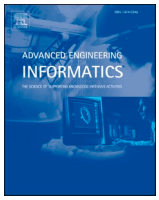
Advanced Engineering Informatics 50 (2021) 101400

Contents lists available at ScienceDirect

Advanced Engineering Informatics

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



Computer vision and long short-term memory: Learning to predict unsafe behaviour in construction

Ting Kong [a,c], Weili Fang [b,c,d,*], Peter E.D. Love [d], Hanbin Luo [c], Shuangjie Xu [e], Heng Li [a]

[a] Department of Building and Real Estate, Faculty of Construction and Environment, Hong Kong Polytechnic University, Hong Kong Special Administrative Region, China
[b] Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Dr., Singapore 117566, Singapore
[c] School of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China
[d] School Civil and Mechanical Engineering, Curtin University, Perth, Western Australia 6845, Australia
[e] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong Special Administrative Region, China

ARTICLE INFO

Keywords:
Deep learning
Computer vision
Long-short term memory
PNpoly algorithm
Unsafe behaviour

ABSTRACT

Predicting unsafe behaviour in advance can enable remedial measures to be put in place to mitigate likely accidents on construction sites. Prevailing safety studies in construction tend to be retrospective and focus on examining the conditions that contribute to unsafe behaviour from a psychological perspective. While such studies are warranted, they can also not visually comprehend the dynamic and complex conditions that influence unsafe behaviour. In this paper, we aim to contribute to filling this void and, in doing so, combine computer vision with Long-Short Term Memory (LSTM) to predict unsafe behaviours from videos automatically. Our proposed approach for predicting unsafe behaviour is based on: (1) tracking people using a SiamMask; (2) predicting the trajectory of people using an improved Social-LSTM; and (3) predicting unsafe behaviour using Franklin's point inclusion polygon (PNPoly) algorithm. We use the Wuhan metro project as a case to evaluate our approach's feasibility and effectiveness. Our adopted SiamMask method outperforms current techniques used for tracking people. Additionally, our improved Social-LSTM can achieve higher accuracy on trajectory prediction than other methods (e.g., Social-GAN). The research findings demonstrate that our developed computer vision approach can be used to accurately predict unsafe behaviour on construction sites.

1. Introduction

Construction has been repeatedly identified as one of the most dangerous industries worldwide [1–4]. In China, for example, in 2019, 904 people were killed during the construction of housing and municipal engineering projects [5]. It is widely understood that unsafe behaviour is the major contributor to accidents on construction sites worldwide [6,7]. Despite, however, the wealth of research that has sought to understand and mitigate unsafe behaviours in construction, they still inevitably materialise, and accidents remain a pervasive problem [8,9].

A new line of inquiry has emerged in construction, focusing on using digital technologies to manage safety performance in real-time [10–14]. The upshot here is that people can be provided with feedback instantaneously, enabling immediate behavioural modifications. Technologies such as computer vision and deep learning, for example, have been implemented to detect if people are wearing their personal protective equipment (PPE) (e.g., safety harness and hard hat) and using the correct posture when lifting items [12,15–20]. A close examination of these studies reveals that attention has solely focused on detecting unsafe behaviour while eschewing the development of a predictive solution-based people's past data.

Traditionally root cause models juxtaposed with the psychological theories (e.g., Theory Reasoned Action (TRA) and Theory of Planned Behaviour (TPB)) have formed the basis to predict people' unsafe behaviour [8,21] as they help us understand how people's behaviour changes under differing working conditions. The models assume that people's behaviour is planned; hence, it predicts deliberate behaviour [22]. However, they are "widely applied without sufficient attention paid to what makes [them] work in its contexts of origin, and without adequate customisation for the specifics" [23]. To this end, root cause models provide a "flawed reductionist view" of safety issues [23]. In a similar vein, machine learning techniques such as artificial neural

* Corresponding author at: Department of Building, School of Design and Environment, National University of Singapore, 4 Architecture Dr., Singapore 117566, Singapore
E-mail address: weili.fang@curtin.edu.au (W. Fang).

https://doi.org/10.1016/j.aei.2021.101400
Received 31 December 2020; Received in revised form 7 August 2021; Accepted 19 August 2021
Available online 6 October 2021

networks (ANN) have also been used to predict unsafe behaviour [24–26]. Studies of this nature do not reflect actual practice as they are often underpinned by literature reviews (e.g., *meta*-analysis). Thus, the data used for predictive modelling is usually derived from studies of unsafe behaviour rendering their predictability and relevance to represent practice questionable.

If we are to make headway to reducing unsafe behaviour on construction sites, then immediate intervention is needed to alert people in real-time of their actions [1,14]. Data acquisition in real-time can provide managers with a better understanding of the nuances and conditions that result in unsafe behaviour occurring on-site. Predictive solutions, such as Radio Frequency Identification (RFID) tags and Global Positioning System (GPS), may provide more accurate results than computer vision-based systems. However, such approaches require people to wear sensors at all times and thus are difficult to implement in practice. Furthermore, many sub-contractors often work across multiple sites and are employed temporarily (e.g., hired labour). Therefore, it is unfeasible to issue a sensor to every person on a given site. In this instance, vision-based approaches are a more practical option for examining unsafe behaviour on construction sites.

Against this backdrop, we aim to develop and present a novel computer vision approach in this paper that can automatically predict people's likely unsafe behaviour by monitoring their actions in real-time. We use a SiamMask to track people in real-time to determine their location on-site to achieve our goal. Then, based on people locations, we develop an improved Social-Long Short-term Memory (LSTM) [27], a form of Recurrent Neural Network (RNN), to learn about their movements. Once we can understand and learn people's movements, we can predict their future trajectories. Finally, the relationships between predicted trajectories and hazardous zones are determined using Franklin's point inclusion in the polygon (PNPoly) algorithm [28]]. Thus, people's likely unsafe behaviour can be inferred and predicted.

We commence our paper by briefly reviewing computer vision and its use to manage safety in construction (Section 2). Then, we introduce and describe our developed research methodology (Section 3). A case study is then used to validate our proposed approach (Section 4). This is followed by identifying the paper's contributions (Section 5) before presenting our conclusions, limitations, and avenues for future research (Section 6).

2. Computer vision and safety

Understanding why and how unsafe behaviour occurs has been a mainstay of safety studies in construction [26,29–31]. For example, in the case of TRA, it assumes that "the behaviour under investigation is under volitional control, that is, that people believe that they can execute the behaviour whenever they are willing to do so" [32]. The TRA was designed to predict volitional behaviours, but its explanatory power is limited. The TPB was extended TRA by including perceived behaviour control to account for internal and external constraints on behaviour [22]. While TPB has enabled us to understand unsafe behaviour in construction better, it only provides a retrospective view of people's actions and may be subject to hindsight bias. As we have indicated above, computer vision and deep learning can overcome these shortcomings. However, there has been no empirical-based work that has yet to predict unsafe behaviour in real-time. It is outside the scope of this paper to provide a detailed review of the use of computer vision and its role in managing safety in construction, as this can be found in the work of Fang *et al.*[1,14].

