Predicting Safety Hazards Among Construction Workers and Equipment Using Computer Vision and Deep Learning Techniques

M. Wang[a], P. Wong[a], H. Luo[a], S. Kumar[b], V. Delhi[b], and J. Cheng[a]

[a]Department of Civil and Environmental Engineering, the Hong Kong University of Science and Technology, Hong Kong
b Department of Civil Engineering, Indian Institute of Technology Bombay, India
E-mail: mwangaz@connect.ust.hk, kywongaz@connect.ust.hk, han.luo@connect.ust.hk, sudip.k.m1997@gmail.com, venkatad@iitb.ac.in, cejcheng@ust.hk

Abstract –

The construction industry is one of the most hazardous industries suffering from a high on-site accident rate. A lot of safety hazards result from dynamic activities of construction workers and equipment. Therefore, tracking the location and motion of workers and equipment as well as identifying the interaction between them are crucial to preventing safety hazards on construction sites. Currently, with the extensive installation of surveillance cameras, computer vision techniques can be applied to process the videos and images captured on construction sites, which can be used to monitor site safety and to identify potential hazards. With the aim to predict and prevent the safety hazards among workers and equipment, this paper proposes a methodology to monitor and analyse the interaction between workers and equipment by detecting their locations and trajectories and identifying the danger zones using computer vision and deep learning techniques. First, workers and construction equipment are automatically located from cameras and classified by a deep region-based convolutional neural network (R-CNN) model. Then, the location and classification results are further processed by another CNN-based model to obtain trajectories of those objects. Based on the detection and trajectories, the spatial-temporal relationship between workers and equipment is analysed, from which the danger zones for the workers are identified and the corresponding safety alarms are generated. Experiments are conducted to demonstrate the capability of the proposed methodology for accurately identifying and predicting safety hazards among construction workers and equipment, which can contribute to the safety conditions on construction sites.

Keywords –

Computer vision; Construction site safety; Convolutional neural network (CNN); Deep learning; Object detection; Spatial-temporal relationship; Trajectory tracking.

1 Introduction

The construction industry is one of the most hazardous industries suffering from a high on-site accident rate. Between 2013 and 2017 in Hong Kong, the construction industry has reported the highest fatality rate among 11 industry sections in every year [1]. In particular, striking against or struck by moving objects has been ranked the top 3 highest number of industrial injuries out of 23 accident types [1]. These figures reflect that the construction industry has been significantly hazardous. Specifically, the dynamic characteristic of the construction site activities such as the movement of construction workers and equipment is one of the major causes for construction safety accidents, such as injury of workers due to the surrounding equipment. Therefore, monitoring the interaction among workers and equipment is essential to predict and prevent safety hazards on construction sites.

Currently, construction safety monitoring mainly relies on observing the real-time site conditions manually through on-site surveillance cameras. Early alerts of potential hazards are judged based on previous experiences and provided based on observations from the cameras. Such manual approaches are labor-intensive and error-prone considering the difficulty of monitoring through multiple cameras simultaneously. Human fatigue could lead to the ignorance of potential hazards, such as workers unconsciously approaching heavy equipment. Furthermore, the alerts based on personal experiences can be subjective or belated, leading to severe consequences. Therefore, a method capable of

automatically monitoring and predicting safety issues on construction sites is desired to reduce the resources and to improve the efficiency for safety monitoring.

Computer vision techniques are adopted to automatically process images or videos to assist with various human activities. In addition, deep learning techniques have been widely applied to facilitate computer vision tasks and achieved promising performance. The objective of this study is to automatically predict the safety hazards among construction workers and equipment through analyzing spatial-temporal interaction among workers and equipment using computer vision and deep learning techniques. In the rest of the paper, related works are reviewed in Section 2 and the proposed methodology is introduced in Section 3. Experiments and results are elaborated in Section 4, followed by conclusions and future work in Section 5.

