was tested using an experiment to detect workers standing on a structural support. For the purpose of this research we used the Python programming language, which enabled the calculation of matrixes and updating of weights to be performed with a Caffe deep learning framework.

4.1. Experimental design and database establishment

In construction, databases for training deep learning models for the purpose of recognizing unsafe behavior, to our knowledge, have not been created and made available for use. Thus, for the purpose of validating the feasibility our proposed computer vison approach, there is a need to create a database to train and test the model we have developed. Using a monocular camera, we created a dataset of 2018 images that contained images of individuals walking and not walking on structural supports over deep foundation-pits, which were from several construction sites in Wuhan, China. The database was randomly divided into two parts, training and testing, with both containing images of individuals and structural supports. For purpose of training we used 1461 images and testing 450 images. Furthermore, a subset of 107 images containing individuals walking on concrete/steel supports were

Fig. 3. Example of image annotation.

Table 1
Detection results of objects and peoples unsafe behavior (detection with p > 0.8).

Performance metric	Detection of person	Steel support detection	Concrete support detection	Unsafe behavior recognition
Correctly detected (TP)	305	818	423	81
Mis-detected (FP)	0	1	4	27
Not detected (FN)	58	282	102	9
Precision	100%	100%	99%	75%
Recall	84%	74%	81%	90%

Fig. 4. Examples of detection of individuals walking on supports.

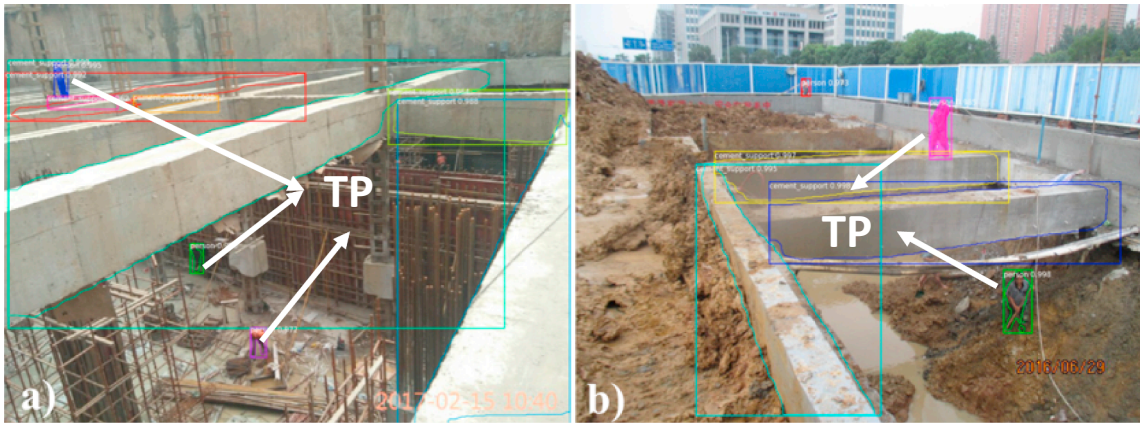

Fig. 5. Examples of detection of an individual walking on supports.

used for testing our proposed model.

To avoid bias, we collected different views, scale, and illuminations for each of the images. The Mask R-CNN was firstly pre-trained by using Microsoft's Common Objects in Context (MS COCO) database, which contains more than 330 k images that can be used for object detection, segmentation, and captioning. Then, a subset of 1461 training images was used to extract and generalize image features to fine tune our pre-trained Mask R-CNN model. A subset of 450 testing images was used for