2.1. Vision-based object detection and segmentation

Computer vision and deep learning approaches have been successfully used for object detection and segmentation for a wide range of purposes in construction, particularly within the context of safety. For example, Fang *et al.*[33] applied a Region-based Convolutional Neural Networks (R-CNN) to detect people and excavators from images. The accuracies to detect people and excavators were 91% and 95%, respectively. Similarly, Fang *et al.* [13] applied a Mask R-CNN, an object segmentation approach, to accurately segment structural support from images. Likewise, An *et al.* [34] developed a large-scale image database with 13 categories of objects (e.g., people and tower cranes). Then seven object detectors (e.g., Yolov3, Faster R-CNN) and four instance segmentation detectors (e.g., Mask R-CNN) are used by An *et al.*[34] to detect and segment the objects. A brief overview of research applying object detection and segmentation in construction can be seen in Table 1.

Methods used for object detection in construction, such as the Faster R-CNN [40], Single Shot-multi-box Detector (SSD) [41], and YOLO [42], utilise single-shot and two-stage detectors, which process entire images and output multiple detections. Such object detections rely on anchor boxes to optimise the speed and efficiency of sliding window detection.

Single-shot detectors can achieve faster detection speeds than their two-stage counterparts. The most popular single-shot models are the SSD and YOLO and their variants. As shown in the work of Luo et al. [35], They used Yolov3 to identify people and plant (i.e., moving or stationary) in a hazardous area. One of the most popular two-stage detection approaches used in construction is the Faster R-CNN due to its ability to balance accuracy and detection speed. Similarly, the Masked R-CNN is widely used for object segmentation [13,43].

2.2. Vision-based object tracking

Visual object tracking (VOT) is used to generate an object's trajectory (e.g., person and equipment) over time by locating its position obtained from videos. Point and kernel tracking have been used widely to track objects on construction sites [16,44]. A particle filter, which uses a sequential Monte Carlo approach based on the particle representation of probability densities, is often used with a point tracking method [16,44]. Contrastingly, kernel tracking methods are typically performed by computing the motion of an object, which is represented by a primitive object region [18]. It has been used to track multiple people or equipment on construction sites effectively and efficiently [45]. Except, when computer vision and deep learning are used to track people on construction sites, the accuracy of results is far superior to point and kernel tracking [10,46].

Most VOT tracking approaches (e.g., deep learning-based, point tracking and kernel tracking) use a rectangular bounding box to initialise the target and estimate its position in the subsequent frames [48]. Such approaches often fail to represent an object from cluttered backgrounds and when a target object moves (i.e., changes its size due to distance) [47,48]. A binary segmentation mask has been developed to address the limitations above and improve tracking performance [48–50]. Noteworthy, tracking objects using a segmentation mask requires more computational power than a simple bounding box-based approach. A fully convolutional Siamese framework (SiamFC)

Table 1
Examples of object detection and segmentation in construction.

Task	Methodology	Target Objects	Author
Object Detection	You Only Look Once (Yolo)	People and excavator	[35]
	RetinaNet	Excavator	[36]
	Yolov3	People	[18]
	Faster R-CNN	People and excavator	[33]
	Fast R-CNN	22 classes of objects	[17]
Object Segmentation	Fully Convolutional Network	Cracks	[37]
	Mask R-CNN	Concrete surface bug hole	[38]
	Mask R-CNN	Structural supports and people	[13]
	Mask R-CNN	Floorboards, joists, air vents and pipework	[39]

developed to track objects with speed in real-time, dubbed the SiamMask, can be used to overcome the requirement for increased computational power [48]. As the SiamMask outperformed the current state-of-the-art tracking approaches tested at the Visual Object Tracking challenge VOT-2018, we employ this approach as the backbone of the model we develop in our study.

3. Proposed framework

Our research framework focuses on: (1) tracking people using a SiamMask; (2) predicting the trajectory of people using an improved Social LTSM; and (3) predicting unsafe behaviour using Franklin's point inclusion of a PNPoly algorithm. We acknowledge studies have used computer vision to detect and segment objects from image/video, and the merits of both the Faster R-CNN and Masked R-CNN have been widely espoused. Thus, our paper focuses on the use of a SiamMask, Social LTSM and a PNPoly algorithm. We describe each of the procedures performed below.

3.1. Tracking

As mentioned above, a SiamMask approach is adopted in this research to simultaneously achieve higher tracking performance and real-time detection speed [48]. Chopra *et al.* [51] first proposed the Siamese network, which was used for face recognition. A sample pair, exemplar image z and search image x , was input to the two branches of the network, where they shared the same weights. Then, the Euclidean distance determines if the exemplar and search images belong to the same category. The SiamFC represents the tracking problem as a similarity marching learning in the embedding space. The structure of SiamFC is presented in Fig. 1.

The SiamFC compares exemplar image z with search image $\times$ to obtain a dense response map. The two inputs are processed by the same networkf_θ, then output two feature maps. Finally, a cross-correlation between these two feature maps is performed, which generates a dense response map, as noted in Eq. [1].

$$g_\theta(z,x) = f_\theta(z) * f_\theta(x) \tag{1}$$

3.1.1. Network architecture and model training

The structure of SiamMask is presented in Fig. 2. A ResNet-50 network is used as a backbone network (f_θ) [53]. To obtain a high spatial resolution in deeper layers, we reduce the output stride to 8 using convolutions with stride 1 [48]. As noted in Fig. 2, the exemplar and search images are inputs into the backbone network (f_θ). The outputs are separated feature maps (15*15*256 and 31*31*256). Then, these two feature maps are used to generate the response map through cross-correlation. Here, we refer to each element of the response map as 'Response of a candidate Window' (ROW). Subsequently, the ROWs are sent into two 1×1 convolutional layers, one with 256 and the other with 63^2 channels, and output three-branch including mask, box generation, and score. The strategy of merging the low- and high-resolution features produces a more accurate object mask. The bounding box is generated from the obtained mask *via* a rotated minimum bounding rectangle [48,54].