2 Related Work

Previous studies for construction safety monitoring were reviewed. In particular, computer vision-based techniques for construction safety monitoring were categorized by [2] into three aspects.

The first category is object detection. Some researches detected workers and equipment by background subtraction algorithm [3], Histograms of Oriented Gradients (HOG) descriptor with Support Vector Machine (SVM) classifier [4], and Scale-Invariant Feature Transform (SIFT) [5]. The approach of SIFT in [5] segmented a wide range of objects on images covering workers, different kinds of equipment and materials. More recently, Fang et al. [6] used a region-based CNN framework named Faster R-CNN to detect workers standing on scaffolds. A deep CNN then classified whether workers are wearing safety belts. Those without safety belt appropriately harnessed were identified to prevent any fall from height. Adoption of deep learning was shown to achieve promising detection accuracy [6].

The second group of techniques is object tracking. Some studies adopted detection-based tracking method, where by definition newly detected objects either initialize new tracks or are mapped to existing tracks for identity maintenance over certain duration. The DeepSORT developed by [7] is a detection-based tracking model. Zhu et al. [8] used SIFT to extract visual features for detecting workers and equipment. Kalman Filter was then used to predict future movement with respect to past measurement. Another study by [9] detected workers and equipment based on HOG features, while their movement were tracked with Particle Filter. For construction site monitoring, deep learning has not been fully studied for target detection in many detection-based tracking approaches.

The third cluster is action recognition. For example, Ding et al. [10] combined CNN with Long-Short-Term-Memory (LSTM) to identify unsafe actions of workers, such as climbing ladders with hand-carry objects, backward-facing or reaching far. While safety hazards of workers were effectively identified, their method only captured single worker and multi-object analysis was not considered. On the other hand, Soltani et al. [11] used background subtraction to estimate posture of an excavator by individually detecting each of its three skeleton parts including dipper, boom and body. Although knowing the operating state of construction equipment would allow safety monitoring nearby, the influence of the equipment on the surrounding objects was not studied [11].

Overall, in terms of automated safety monitoring with computer vision techniques, previous studies focused on different parts accounting for the safety issues separately, such as identifying working status of construction equipment or tracking the movement of workers. There is a lack of an integrated analysis of the spatial-temporal interaction among workers and equipment considering the potential influences from different aspects. A robust mechanism that analyzes the spatial-temporal interaction among workers and equipment is desired for automated and real-time monitoring of on-site safety.

3 Methodology

In this study, an integrated approach is proposed for predicting safety hazards among construction workers and equipment using computer vision and deep learning techniques. The proposed approach involves three parts, as shown in Figure 1. First, images are extracted from videos, and construction workers and equipment in video frames are extracted by the detection model. Then, the detection results are imported into the second part. For construction equipment, the danger zone is identified

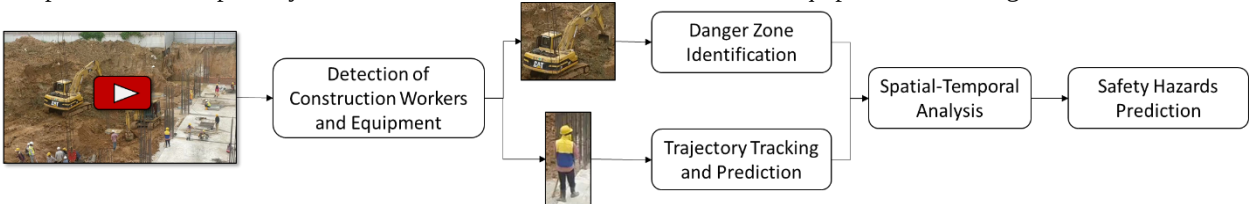

Figure 1. Overall workflow for predicting the safety hazards among construction workers and equipment

while for construction workers, the trajectory is tracked and predicted. By combining the identified danger zone and predicted trajectory, the spatial-temporal relationship among construction workers and equipment is analyzed. In the end, safety hazards are predicted based on the spatial-temporal relationship analysis.