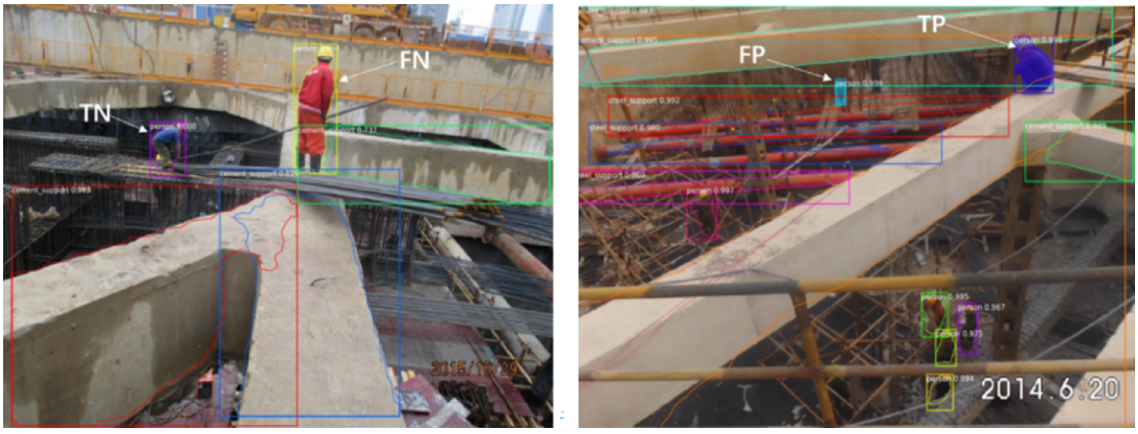

Fig. 6. Examples of missed detections.

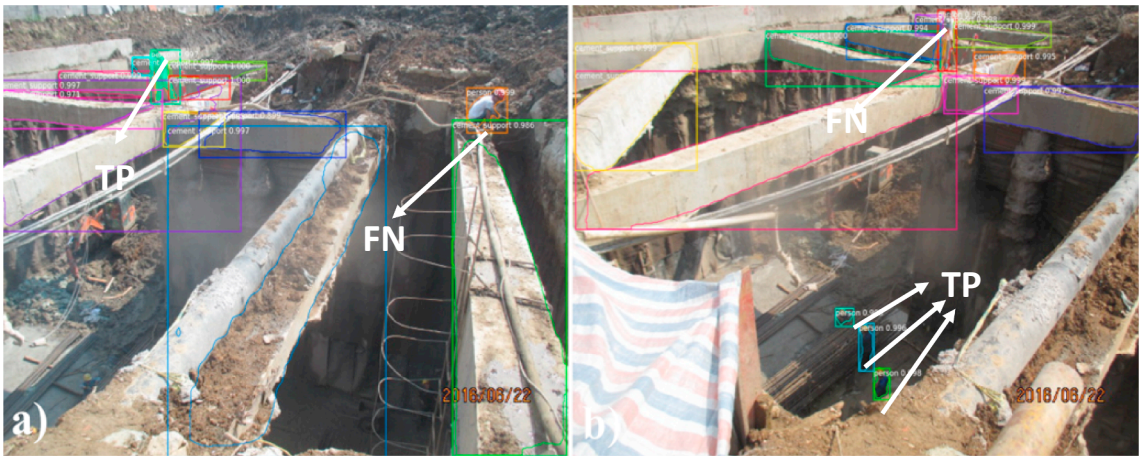

Fig. 7. Examples of missed detections.