The output of the model includes mask, bounding box, and score. Hence, the training loss in this research contains three parts, which the following equations can compute.

$$L_{loss} = \lambda_1 L_{mask} + \lambda_2 L_{score} + \lambda_3 L_{box} \tag{2}$$

$$L_{mask} = \sum_n \left(\frac{1+y_n}{2wh} \sum_{ij} log\left(1 + e^{-c_n^{ij} m_n^{ij}}\right) \right) \tag{3}$$

$$L_{score} = -logp_x \tag{4}$$

$$L_{box} = \sum_{i=0}^{3} smooth_{L_1} = \sum_{i=0}^{3} smooth_{L_1}(\delta[i], \sigma) \tag{5}$$

$$smooth_{L_1}(x,\sigma) = \begin{cases} 0.5\sigma^2 x^2 & |x|. < \dfrac{1}{\sigma^2} \\ |x| - \dfrac{1}{2\sigma^2} & |x| \geq \dfrac{1}{\sigma^2} \end{cases} \tag{6}$$

where, λ_1, λ_2, and λ_3 are hyper-parameters [48,55], set to 32, 1 and 1, respectively; y_n is the ground-truth binary label of each ROW, wherey_n is set to 1 following Wang *et al.*[48]; c_n^{ij} The pixel-wise ground-truth label corresponds to a pixel (i, j) of the object mask in the n-th candidate ROW; w and h are the sizes of a pixel-wise ground-truth mask.

3.2. Trajectory prediction

After tracking people in real-time, the next step is to predict their working trajectory. A Social-LSTM is adopted as it achieved state-of-the-art methods tested on the database, for example, the ETH [56] and UCY [27,57]. The Social-LSTM considers the people-people interaction in predicting their trajectory, improving the robustness and accuracy of multi-people tracking. The Kalman filter is adopted to correct the Social-LSTM results to enhance the robustness of the prediction method. A Kalman filter is an optimal estimator that can infer parameters of interest from indirect, inaccurate and uncertain observations [58,59], which can estimate the people's walking based on their historical trajectory. Therefore, we correct the Social-LSTM results with the Kalman filter once a person's predicted walking speed is inconsistent with the

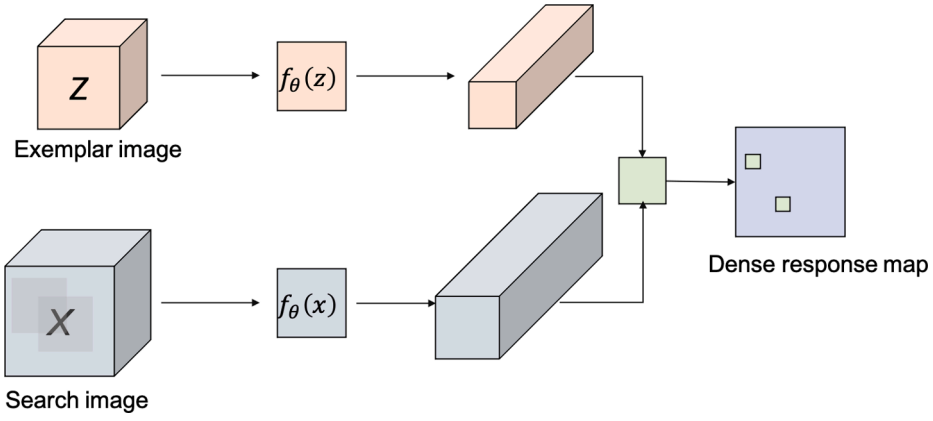


Fig. 1. Structure of a SiamFC.
Adapted from [52]

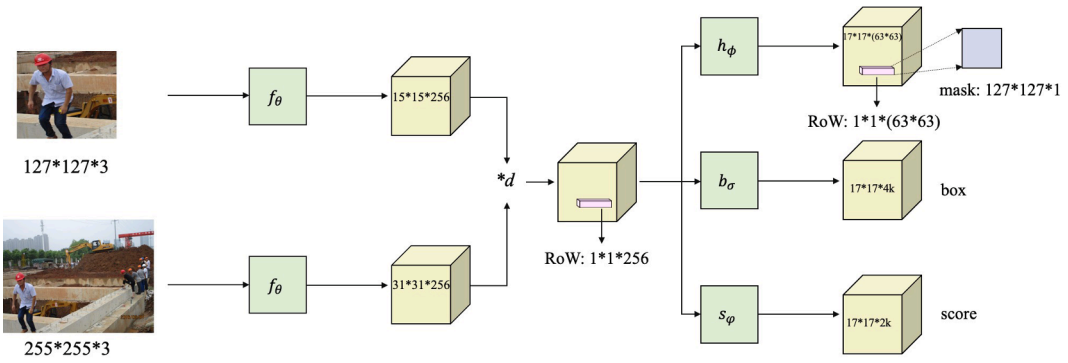


Fig. 2. Structure of a SiamMask.
Adapted from [60]

recorded speed. The process of improved Social LSTM is presented in Fig. 3.

A social pooling layer is added between time-step t and time-step $t+1$. The output is a three-dimensional state tensor. The two dimensions are plane coordinates, and the third dimension is the state tensor output by the Social-LSTM model at time-step t. Thus, the target person's trajectory can be predicted by gathering the Social-LSTM information of the neighbours'. The procedure is as follows:

Step 1: Capture the hidden state information (h_i^t) of the i-$^{\text{th}}$ person at time-step t;

Step 2: Share hidden state information with neighbours by building a 'social' hidden-state tensor H_t^i, as noted in Eq. [7]:

$$H_t^i(m,n,:) = \sum_{j \in N_i} 1_{mn}\left[x_t^j - x_t^i, y_t^j - y_t^i\right] h_{t-1}^j \tag{7}$$

where, D is the hidden-state dimension; N_o is the neighbourhood size; h_{t-1}^j is the hidden state of the Social-LSTM corresponding to the j^{th} person at time-step t-1; $1_{mn}[x,y]$ is an indicator function to check if (x,y) is in the (m,n) cell of the grid; N_i is the set of neighbours corresponding to the person i.

Step 3: Embedding the pooled social hidden-state tensor into the vector a_t^i and embedding coordinates into vector e_t^i, as noted in Eq. [8], Eq. [9] and Eq. [10].

$$h_i^t = LSTM\left(h_i^{t-1}, e_i^t, a_t^i, W_l\right) \tag{8}$$

$$a_t^i = \phi\left(H_t^i; W_a\right) \tag{9}$$

$$e_t^i = \phi\left(x_t^i, y_t^i; W_e\right) \tag{10}$$

where, $\phi(\bullet)$ is an embedding function with ReLU non-linearity; W_e and W_a embedding weights; W_l is the Social-LSTM weight.

3.2.1. Trajectory estimation

The hidden-state information at time-step t is used to predict the distribution of the trajectory position $(\widehat{x},\widehat{y})_{t+1}^i$ at the next time-step $t+1$ [27]. Following Graves [60], a bivariate Gaussian distribution parametrized is assumed with the mean $\mu_{t+1}^i = (\mu_x,\mu_y)_{t+1}^i$, standard deviation $\sigma_{t+1}^i = (\sigma_x,\sigma_y)_{t+1}^i$ and correlation coefficient ρ_{t+1}^i. The predicted coordinates $(\hat{x}_t^i,\hat{y}_t^i)$ at time-step t can be computed by flowing Equations.