3.1 Equipment and Worker Detection

The detection of equipment and workers are realized through Faster R-CNN, in view of its higher accuracy demonstrated by previous studies [12,13]. The architecture of Faster R-CNN includes (1) backbone network to extract image features; (2) region proposal generate (RPN) network for generating region of interest (ROI), and (3) classification network for producing class scores and bounding boxes for objects. Details of each part are introduced as follows.

3.1.1 Backbone Network

The backbone of the Faster R-CNN is used for feature extraction through a stack of convolution, activation, and max pooling layers. After pre-processing, each image is fed into the network as a three-dimensional array representing the pixel values on the RGB channels. In the beginning, a certain number of filters are assigned with random weights as initialization. During convolution, each filter slide across the input volume and the dot product between the filter and the corresponding image patch convolved is calculated and added with a bias value to obtain the convolution result. To add non-linearity to the network, the convolution result is fed into an activation function, for which rectified linear units (ReLU) is used in this study. After activation, max pooling is performed to reduce the dimension of feature maps by selecting the maximum value from each image patch covered by the filter and use the maximum value as the new feature value. Through setting the max pooling filter size and stride, the dimension of feature maps is down-sampled by a factor of 2 each time, such that the computational cost is reduced. A stack of convolution, ReLU and max pooling layers are performed to generate the feature maps, which are then fed into the RPN network for generating candidate ROIs (i.e. the potential regions with objects). As training the model from scratch is time-consuming and requires a large number of images, transfer learning is applied in this study as the low-level features of objects such as edges and corners are transferable among different datasets. Specifically, the weight of a model namely VGG16, which is pre-trained on another large dataset, is utilized to initialize the backbone network. The model is then fine-tuned with our dataset of construction equipment and workers such that high-level features (i.e. the characteristics of equipment and workers) are learned in the applied model.

3.1.2 RPN Network

The generated feature maps from the backbone network are passed through the RPN network which consists of a convolutional layer and a ReLU layer. The results are then fed into two parallel layers for label classification and bounding box regression. The class labels include foreground and background, indicating whether an object is contained in the region or not while the bounding box regression layer is used for refining the location of bounding boxes. In the beginning, a number of small windows, named anchors, with different sizes and aspect ratios are designed at each pixel location such that multi-scale features can be extracted. During the training, the class scores and the four coordinates of each bounding box are predicted for each anchor, after which the loss for both class scores and coordinates are computed. The number of anchors is reduced by non-maximum suppression (NMS) based on the foreground scores.

3.1.3 Classification Network

After obtaining the potential ROIs from the RPN, the regions on the feature maps corresponding to the ROIs are extracted through a crop pooling method. The extracted regions of the feature maps are reshaped and fed into the classification network, where two parallel layers are used for predicting the class labels with probability of each class and bounding boxes with more accurate locations. There are 7 classes included in this study including 5 types of equipment, the worker and the background. In the end, the loss of the predicted classes and the bounding boxes are computed based on the ground truth labels. The model weights are then updated by back-propagation during training process.

The layers of the three networks are combined together and the weights are trained in an end-to-end manner. After training the model on our dataset, the model is capable of detecting construction equipment and workers accurately during inference using the optimal trained weights.

3.2 Worker Trajectory Tracking

For trajectory tracking and prediction, the DeepSORT framework proposed by Wojke et al. [7] is used as our baseline because it demonstrated robustness of identity preservation upon arbitrary duration of occlusion. As a detection-based tracking framework, the worker detection results at a frame either initialize new tracks or are mapped to the most similar identities being tracked. Kalman Filter is used to predict future position of the target for position proximity matching. Apart from position constraint, identity assignment of the set of new detection also considers appearance similarity against the targets being tracked. Readers are referred to the original

publication for more detailed mathematical formulation of DeepSORT.