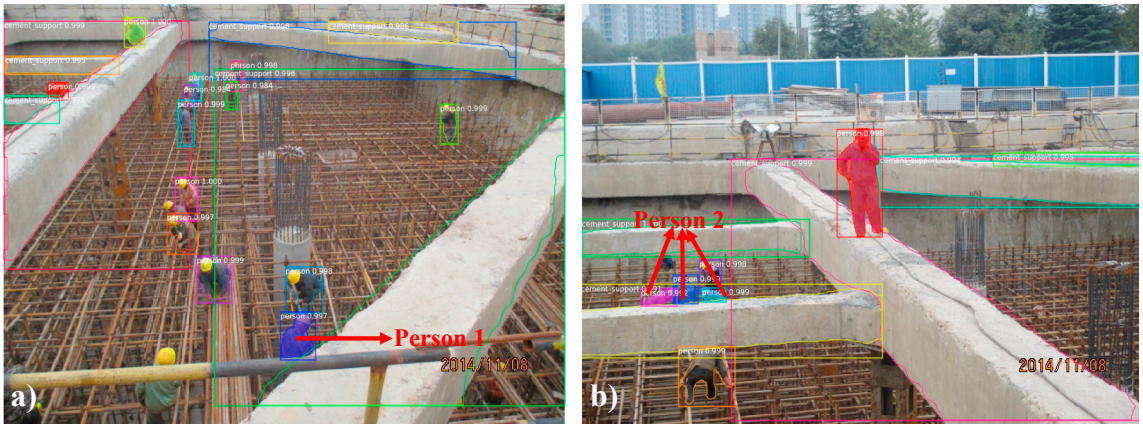

Fig. 8. Examples of people that are occluded by supports.

testing Mask R-CNN model.

4.2. Database annotation

Prior to testing the performance of the developed model to automatically detect people that traverse the structural supports, data needed to be labelled to train the Mask R-CNN network. The "Labelling Tool example Flask App" was used to manually annotate training images, which were also developed using Python. In the Labeling Tool interface, three labels were defined: (1) steel support; (2) concrete support; and (3) person. An example of manual labeling undertaken for people and supports using the "Labelling tool example Flask App" can be seen in Fig. 3.

4.3. Performance evaluation and results

4.3.1. Performance evaluation metrics

The purpose of the evaluation is to test the algorithm's ability to recognize people walking on structural supports over deep foundation-pits from images. We use two key performance indexes the: (1) precision rate; and (2) recall rate. The following equation can be applied to determine the performance indexes:

$$\begin{cases} Precision = \frac{TP}{TP + FP} \\ Recall = \frac{TP}{TP + FN} \end{cases} \quad (8)$$

A "true positive (TP)" refers to an event where the system makes an accurate recognition of a worker traversing a steel support. A "false positive (FP)" occurs when the system recognizes a worker not traversing a structural support as standing/walking on it. However, a "false negative (FN)" may arise when an unsafe behavior cannot be correctly recognized.

4.3.2. Implementation details of proposed approach

We developed the Mask R-CNN module using Python, which was performed on a sever equipped with a 2.5 GHz Intel® Xeon® E5-2680 CPU, a NVIDIA(R) Tesla(TM) K80GPU and 64 RAM. We resized images to 800 pixels to obtain a shorter edge. Each mini-batch has two images per GPU and has 64 samples ROIs with a ratio of 1:3 of positive to negatives. Where an RoI is considered positive if it has an IoU with a ground-truth box of at least 0.5, otherwise it is negative. The model was trained with a learning rate of 0.001, which is decreased by 10 when all layers have been fined tuned. At the same time, a weight decay of 0.0001 and a momentum of 0.9 are used. In this paper, the RPN with five scales (32×32, 64×64, 128×128, 256×256 and 512×512 pixel) and three aspect ratios (1:1, 1:2, and 2:1) are used, with the stride of anchors being set to 1.

4.3.3. Results performance

Table 1 presents the detection results for the: (1) object detection performance; and (2) unsafe behaviour recognition performance. Here we can conclude that our developed model is able to successfully detect people, the structural support (i.e. concrete and steel) and their physical presence on them. The precision and recall rates for detecting people, and structural supports are 100%, 84%; 100%, 74%; and 99%, 81% respectively. The precision and recall rates of detecting people traversing supports are 75% and 90% respectively. Several examples of the detection results are presented in Figs. 4 and 5. When a person's bound box is full of color, it indicates that they are traversing the structural support.

However, some people traversing structural supports were unable to be recognized using this proposed approach. Figs. 6 and 7 present some examples of missed and undetected results.