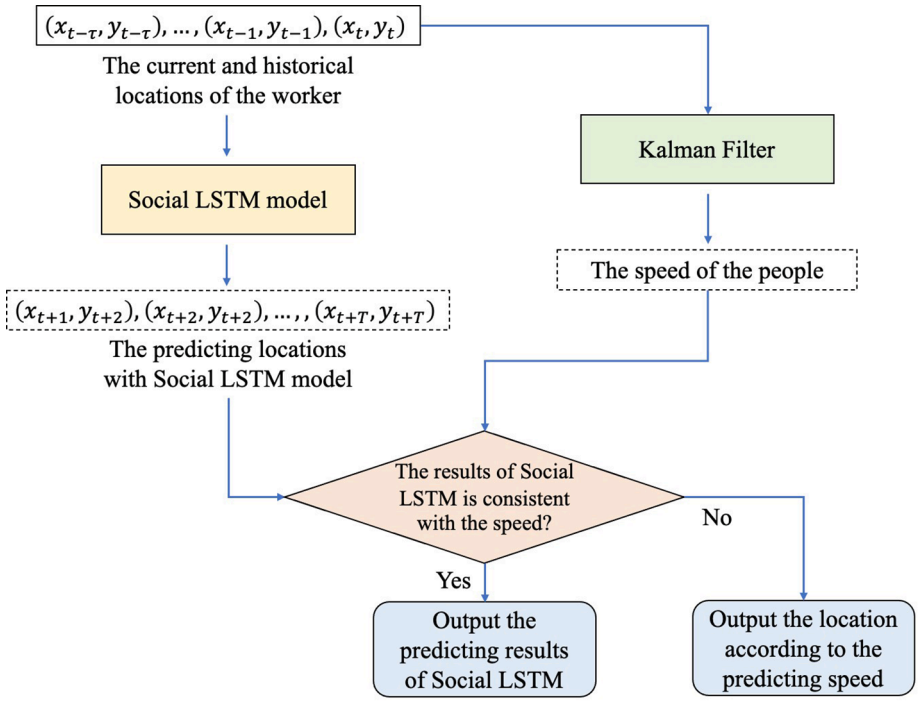


Fig. 3. Process of improved Social LSTM.

$$(\widehat{x},\widehat{y})_t^i \sim \mathrm{N}\left(\mu_t^i,\sigma_t^i,\rho_t^i\right) \tag{11}$$

$$\left[\mu_t^i,\sigma_t^i,\rho_t^i\right] = W_p h_i^{t-1} \tag{12}$$

$$L^i\left(W_e,W_l,W_p\right) = -\sum_{t=T_{obs}+1}^{T_{pred}} log\left(\mathbb{P}\left(x_t^i,y_t^i\middle|\sigma_t^i,\mu_t^i,\rho_t^i\right)\right) \tag{13}$$

where, L^i is the i^{th} trajectory.

3.3. Prediction of unsafe behaviour

After predicting a person's future trajectory, a PNPoly algorithm determines if their coordinates are in the hazardous area so that unsafe behaviour can be envisaged. As shown in Fig. 4, if the predicted trajectories (P_m and P_k) are within the polygon, then a person's behaviour will be envisaged as unsafe.

The PNPoly algorithm tests whether a point is inside a polygon (convex or concave) by counting how many times the ray from the test point crosses its edge. If the count is an odd number, the point is in the polygon area; otherwise, it is outside.

4. Experiment and results

The Wuhan Metro project is selected and used to evaluate the feasibility and effectiveness of our proposed approach. We specifically focus on foundation pits and predict people approaching their edge. Our previous study adopted a Mask R-CNN to detect structural support [13]. Consequently, we reframe from explaining 'how' the Mask R-CNN is implemented in this case. Instead, we limit the scope of our research to tracking and predicting a person's trajectory.

4.1. Data collection

In collaboration with a contractor involved with constructing the Wuhan subway, several Close-Circuit Television (CCTV) cameras with two million pixel high-definition were installed on construction sites. The video data were instore in a web-based real-time monitoring system (Fig. 5). All image/video data used for testing our model are obtained from the real-time monitoring system.

4.2. Tracking performance

Three key performance indicators (KPIs) are used to evaluate the tracking performance of our approach: (1) mean intersection over union (mIOU); (2) mean average precision (mAP)@0.5 IOU; and (3) mAP@0.7 IOU. The AP is the area under the precision and recall curve. Here, the IOU, precision and recall can be computed in Eq. [14], Eq. [15], and Eq. [16]:

$$IOU = \frac{Detectionresult \cap Groundtruth}{Detectionresult \cup Groundtruth} \tag{14}$$

$$precision = \frac{TP}{TP+FP} \tag{15}$$

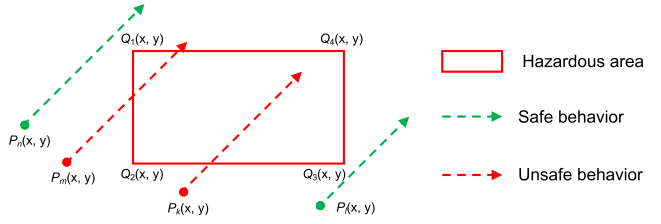


Fig. 4. Example of safe and unsafe behaviour based on the predicted trajectory.

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

where, A 'true positive' (TP) refers to a person who has been correctly tracked. A 'false positive (FP)' occurs when a tracked worker is some other object. A 'false negative (FN)' refers to a failure in tracking a person in the image.

The SiamMask model was pre-trained on the DAVIS database [61]. Then a video from a real-time monitoring system is collected for testing. The duration of this video is 222 s, including 5550 frames and includes four people. Fig. 6 presents an example of people being tracked using the SiamMask approach. The mIOU, mAP@0.5, and mAP@0.7 tracking these four people are 73.4%, 92.9%, and 68.1%, respectively. The testing speed of tracking each person is 0.018 s.

4.2.1. Evaluation comparison

Two representative methods, the ASLA (visual tracking via Adaptive Structural Local Sparse Appearance Model) [62] and SCM (tracking via a Sparse Collaborative appearance Model) [63], are compared to evaluate the tracking performance of our model as they have been found to provide high levels of accuracy [64]. The average sequence overlap score (AOS) and tracking length (TL) in Xiao and Zhu [64]'s work for ASLA and SCM were 84%, 72%, and 83%, 68%, respectively. In this comparison study, the parameters of ASLA and SCM accord with Jia *et al.* [62] and Zhong *et al.* [63], respectively. The comparison study results are presented in Table 2, where it can be seen that our adopted SiamMask approach improves our ability to track people on construction sites.