There are several parameters in DeepSORT that define its track management mechanism. Two of them are highlighted since their values are adapted to our site analytics. 'Lambda' controls the relative influence of position and appearance constraints in identity matching. For our analysis, Lambda is set as 0.5 for a complementary effect among these two factors. On one hand, construction sites tend to be crowded such that many workers may walk very close to others. Position proximity alone may make identity matching inaccurate when there are many candidates near a target. On the other hand, construction workers tend to have very similar appearance when wearing reflective vest and safety helmet, such that appearance may not explicitly distinguish multiple worker. Therefore, a balanced reliance among these two aspects is granted.

Another parameter 'Max_age' is set to be 200 frames in our study. This means that unique identity of a disappeared worker is remembered for 200 consecutive frames, during which it can be retrieved upon reappearance of the target. A long memory increases the computational burden for handling many targets, while a short memory leads to discontinuous trajectory history and subsequently inaccurate prediction of future trajectory. A 200-frames memory is considered reasonable for obtaining trajectory history. Acquiring complete trajectory history would support the prediction of future trajectory.

3.3 Worker Trajectory Prediction

Trajectory prediction for workers is based on the inference output from Kalman Filter used in DeepSORT. The translation of the 'foot' position of a target is considered, since it represents the area on which he/she walks. Therefore, the bottom center coordinate of a box is obtained by relating the corner coordinates with the width and height of a box. Since Kalman Filter only infers target movement at next timestamp, a mechanism is proposed to predict movement in a longer future. In particular, the future position of the 'foot' of a target (x_2, y_2) is linearly extrapolated with respect to its current position (x_1, y_1) and velocity:

$$x_2 = x_1 + (\alpha_x * \dot{x}) \quad (1)$$

$$y_2 = y_1 + (\alpha_y * \dot{y}) \quad (2)$$

where $\dot{x}$ and $\dot{y}$ are velocities along individual direction output by Kalman Filter, α_x and α_y are number of future frames for position prediction.

With the current and future positions, $P = (x_1, y_1)$ and $P' = (x_2, y_2)$, respectively, the predicted trajectory of a target is defined to be the directed line segment PP' pointing from P to P'. On the image plane, this trajectory line is represented by a linear equation, while the coordinates of P and P' define its boundary. An assumption is that targets tend to produce smooth motion within the inference period, like standing still or walking constantly with current velocities $\dot{x}$ and $\dot{y}$.

3.4 Definition of Danger Zone

When a construction equipment (e.g. excavator) is in its working state, there is an activity region around the construction equipment. There is a high potential of safety hazards when workers or other construction machines enter this activity region. In this study, this kind of activity region is defined as a danger zone.

For an excavator, the exact danger zone could be defined as a located circle in a real construction site, which is an ellipses when projected to 2-D images, based on danger parameters including the position (pos) of the excavator, the lengths (l), working directions (d) and ranges (r) of arms, the size (s) of its body, and other factors (o). An exact danger zone (edz) is defined in Equation (3)

$$edz = f(pos, l, d, r, s, o) \quad (3)$$

On a construction site with a working schedule for on-site machines, most danger parameters of a construction machine are deterministic during a certain time interval, such as the lengths, working directions and ranges of arms, and the size of the construction equipment. On the other hand, the location of the excavator is obtained from the bounding box coordinates through the equipment and worker detection model, as introduced in Section 3.1. With those danger parameters, the size of the danger zone around an excavator could be determined.

3.5 Spatial-Temporal Analysis

Safety status of each worker is categorized into 3 types, based on the geometric relationship between the predicted trajectory and danger zone. As shown in Figure 2, 'Normal' is marked if the starting point of a trajectory (current position) is outside any danger zone while the predicted trajectory does not touch any danger zone (e.g. T1, T6). 'Danger Now' is marked as long as the start point lies within a danger zone (e.g. T2, T3). 'Potential Danger' is marked if the start point is outside any danger zone while the predicted trajectory intercepts a danger zone at least once (e.g. T4, T5, T7).