5. Limitations

While this research provides a contribution to recognizing unsafe behavior on construction sites, it has several limitations that we need to acknowledge and address if effective real-time monitoring can be implemented in practice. Firstly, our study focused on a limited number of activities that were associated with the construction of deep foundation-pits. Thus, the scope of research needs to be extended to examine a wider range of activities and behaviors to enable an effective and reliable system for real-time monitoring to be developed and implemented. Secondly, our model was not able to detect all people traversing structural supports due to the presence of occlusions as we have denoted in Fig. 8. Finally, to improve the process of detection we need larger sample sizes to train and test the Mask R-CNN and prevent 'over fitting'.

6. Conclusions

In this paper we have presented a new computer vision approach to recognize the unsafe behavior of people traversing structural supports used during the construction of deep-pit foundations. The approach that we developed utilized: (1) a Mask R-CNN to detect and segment supports and people; and (2) an Overlapping Detection Module to determine relevant positioning and relationship between people and structural supports. Using an experiment our research demonstrated that the use of computer vision can enable the detection of an individual's unsafe behavior with a high degree of accuracy as the precision and recall rates were found to be 75% and 90%, respectively. Considering the accuracy of our computer vision approach, we believe that there is considerable potential to automatically recognize the unsafe behavior of people, particularly in the case of traversing structural supports. Being able to identify unsafe acts in real-time during construction and lead to perfunctory intervention by management, which can result in immediate behavior modification. In addition, the acquired video can be used to provide people with direct visual feedback and be used as a tool for safety education.

To enact real-time monitoring on-site to mitigate unsafe behavior will require further research to be undertaken to develop an optimum algorithm to improve the generalization of our proposed approach to detect unsafe behavior with varying backdrops that exist. Moreover, by reconstructing a three-dimensional construction model, the spatial information would be computed by detecting construction objects from the as-built three-dimensional construction model. Thus, the accuracy on detecting unsafe behavior would be improved.

Acknowledgments

The authors would like to acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 51878311, No. 71732001, No. 51678265, No. 718210011301059 and China Scholar Council (CSC).

References

[1] P.E.D. Love, P. Teo, J. Morrison, Unearthing the nature and interplay of quality and safety in construction projects: an empirical study, Saf. Sci. 103 (2018) 270–279, https://doi.org/10.1016/j.ssci.2017.11.026.
[2] E.A. Nadhim, C. Hon, B. Xia, I. Stewart, D. Fang, Falls from height in the construction industry: a critical review of the scientific literature, Int. J. Environ. Res. Publ. Health 13 (7) (2016) 638, https://doi.org/10.3390/ijerph13070638.
[3] Technical Specification for Safety Construction of Deep Building Foundation Pits, Ministry of Housing and Urban-Rural Development of People's Public of China, China Architecture & Building Press, 2014.
[4] Technical Code for Safety of Working at Height of Building Construction, Ministry of Housing and Urban-Rural Development of People's Public of China, China Architecture & Building Press, 2016.
[5] M. Zhang, D. Fang, A cognitive analysis of why Chinese scaffolders do not use safety harnesses in construction, Constr. Manage. Econ. 31 (3) (2013) 207–222, https://doi.org/10.1080/01446193.2013.764000.
[6] X.S. Dong, J.A. Largay, S.D. Choi, X. Wang, C.T. Cain, N. Romano, Fatal falls and PFAS use in the construction industry: Findings from the NIOSH FACE reports, Accid. Anal. Prevent. 102 (Suppl. C) (2017) 136–143, https://doi.org/10.1016/j.aap.2017.02.028.
[7] Y.M. Goh, P.E.D. Love, Adequacy of personal fall arrest energy absorbers in relation to heavy workers, Saf. Sci. 48 (6) (2010) 747–754, https://doi.org/10.1016/j.ssci.2010.02.020.
[8] G.J. Fogarty, A. Shaw, Safety climate and the Theory of Planned Behavior: towards the prediction of unsafe behavior, Accid. Anal. Prev. 42 (5) (2010) 1455–1459, https://doi.org/10.1016/j.aap.2009.08.008.
[9] R.M. Choudhry, D. Fang, Why operatives engage in unsafe work behavior: Investigating factors on construction sites, Saf. Sci. 46 (4) (2008) 566–584, https://doi.org/10.1016/j.ssci.2007.06.027.
[10] P.E.D. Love, P. Teo, F. Ackermann, J. Smith, J. Alexander, E. Palaneeswaran, J. Morrison, Reduce rework, improve safety: an empirical inquiry into the precursors to error in construction, Prod. Plann. Contr. 29 (5) (2018) 353–366, https://doi.org/10.1080/09537287.2018.1424961.
[11] P.E.D. Love, S. Veli, P.R. Davis, P. Teo, J. Morrison, 'See the Difference' in a precast facility: changing mindsets with an experiential safety program, J. Constr. Eng. Manage. 143 (2) (2016), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001224.
[12] J. Surry, Industrial accident research: a human engineering approach, J. Occup. Environ. Med. 11 (9) (1969), https://doi.org/10.2307/2518508.
[13] org.cambridge.ebooks.online.book, Human Error, 1990.
[14] J. Bohm, D. Harris, Risk perception and risk-taking behaviour of construction site dumper drivers, Int. J. Occup. Saf. Ergon. 16 (1) (2010) 55–67, https://doi.org/10.1080/10803548.2010.11076829.
[15] R.A. Haslam, S.A. Hide, A.G.F. Gibb, D.E. Gyi, T. Pavitt, S. Atkinson, A.R. Duff, Contributing factors in construction accidents, Appl. Ergon. 36 (4) (2005) 401–415, https://doi.org/10.1016/j.apergo.2004.12.002.
[16] G.A. Howell, T.S. Abdelhamid, P. Mitropoulos, Systems model of construction accident causation, J. Constr. Eng. Manage. 131 (7) (2005) 816–825, https://doi.org/