4.3. Trajectory prediction performance

Two key performance indictors (KPIs) are used to evaluate the trajectory prediction performance: (1) average displacement error (ADE); and (2) final displacement error (FDE). These two KPIs can be computed in Eq. [17] and Eq. [18]:

$$ADE = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{t_{pred}}\sum_{t=t_{obs}+1}^{t_{obs}+t_{pred}}\sqrt{\left(x_i^t-\widehat{x}_i^t\right)^2+\left(y_i^t-\widehat{y}_i^t\right)^2} \tag{17}$$

$$FDE = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\left(x_i^{t_{pred}}-\widehat{x}_i^{t_{pred}}\right)^2+\left(y_i^{t_{pred}}-\widehat{y}_i^{t_{pred}}\right)^2} \tag{18}$$

A video from our real-time monitoring system (Fig. 5) is selected and used to validate the performance of Social-LSTM for trajectory prediction. In this experiment, the hyperparameter of the Social-LSTM is set as follows: the spatial pooling size (N_0) is 32, the pooling window size is 8*8, and the learning rate is 0.003.

The ADE and FDE of our model to 1 s, 2 s, 3 s, 4 s, 5 s, 6 s, 7 s, 8 s, 9 s, and 10 s are used to examine the performance of our model to predict future trajectory. The results presented in Fig. 7 indicate that our improved Social-LSTM can achieve higher accuracy (e.g., ADE and FDE) on trajectory prediction except at the length of 1 s. The ADE and FDE for our improved method at 1 s were 7.019 and 9.696 pixels, respectively. While the ADE and FDE for the Social-LSTM at 1 s were 6.303 pixels and 11.505 pixels. Therefore, we can suggest that our improved Social-LSTM achieved robust results for long-term prediction. The testing speed of prediction of each person's trajectory is 0.030 s.

4.3.1. Evaluation comparison

Two deep learning models, the Social-LSTM and GANs are used to examine the tracking performance [65]. The Social-GAN model has been previously used by Kim *et al.* [66] to predict people's paths. The hyperparameters of the Social-LSTM and GANs used in this comparison study were followed by Alahi *et al.* [27] and Gupta *et al.* [65], respectively. The results of these three models are presented in Table 3, and we can conclude that our improved Social-LSTM method outperforms both the

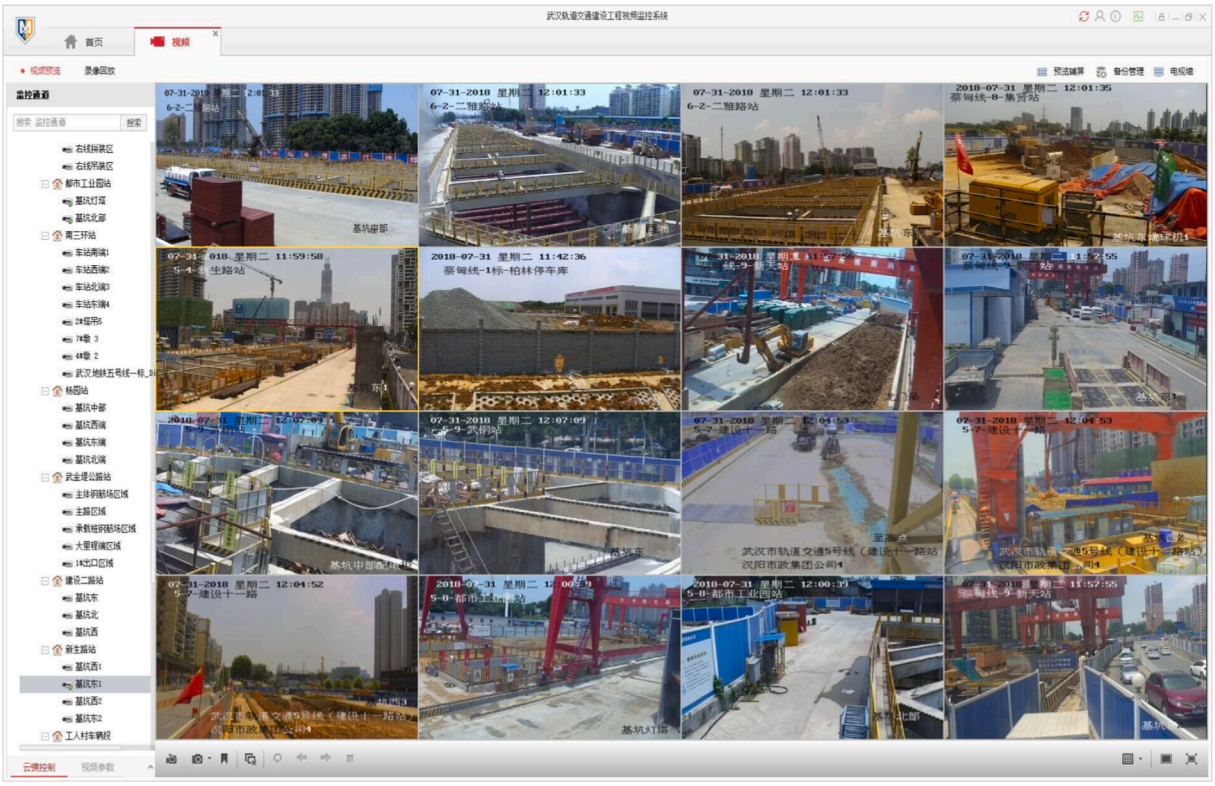


Fig. 5. A web-based real-time monitoring system.

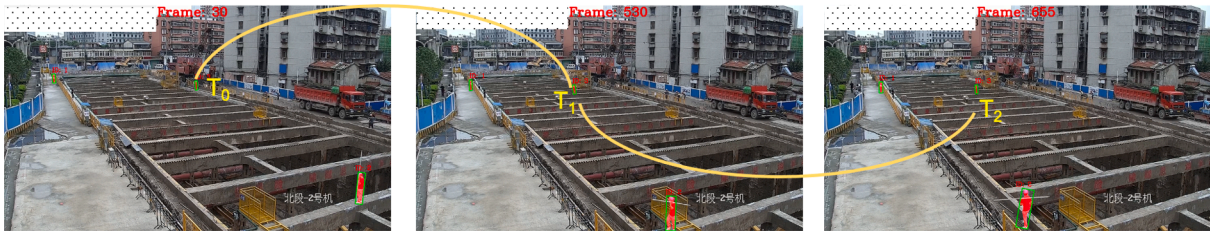


Fig. 6. Example of people tracking results.

Table 2
Comparison of results to track people.