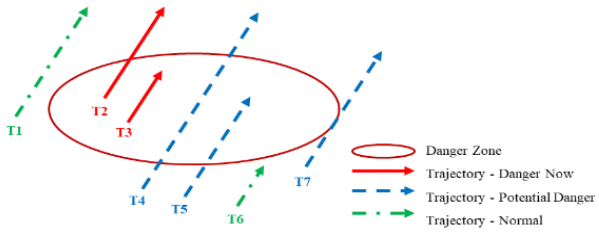

Figure 2. Examples of categorizing safety statuses based on predicted trajectory of a worker

The mathematical relationship between an ellipse and a straight line is derived to check for interception. Equation (4) and Equation (5) show the general forms of an ellipse and an infinite straight line on x-y plane respectively, while Figure 3 summarizes the notation of all related symbols.

$$\frac{(x-x_c)^2}{{r_x}^2}+\frac{(y-y_c)^2}{{r_y}^2}=1 \quad (4)$$

$$y=\frac{y_2-y_1}{x_2-x_1}*x+y_0 \quad (5)$$

$$\text{where, } y_0=y_1-\left(\frac{y_2-y_1}{x_2-x_1}\right)x_1 \quad (6)$$

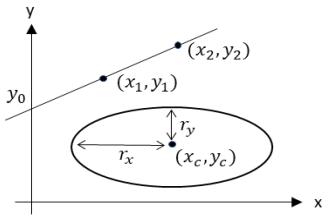

Figure 3. Notation for general representation of ellipse and straight line

A danger zone is centered at (x_c, y_c) with radiuses r_x and r_y. As for the predicted trajectory of a worker, an infinite straight line is defined by the current position as (x_1, y_1) and future position as (x_2, y_2). To find the intercept(s) between an ellipse and a straight line, Equation (4) and Equation (5) are solved simultaneously for x and y. After substitution, a quadratic equation is obtained, as shown in Equation (7).

$$Ax^2+Bx+C=0 \quad (7)$$

$$A=\frac{1}{{r_x}^2}+\frac{m^2}{{r_y}^2} \quad (8)$$

$$B=-\frac{2x_c}{{r_x}^2}+\frac{2m(y_0-y_c)}{{r_y}^2} \quad (9)$$

$$C=\frac{{x_c}^2}{{r_x}^2}+\frac{(c-y_c)^2}{{r_y}^2}-1 \quad (10)$$

$$m=\frac{y_2-y_1}{x_2-x_1} \quad (11)$$

The solution(s) to Equation (7), if any, correspond to the x-coordinate(s) at which the straight line intercepts the ellipse. The expression in Equation (12) is used to check whether interception exist. If $\Delta \geq 0$, the line either passes through the ellipse at two distinct points, or touches the ellipse at one point. In this case, Equation (13) provides the corresponding intercepted x-coordinate(s). If $\Delta < 0$, the line is always outside the ellipse.

$$\Delta=B^2-4AC \quad (12)$$

$$x^+=\frac{-B+\sqrt{\Delta}}{2A} \; and \; x^-=\frac{-B-\sqrt{\Delta}}{2A} \quad (13)$$

The key consideration when checking interception is that each trajectory is a definite segment spanning from the current to future position of the target. It is possible that the interception point(s) is/are outside the definite segment which is only a portion of its associated infinite line. Therefore, an additional condition is imposed to check whether the definite segment intercepts the ellipse. Table 1 illustrates the decision logic. In case of no interception at all, 'Normal' is assigned to a target (e.g. T1 in Figure 2). Otherwise, if the current position lies between the intercepted points, it lies within the danger zone and 'Danger Now' is assigned (e.g. T2, T3 in Figure 2). If the current position lies outside the zone while at least one of the intercepted point(s) is within the definite segment, the trajectory passes through the zone and 'Potential Danger' is assigned (e.g. T4, T5, T7 in Figure 2). Otherwise, the trajectory does not enter the zone at all and 'Normal' is assigned (e.g. T6 in Figure 2).