10.1061/(asce)0733-9364(2005) 131:7(816).
[17] P.E.D. Love, D.J. Edwards, J. Smith, Rework causation: emergent theoretical insights and implications for research, 965275–965275, J. Constr. Eng. Manage. 142 (6) (2016), https://doi.org/10.1061/(ASCE)CO.1943-7862.0001114.
[18] P.E.D. Love, J. Smith, P. Teo, Putting into practice error management theory: unlearning and learning to manage action errors in construction, Appl. Ergon. 69 (2018) 104–111, https://doi.org/10.1016/j.apergo.2018.01.007.
[19] W. Fang, L. Ding, H. Luo, P.E.D. Love, Falls from heights: a computer vision-based approach for safety harness detection, Autom. Constr. 91 (2018) 53–61, https://doi.org/10.1016/j.autcon.2018.02.018.
[20] L. Ding, W. Fang, H. Luo, P.E.D. Love, B. Zhong, Q. Xi, A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory, Autom. Constr. 86 (118) (2018) 124, https://doi.org/10.1016/j.autcon.2017.11.002.
[21] W. Fang, L. Ding, B. Zhong, P.E.D. Love, H. Luo, Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach, Adv. Eng. Inf. 37 (2018) 139–149, https://doi.org/10.1016/j.aei.2018.05.003.
[22] H. Guo, Y. Yu, Q. Ding, M. Skitmore, Image-and-skeleton-based parameterized approach to real-time identification of construction workers' unsafe behaviors, J. Constr. Eng. Manage. 144 (6) (2018) 04018042, https://doi.org/10.1061/(ASCE)CO.1943-7862.0001497.
[23] Y. Yu, H. Guo, Q. Hua, H. Li, M. Skimore, An experimental study of real-time identification of construction workers' unsafe behaviors, Autom. Constr. 82 (2017) 193–206, https://doi.org/10.1016/j.autcon.2017.05.002.
[24] H. Guo, Y. Yu, M. Skitmore, Visualization technology-based construction safety management: a review, Autom. Constr. 73 (2017) 135–144, https://doi.org/10.1016/j.autcon.2016.10.004.
[25] H.W. Heinrich, D. Petersen, N.R. Roos, Industrial Accident Prevention : A Safety Management Approach, 1980.
[26] S. Han, S. Lee, A vision-based motion capture and recognition framework for behavior-based safety management, Autom. Constr. 35 (2013) 131–141, https://doi.org/10.1016/j.autcon.2013.05.001.
[27] L. Ma, R. Sacks, U. Kattel, T. Bloch, 3D object classification using geometric features and pairwise relationships, Comput.-Aided Civ. Infrastruct. Eng. 33 (2) (2017) 152–164, https://doi.org/10.1111/mice.12336.
[28] H.B. Luo, C.H. Xiong, W.L. Fang, P.E.D. Love, B.W. Zhang, X. Ouyang, Convolutional neural network: computer vision-based workforce activity assessment in construction, Autom. Constr. 94 (2018) 281–289, https://doi.org/10.1016/j.autcon.2018.06.007.
[29] R. Sacks, L. Ma, R. Yosef, A. Borrmann, S. Daum, U. Kattel, Semantic enrichment for building information modeling: procedure for compiling inference rules and operators for complex geometry, J. Comput. Civ. Eng. 31 (6) (2017) 04017062, https://doi.org/10.1061/(asce)cp.1943-5487.0000705.
[30] S. Han, S. Lee, F. Peña-Mora, Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing, J. Comput. Civ. Eng. 27 (6) (2013) 635–644, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000279.