Methods	mIoU					mAP@0.5					mAP@0.7				
	#1*	#2	#3	#4	mean	#1	#2	#3	#4	mean	#1	#2	#3	#4	mean
ASLA	0.453	0.061	0.649	0.621	0.399	0.523	0.055	0.924	0.914	0.545	0.233	0.031	0.328	0.712	0.184
SCM	0.275	0.102	0.556	0.68	0.368	0.344	0.087	0.653	0.864	0.431	0.062	0.08	0.173	0.553	0.174
SiamMask	0.734	0.682	0.779	0.755	**0.734**	0.906	0.873	0.987	0.942	**0.929**	0.677	0.550	0.797	0.731	**0.681**

* Note: #1, #2, #3, and #4 are the identifier of people in our video; "mean" is the average result of these four people.

Social-LSTM and GANs.

Fig. 8 present an example of the trajectory prediction by using these three methods. Here it can be seen that the: (1) blue points are the ground truth of frames before current time; (2) purple points are the ground truth of frames after current time; (2) yellow points are predicted by using our improved Social-LSTM; (4) the cyan points are predicted by the Social-LSTM; and (5) red points are predicted by the Social-GAN.

4.4. Unsafe behaviour prediction results

A testing database is used to evaluate the effectiveness and feasibility of our approach that was used to predict unsafe behaviour. The hazardous area is manually defined and labelled, as noted in Fig. 9. An example of unsafe behaviour prediction is presented in Fig. 9. Based on the people's predicted trajectories, we observe that two people (ID2 and ID3) have entered a hazardous area, while person #4 is predicted to arrive. Notably, our approach fails to predict unsafe behaviour as their trajectories are not correctly determined. Possible reasons influencing the performance trajectory are occlusions and their distance from the video camera.

5. Discussion

Predicting unsafe behaviour in real-time is an area of research that

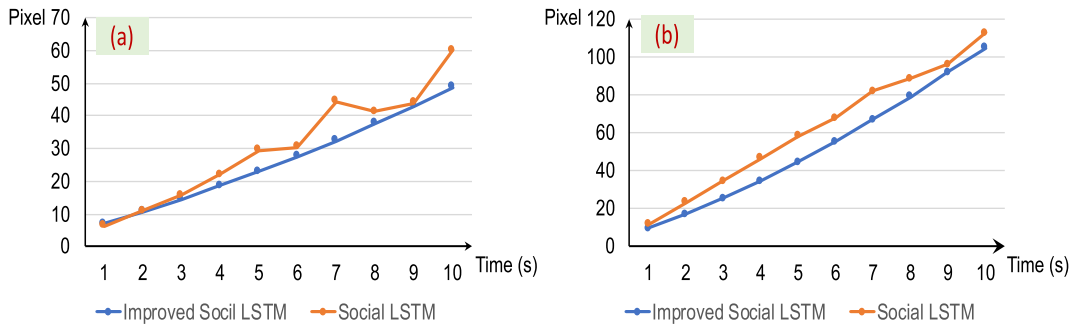


Fig. 7. Results of trajectory prediction: (a) ADE; (b) FDE.

Table 3
Result of comparison study on trajectory prediction.

Model	1.6 s		2.4 s	
	ADE (pixel)	FDE (pixel)	ADE (pixel)	FDE (pixel)
Social LSTM + Kalman filtering	9.165	**14.161**	**12.132**	**20.470**
Social-LSTM	**9.014**	18.535	12.739	27.740
Social-GAN	12.23	19.64	16.15	27.35

has yet to receive attention in construction. The ability to predict unsafe behaviour enables remedial measures to be put in place to mitigate potential accidents, yet this has been a challenge as information used for decision-making is retrospective. Our research addresses this problem by utilising computer vision and deep learning and, therefore, can be used to re-examine the relevancy of TPB to construction as direct observations of behaviour can now be made in real-time. For example, research has shown safety incidents generally materialise while performing rework [3,31]. In this instance, unplanned work may result in unplanned behaviour, not planned as expected under the TPB. Thus, viewing behaviour in real-time opens doors for us to challenge prevailing theories such as TPB and develop new ones that reflect actual practice in dynamic and complex work environments.

In sum, the contributions of the research we have presented in this paper is threefold. Firstly, our computer vision and deep learning framework can automatically determine unsafe behaviour by predicting people's trajectories. In doing so, managers can take action immediately to prevent accidents from occurring. For example, site managers can send out warning alerts using sensors to a person entering a hazardous area [67]. Our proposed framework also provides a solution to enhance people's safety awareness by notifying them of their unsafe action and its likely outcome if not addressed. We acknowledge that predicting unsafe behaviour is a complex and challenging task. Indeed, the

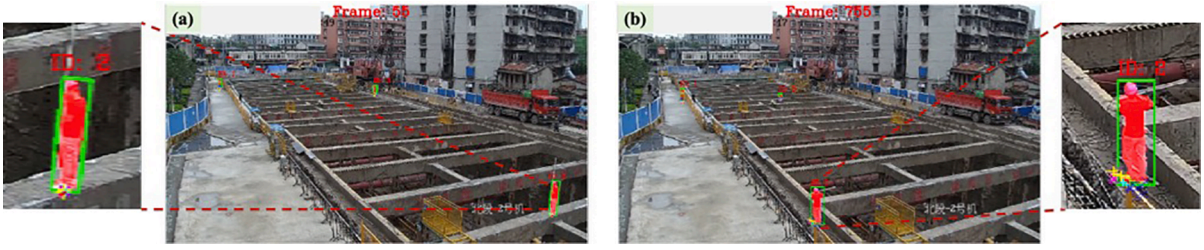


Fig. 8. Example of people trajectory prediction results: (a) T = frame 55; (b) T = frame 755.

Fig. 9. Examples of unsafe behaviour prediction result.

management of safety is a wicked problem. But, our approach can enable management and the site workforce to work together to prevent injuries and accidents. We stress that our model provides a 'means to an end' and is not a panacea for addressing and determining the conditions that contribute to unsafe behaviour. Our future research will focus on the activity-level unsafe behaviour prediction to further understand why unsafe behaviour is occurring.

Secondly, we empolyed empl a SiamMask approach to track people in construction sites. As we have previously noted, current tracking approaches (e.g., point tracking or deep learning-based) in construction cannot represent an object from cluttered and dynamic backgrounds. Our employed SiamMask approach produced a pixel-wise mask in each frame while tracking a person's trajectory. The results presented in our experiment demonstrated its effectiveness and feasibility to track people while they are moving. Additionally, we compared our approaches to tracking performance with the ASLA and SCM and shown it is more accurate.