Table 1. Algorithm for assigning safety status based on worker trajectory and danger zone

If $\Delta < 0$:
Assign '*Normal*'
Else:
If $x_1 \in [x^-, x^+]$:
Assign '*Danger Now*'
Else if $x^- \in [x_1, x_2]$ or $x^+ \in [x_1, x_2]$:
Assign '*Potential Danger*'
Else:
Assign '*Normal*'

4 Experiments and Results

Experiments are performed to demonstrate the capability of the proposed methodology for automatically predicting safety hazards on construction sites through the analysis of spatial-temporal relationship between construction workers and equipment based on the captured images and videos from surveillance cameras. The spatial-temporal relationship is analysed based on two inputs: (1) the danger zone obtained from the detected location of the construction equipment and other

parameters, and (2) the predicted worker trajectory based on the historical trajectory records.

4.1 Experiment Dataset

Several videos captured from surveillance cameras on construction sites are collected and images are extracted from those videos. 2410 images containing 5 types of construction equipment (i.e. dump trucks, excavators, loaders, mixer trucks, and rollers) and construction workers are extracted and each image is annotated with ground truth labels and bounding boxes. 90% of the images are used for model training and validation while 10% are for model testing. In the end, the methodology is also applied on a new construction site video to demonstrate the real-time prediction of the hazards.

4.2 Experiment Implementation

The danger parameters of a construction equipment are assumed to be known in the experiment. Based on this assumption, the size of the danger zone around a construction equipment such as excavator is pre-defined. Then, the exact danger zones on a construction site are determined with the result of equipment detection, and the identified danger zones are used for spatial-temporal analysis.

Firstly, the architecture of the model for construction worker and equipment detection, as introduced in Section 3.1, is constructed using Pytorch, which is a common platform for implementing deep learning models. The model is trained using the annotated images for 40 epochs and the training loss is plotted to monitor the learning progress of the model. The model is evaluated using average precision (AP) for each class and the mean AP (mAP) for all the classes.

As for the worker trajectory prediction introduced in Section 3.3, α_x and α_y are both set to be 60 frames, such that for each target his/her position after 60 frames is inferred. This is a reasonable period because it predicts the target movement in the next 2 seconds, if considering a typical video with 30 frames per seconds. For safety monitoring on construction sites, 2-second movement prediction would allow identification of potential hazards.

4.3 Experiment Results and Analysis

The results include two parts – (1) the accuracy of equipment and worker detection; (2) the prediction results of safety hazard based on the analysis of spatial-temporal relationship among workers and equipment.

4.3.1 Accuracy of the detection model

The accuracy of the Faster R-CNN model on the testing dataset is summarized in Table 2. The model achieved high detection accuracy for both workers and construction equipment. The AP values of all the classes achieved at least 85% and even exceed 95% for most classes except for dump trucks, rollers and workers. With a mAP of 92.55%, the model is demonstrated to be promising for detecting workers and equipment accurately on the construction site.

Table 2. Accuracy of worker and equipment detection

Class Name	AP (%)	mAP (%)
Dump truck	85.12	
Excavator	95.94	
Loader	96.19	
Mixer truck	97.73	
Roller	86.91	
Worker	93.40	
		92.55

Nevertheless, as shown in Figure 4, the detection model tended to miss the excavator when its arm was not hanged horizontally, and those workers squatting or carrying objects. These cases could be attributed to our limited training dataset, which may not have covered a variety of view angles or human gesture. An enriched dataset would further enhance the detection accuracy.