[31] K. Yang, C.R. Ahn, M.C. Vuran, S.S. Aria, Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit, Autom. Constr. 68 (2016) 194–202, https://doi.org/10.1016/j.autcon.2016.04.007.
[32] H. Jebelli, C.R. Ahn, T.L. Stentz, Fall risk analysis of construction workers using inertial measurement units: validating the usefulness of the postural stability metrics in construction, Saf. Sci. 84 (2016) 161–170, https://doi.org/10.1016/j.ssci.2015.12.012.
[33] R. Akhavian, A.H. Behzadan, Smartphone-based construction workers' activity recognition and classification, Autom. Constr. 71 (2016) 198–209, https://doi.org/10.1016/j.autcon.2016.08.015.
[34] N.D. Nath, R. Akhavian, A.H. Behzadan, Ergonomic analysis of construction worker's body postures using wearable mobile sensors, Appl. Ergon. 62 (2017) 107–117, https://doi.org/10.1016/j.apergo.2017.02.007.
[35] A. Costin, N. Pradhananga, J. Teizer, Leveraging passive RFID technology for construction resource field mobility and status monitoring in a high-rise renovation project, Autom. Constr. 24 (24) (2012) 1–15, https://doi.org/10.1016/j.autcon.2012.02.015.
[36] T. Cheng, M. Venugopal, J. Teizer, P.A. Vela, Performance evaluation of ultra wideband technology for construction resource location tracking in harsh environments, Autom. Constr. 20 (8) (2011) 1173–1184, https://doi.org/10.1016/j.autcon.2011.05.001.
[37] J. Teizer, D. Lao, M. Sofer, Rapid automated monitoring of construction site activities using ultra-wideband, 24th International Symposium on Automation and Robotics in Construction, 2007, https://doi.org/10.22260/isarc2007/0008.
[38] A. Giretti, A. Carbonari, B. Naticchia, M. DeGrassi, Design and first development of an automated real-time safety management system for construction sites, J. Civ. Eng. Manage. 15 (4) (2009) 325–336, https://doi.org/10.3846/1392-3730.2009.15.325-336.
[39] N. Pradhananga, J. Teizer, Automatic spatio-temporal analysis of construction site equipment operations using GPS data, Autom. Constr. 29 (1) (2013) 107–122, https://doi.org/10.1016/j.autcon.2012.09.004.
[40] J. Hildreth, M. Vorster, J. Martinez, Reduction of short-interval GPS data for construction operations analysis, J. Constr. Eng. Manage. 131 (8) (2005) 920–927, https://doi.org/10.1061/(asce)0733-9364(2005)131:8(920).
[41] J. Teizer, B.S. Allread, C.E. Fullerton, J. Hinze, Autonomous pro-active real-time construction worker and equipment operator proximity safety alert system, Autom. Constr. 19 (5) (2010) 630–640, https://doi.org/10.1016/j.autcon.2010.02.009.
[42] H. Kim, K. Kim, H. Kim, Vision-based object-centric safety assessment using fuzzy inference: monitoring struck-by accidents with moving objects, J. Comput. Civ. Eng. 30 (4) (2016) A04015075, https://doi.org/10.1061/(asce)cp.1943-5487.0000562.
[43] S. Chi, C.H. Caldas, Automated object identification using optical video cameras on construction sites, Comput.-Aid. Civ. Infrastruct. Eng. 26 (5) (2011) 368–380, https://doi.org/10.1111/j.1467-8667.2010.00690.x.