Finally, we improve the ability to predict a person's trajectory performance by integrating a Social-LSTM with Kalman filtering. We corrected the Social-LSTM results with the Kalman filter once the predicting walking speed is not consistent with the historical speed. The results presented in Table 3 have demonstrated that our improved method can outperform the Social-LSTM and GANs.

6. Conclusion

This study proposes a computer vision approach with long short-term memory to automatically predict people' unsafe behaviour from videos in an early time manner. Our proposed approach consisted of: (1) tracking people using SiamMask; (2) predicting the trajectory of people using improved Social-LSTM; and (3) predicting unsafe behaviour using a PNPoly algorithm.

The case of the Wuhan railway construction project was selected to evaluate the feasibility and effectiveness of our proposed approach. The mIOU, mAP@0.5, and mAP@0.7 of our adopted tracking model are 73.4%, 92.9%, and 68.1%, respectively, outperforming ASLA and SCM methods. The experiment presented in this study demonstrated that our proposed approach could accurately track and predict people's trajectory so that unsafe behaviour can be accurately predicted. Predicting unsafe behaviour early during construction can help site managers take action for immediate behaviour intervention.

While we have developed and presented a novel way of predicting unsafe behaviour, our research has several limitations that need to be acknowledged. Firstly, the hazardous area was manually defined and labelled and remain static, but they can change as site works progress. However, we suggest that this limitation will be addressed in our future work by integrating the object segmentation methods (e.g., Mask R-CNN) with our model.

Secondly, we have been unable to develop the database to train our Social-LSTM and SiamMask models. Our model was pre-trained using public database (e.g., DAVIS), then our database from construction sites was used to test the model's performance. While this limitation may have affected our model's accuracy, we suggest it can be overcome by creating an extensive video database, which we aim to do in future research. Finally, our approach was tested on a few scenarios (e.g., people approaching a foundation pit without safety protection) on construction sites. Thus, our model is not generalizable, but by examining different scenarios across a broader range of construction sites, we can address this issue in our forthcoming studies.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

The authors would like to acknowledge National Natural Science Foundation of China (Grant No.71732001, No. 51978302). The authors would like to thank the Editor and three anonymous reviewers for the constructive comments. Additionally, access to the data and the deep learning codes presented in this paper may be made available upon request from the corresponding author.

References

[1] W. Fang, P.E.D. Love, H. Luo, L. Ding, Computer vision for behaviour-based safety in construction: A review and future directions, Adv. Eng. Inf. 43 (2020) 100980, https://doi.org/10.1016/j.aei.2019.100980.
[2] J.W. Hinze, J. Teizer, Visibility-related fatalities related to construction equipment, Saf. Sci. 49 (2011) 709–718.
[3] 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.
[4] Occupation Safety and Health Administration. https://www.osha.gov/data/commonstats, 2018. (Asessed 3rd October 2020).
[5] Ministry of Housing and Urban-Rural Development of the People's Republic of China.http://www.mohurd.gov.cn/wjfb/202006/t20200624_246031.html, 2019. (Asessed 6th August 2021).
[6] H.W. Heinrich, Industrial Accident Prevention. A Scientific Approach, Industrial Accident Prevention. A Scientific Approach., (1941).
[7] P.E.D. Love, P. Teo, J. Smith, F. Ackermann, Y. Zhou, The nature and severity of workplace injuries in construction: engendering operational benchmarking, Ergonomics 62 (2019) 1273–1288.
[8] B.H.W. Guo, T.W. Yiu, V.A. González, Predicting safety behavior in the construction industry: Development and test of an integrative model, Saf. Sci. 84 (2016) 1–11.
[9] P.E.D. Love, L. Ika, H. Luo, Y. Zhou, B. Zhong, W. Fang, Rework, Failures, and Unsafe Behavior: Moving Toward an Error Management Mindset in Construction, IEEE Trans. Eng. Manage. (2020) 1–13, https://doi.org/10.1109/TEM.2020.2982463.
[10] O. Angah, A.Y. Chen, Tracking multiple construction workers through deep learning and the gradient based method with re-matching based on multi-object tracking accuracy, Autom. Constr. 119 (2020), 103308.
[11] 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.
[12] 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.
[13] W. Fang, B. Zhong, N. Zhao, P.E.D. Love, H. Luo, J. Xue, S. Xu, A deep learning-based approach for mitigating falls from height with computer vision: Convolutional neural network, Adv. Eng. Inf. 39 (2019) 170–177.
[14] W. Fang, L. Ding, P.E.D. Love, H. Luo, H. Li, F. Peña-Mora, B. Zhong, C. Zhou, Computer vision applications in construction safety assurance, Autom. Constr. 110 (2020), 103013.
[15] J. Cai, L. Yang, Y. Zhang, S. Li, H. Cai, Multitask Learning Method for Detecting the Visual Focus of Attention of Construction Workers, J. Construct. Eng. Manage. 147 (2021) 04021063.
[16] H. Kim, K. Kim, H. Kim, Vision-Based Object-Centric Safety Assessment Using Fuzzy Inference: Monitoring Struck-By Accidents with Moving Objects, J. Comput. Civil Eng. 30 (4) (2016) 04015075, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000562.
[17] X. Luo, H. Li, D. Cao, F. Dai, JoonOh Seo, SangHyun Lee, Recognizing Diverse Construction Activities in Site Images via Relevance Networks of Construction-Related Objects Detected by Convolutional Neural Networks, J. Comput. Civil Eng. 32 (3) (2018) 04018012, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000756.
[18] X. Luo, H. Li, H. Wang, Z. Wu, F. Dai, D. Cao, Vision-based detection and visualization of dynamic workspaces, Autom. Constr. 104 (2019) 1–13.
[19] W. Fang, P.E.D. Love, L. Ding, S. Xu, T. Kong, H. Li, Computer Vision and Deep Learning to Manage Safety in Construction: Matching Images of Unsafe Behavior and Semantic Rules, IEEE Transactions on Engineering Management (2021) 1–13, https://doi.org/10.1109/TEM.2021.3093166.
[20] B.E. Mneymneh, M. Abbas, H. Khoury, Evaluation of computer vision techniques for automated hardhat detection in indoor construction safety applications, Front. Eng. Manage. 5 (2018) 227–239.
[21] B.H. Sheppard, J. Hartwick, P.R. Warshaw, The Theory of Reasoned Action: A Meta-Analysis of Past Research with Recommendations for Modifications and Future Research, J. Consumer Res. 15 (1988) 325–343.
[22] I. Ajzen, The theory of planned behavior, Organ. Behav. Hum. Decis. Process. 50 (1991) 179–211.
[23] M.F. Peerally, S. Carr, J. Waring, M. Dixon-Woods, The problem with root cause analysis, BMJ Quality Safety 26 (2017) 417.
[24] B.U. Ayhan, O.B. Tokdemir, Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial Neural Networks, J. Const. Eng. Manage. 146 (2020) 04019114.
[25] F. Davoudi Kakhki, S.A. Freeman, G.A. Mosher, Use of Neural Networks to Identify Safety Prevention Priorities in Agro-Manufacturing Operations within Commercial Grain Elevators, Applied Sciences 9 (2019) 4690.