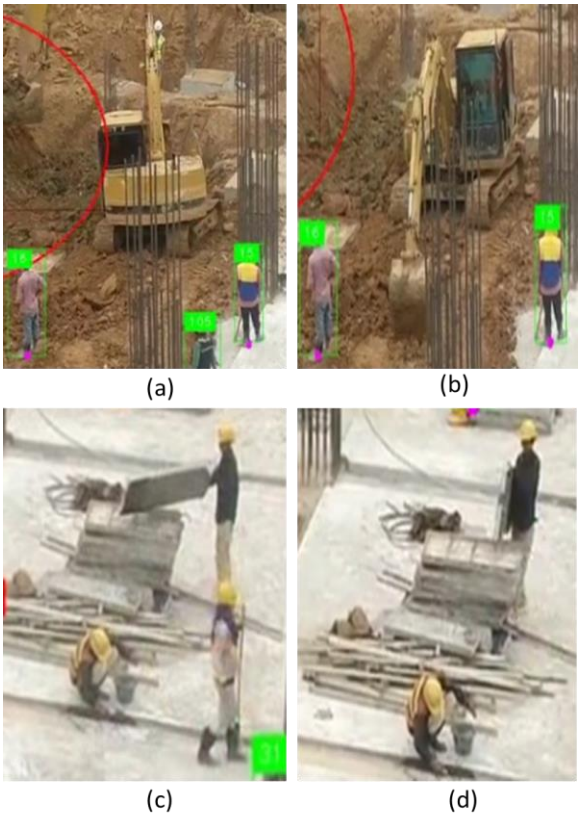

Figure 4. Examples of incorrect detection of equipment and workers

4.3.2 Accuracy of the spatial-temporal analysis

The detection model was applied on a construction site video to detect the equipment and workers. Trajectories of workers were also tracked and predicted. Danger zones were then defined to monitor safety statuses of workers based on spatial-temporal relationship among their trajectories and the danger zones. As shown in Figure 5, an ellipse around the excavator bounded a danger zone. In addition, historical trajectories of workers were displayed in purple, while their predicted trajectories after 60 frames were visualized as blue arrows. Safety statuses were then assigned to individual worker, based on the mathematical formulation in Section 3.5. Different safety statuses were categorized by colorized bounding boxes.

Figure 5. Examples of identified danger zones and different safety statuses of workers

The accuracy of assigning safety statuses is evaluated by Average Precision of Status (AP_{status}), as defined by Equation (14).

$$AP_{status} = \frac{TS}{TP} \tag{14}$$

where TS counts the number of workers assigned with true safety status, TP counts the number of detected workers as true positives from the detection model. In our experiment, assignment of safety statuses achieved an 87.45% AP_{status}. This suggests that our framework accurately revealed the safety status of individual worker against the danger zones. For example, in Figure 5, worker number 15 was alerted in red with 'danger now' since he lied within the danger zone, while worker number 16 was labelled in blue as 'potential danger' since he was about to enter the danger zone.

Nevertheless, dangerous conditions were sometimes not identified. As shown in Figure 6, red alerts were not issued even when workers 15 and 16 stood close to the excavator, possibly because the equipment was not detected and hence no danger zone was defined. Moreover, blue alert was not issued to the worker walking towards the operating area of excavator, possibly because he was occluded by other objects. These cases reflect that the accuracy of assigning safety status heavily relies on the performance of equipment and worker detection. The spatial-temporal analysis would be further supported with a more robust detection model and an enriched training dataset.

(a) (b)

Figure 6. Examples of incorrect prediction of safety statuses

5 Conclusion and Future Work

Construction industry is reported to be the most hazardous with a high rate of accidents on construction site. There are various dynamic activities on construction sites such as the operation of various construction equipment and the movement of workers. The interaction between construction workers and equipment is one important reason resulting in on-site safety hazards. Therefore, to avoid potential safety hazards, it is necessary to monitor the working status of construction workers and equipment, and analyse the spatial-temporal interactions between them. Currently, on-site conditions are monitored and analysed manually from the surveillance cameras, which is labour-intensive and error-prone. Furthermore, the alerts for safety hazards may be subjective and belated, leading to severe consequences.