[44] M.W. Park, I. Brilakis, Construction worker detection in video frames for initializing vision trackers, Autom. Constr. 28 (2012) 15–25, https://doi.org/10.1016/j.autcon.2012.06.001.
[45] E. Rezazadeh Azar, B. McCabe, Automated visual recognition of dump trucks in construction videos, J. Comput. Civ. Eng. 26 (6) (2012) 769–781, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000179.
[46] M. Memarzadeh, A. Heydarian, M. Golparvarfard, J.C. Niebles, Real-time and automated recognition and 2D tracking of construction workers and equipment from site video streams, Comput. Civ. Eng. 2012 (2012) 429–436, https://doi.org/10.1061/9780784412343.0054.
[47] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.
[48] J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw. 61 (2015) 85–117, https://doi.org/10.1016/j.neunet.2014.09.003.
[49] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, https://doi.org/10.1109/cvpr.2017.690.
[50] P. VOC, PASCAL VOC, Detection Results: VOC2012, 2017. < http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?cls=mean&challengeid=11&compid=4&submid=9222 > .
[51] G.G. Kaiming He, Piotr Dollár, Ross Girshick, Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), 2017, https://doi.org/10.1109/iccv.2017.322.
[52] M.E. Mneymneh, M. Abbas, H. Khoury, Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications, Front. Eng. Manage. (2018), https://doi.org/10.15302/j-fem-2018071.
[53] J. Holmström, M. Ketokivi, A.P. Hameri, Bridging practice and theory: a design science approach, Decis. Sci. 40 (1) (2009) 65–87, https://doi.org/10.1111/j.1540-5915.2008.00221.x.
[54] M. Pournader, A.A. Tabassi, P. Baloh, A three-step design science approach to develop a novel human resource-planning framework in projects: the cases of construction projects in USA, Europe, and Iran, Int. J. Proj. Manage. 33 (2) (2015) 419–434, https://doi.org/10.1016/j.ijproman.2014.06.009.
[55] R. Weber, Design-Science Research, Research Methods, second ed., Chandos Publishing, 2018, pp. 267–288 https://doi.org/10.1016/B978-0-08-102220-7.00011-X.
[56] K. Peffers, T. Tuunanen, M.A. Rothenberger, S. Chatterjee, A Design science research methodology for information systems research, J. Manage. Inform. Syst. 24 (3) (2007) 45–77, https://doi.org/10.1016/j.accinf.2011.02.004.
[57] M. Chu, J. Matthews, P.E.D. Love, Integrating mobile building information modelling and augmented reality systems: an experimental study, Autom. Constr. 85 (2018) 305–316, https://doi.org/10.1016/j.autcon.2017.10.032.
[58] X.Z. Kaiming He, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, , https://doi.org/10.1109/cvpr.2016.90.
[59] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inform. Process. Syst. (2012) 1097–1105, https://doi.org/10.1145/3065386.
[60] R. Girshick, J. Donahue, T. Darrew, J. Mailk, Rich feature hierarchies for accurate object detection and semantic segmentation, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, https://doi.org/10.1109/cvpr.2014.81.