[26] D.A. Patel, K.N. Jha, Neural Network Model for the Prediction of Safe Work Behavior in Construction Projects, J. Const. Eng. Manage. 141 (2015) 04014066.
[27] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social LSTM: Human Trajectory Prediction in Crowded Spaces, IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) 2016 (2016) 961–971.
[28] W.R. Franklin, Pnpoly-point inclusion in polygon test, Web site: http://www. ecse. rpi. edu/Homepages/wrf/Research/Short_Notes/pnpoly. html, (2006).
[29] J. Choi, B. Gu, S. Chin, J.-S. Lee, Machine learning predictive model based on national data for fatal accidents of construction workers, Autom. Constr. 110 (2020), 102974.
[30] 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.
[31] 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. Plan. Control 29 (5) (2018) 353–366.
[32] C. Spielberger, Encyclopedia of applied psychology, Academic press2004.
[33] 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.
[34] X. An, L. Zhou, Z. Liu, C. Wang, P. Li, Z. Li, Dataset and benchmark for detecting moving objects in construction sites, Autom. Constr. 122 (2021), 103482.
[35] H. Luo, J. Liu, W. Fang, P.E.D. Love, Q. Yu, Z. Lu, Real-time smart video surveillance to manage safety: A case study of a transport mega-project, Adv. Eng. Inf. 45 (2020), 101100.
[36] D. Roberts, M. Golparvar-Fard, End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level, Autom. Constr. 105 (2019), 102811.
[37] X. Yang, H. Li, Y. Yu, X. Luo, T. Huang, X. Yang, Automatic Pixel-Level Crack Detection and Measurement Using Fully Convolutional Network, Comput.-Aided Civ. Infrastruct. Eng. 33 (2018) 1090–1109.
[38] F.-J. Wei, G. Yao, Y. Yang, S. Yujia, Instance-level recognition and quantification for concrete surface bughole based on deep learning, Autom. Constr. 107 (2019), 102920.
[39] G.A. Atkinson, W. Zhang, M.F. Hansen, M.L. Holloway, A.A. Napier, Image segmentation of underfloor scenes using a mask regions convolutional neural network with two-stage transfer learning, Autom. Constr. 113 (2020), 103118.
[40] 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 (6) (2017) 1137–1149.
[41] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: Single Shot MultiBox Detector, Springer International Publishing, Cham, 2016, pp. 21–37.
[42] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You Only Look Once: Unified, Real-Time Object Detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779–788.
[43] B. Kim, S. Cho, Automated Multiple Concrete Damage Detection Using Instance Segmentation Deep Learning Model, Applied Sciences 10 (2020) 8008.
[44] Z. Zhu, X. Ren, Z. Chen, Visual Tracking of Construction Jobsite Workforce and Equipment with Particle Filtering, J. Comput. Civil Eng. 30 (2016) 04016023.
[45] M. Bügler, A. Borrmann, G. Ogunmakin, P.A. Vela, J. Teizer, Fusion of Photogrammetry and Video Analysis for Productivity Assessment of Earthwork Processes, Comput.-Aided Civ. Infrastruct. Eng. 32 (2017) 107–123.
[46] J. Cai, H. Cai, Robust Hybrid Approach of Vision-Based Tracking and Radio-Based Identification and Localization for 3D Tracking of Multiple Construction Workers, J. Comput. Civil Eng. 34 (2020) 04020021.
[47] A. He, C. Luo, X. Tian, W. Zeng, A Twofold Siamese Network for Real-Time Object Tracking, CoRR (2018).
[48] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, P.H.S. Torr, Fast Online Object Tracking and Segmentation: A Unifying Approach, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1328–1338.
[49] A. Khoreva, F. Perazzi, R. Benenson, B. Schiele, A. Sorkine-Hornung, Learning Video Object Segmentation from Static Images, CoRR (2016).
[50] H. Ci, C. Wang, Y. Wang, Video Object Segmentation by Learning Location-Sensitive Embeddings: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI, 2018, pp. 524-539.
[51] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 539-546 vol. 531.
[52] L. Bertinetto, J. Valmadre, J.F. Henriques, A. Vedaldi, P.H.S. Torr, Fully-Convolutional Siamese Networks for Object Tracking, Springer International Publishing, Cham, 2016, pp. 850–865.
[53] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, 2016, pp. 770–778.
[54] P.O. Pinheiro, T.-Y. Lin, R. Collobert, P. Dollár, Learning to Refine Object Segments, Springer International Publishing, Cham, 2016, pp. 75–91.
[55] P.O.O. Pinheiro, R. Collobert, P. Dollar, Learning to Segment Object Candidates, Adv. Neural Info. Process. Syst. 28 (2015).
[56] S. Pellegrini, A. Ess, K. Schindler, L.v. Gool, You'll never walk alone: Modeling social behavior for multi-target tracking, in: 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 261–268.
[57] A. Lerner, Y. Chrysanthou, D. Lischinski, Crowds by Example, Comput. Graphics Forum 26 (2007) 655–664.
[58] S. Citrin, Two dimensional Block Kalman Filtering using multiple estimators, Colorado State University, 1990.
[59] 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.
[60] A. Graves, Generating Sequences With Recurrent Neural Networks, CoRR (2013).
[61] F. Perazzi, J. Pont-Tuset, B. McWilliams, L.V. Gool, M. Gross, A. Sorkine-Hornung, A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, 2016, pp. 724–732.
[62] X. Jia, H. Lu, M. Yang, Visual tracking via adaptive structural local sparse appearance model, in: IEEE Conference on Computer Vision and Pattern Recognition 2012, 2012, pp. 1822–1829.
[63] W. Zhong, H. Lu, M. Yang, Robust Object Tracking via Sparse Collaborative Appearance Model, IEEE Trans. Image Process. 23 (2014) 2356–2368.
[64] Bo Xiao, Zhenhua Zhu, Two-Dimensional Visual Tracking in Construction Scenarios: A Comparative Study, J. Comput. Civil Eng. 32 (3) (2018) 04018006, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000738.
[65] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, A. Alahi, Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks, CoRR (2018).
[66] V.R. Kamat, Trajectory Prediction of Mobile Construction Resources Toward Pro-active Struck-by Hazard Detection, in: Proceedings of the 36th International Symposium on Automation and Robotics in Construction (ISARC), International Association for Automation and Robotics in Construction (IAARC), 2019, pp. 982–988.
[67] 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 (2010) 630–640.