In this study, an integrated approach based on computer vision and deep learning techniques is proposed to predict safety hazards on construction site

through the spatial-temporal analysis of construction workers and equipment. Specifically, Faster R-CNN is first applied to detect construction workers and equipment from surveillance videos. Based on the detection results, danger zones of the construction equipment are identified while worker trajectories are tracked and predicted as well. Finally, the spatial-temporal interaction between the danger zone and the predicted worker trajectory is analysed, based on which the potential hazards are predicted and corresponding alerts are issued.

Experiments are performed with videos captured from surveillance cameras on construction sites to validate the capability of the proposed methodology. The detection model obtained high accuracy on detecting workers and equipment, with a mAP of 92.55% and AP values above 95% for most classes. As for the spatial-temporal analysis, the precision of assigning safety statuses to workers achieved 87.45%, based on which the safety hazard alerts are provided. The experiment results demonstrated that the proposed approach is capable of automatically predicting potential safety hazards through detecting construction workers and equipment, identifying danger zones, tracking and predicting worker trajectories, and analysing spatial-temporal interactions on construction sites. Even though there are still some negative examples in experiments, the overall experiment results demonstrate promising performance of the proposed integrated approach for predicting safety hazards among construction workers and equipment on construction sites. Future work will focus on exploring better methods to obtain danger zones around construction equipment and to improve the performance of the proposed approach.

References

[1] Occupational Safety and Health Statistics Bulletin, Hong Kong, 2018. https://www.labour.gov.hk/eng/osh/pdf/Bulletin2017.pdf (accessed February 4, 2019).

[2] J. Seo, S. Han, S. Lee, H. Kim, Computer vision techniques for construction safety and health monitoring, Adv. Eng. Informatics. 29 (2015) 239–251. doi:10.1016/j.aei.2015.02.001.

[3] S. Chi, C.H. Caldas, Automated Object Identification Using Optical Video Cameras on Construction Sites, Comput. Civ. Infrastruct. Eng. 26 (2011) 368–380. doi:10.1111/j.1467-8667.2010.00690.x.

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

[5] H. Kim, K. Kim, H. Kim, Data-driven scene parsing method for recognizing construction site objects in the whole image, Autom. Constr. 71 (2016) 271–282. doi:10.1016/j.autcon.2016.08.018.

[6] 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. doi:10.1016/j.autcon.2018.02.018.

[7] N. Wojke, A. Bewley, D. Paulus, Simple online and realtime tracking with a deep association metric, in: Proc. - Int. Conf. Image Process. ICIP, 2018: pp. 3645–3649. doi:10.1109/ICIP.2017.8296962.

[8] Z. Zhu, M.W. Park, C. Koch, M. Soltani, A. Hammad, K. Davari, Predicting movements of onsite workers and mobile equipment for enhancing construction site safety, Autom. Constr. 68 (2016) 95–101. doi:10.1016/j.autcon.2016.04.009.

[9] Z. Zhu, X. Ren, Z. Chen, Integrated detection and tracking of workforce and equipment from construction jobsite videos, Autom. Constr. 81 (2017) 161–171. doi:10.1016/j.autcon.2017.05.005.

[10] L. Ding, W. Fang, H. Luo, P.E.D. Love, B. Zhong, X. Ouyang, A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory, Autom. Constr. 86 (2018) 118–124. doi:10.1016/j.autcon.2017.11.002.

[11] M.M. Soltani, Z. Zhu, A. Hammad, Skeleton estimation of excavator by detecting its parts, Autom. Constr. 82 (2017) 1–15. doi:10.1016/j.autcon.2017.06.023.

[12] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149. doi:10.1109/TPAMI.2016.2577031.

[13] J.C.P. Cheng, M. Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171. doi:10.1016/j.autcon.2018.08.006.