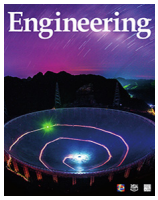
Engineering xxx (xxxx) xxx

Contents lists available at ScienceDirect

Engineering

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



Research
Civil Engineering—Article

Construction Activity Analysis of Workers Based on Human Posture Estimation Information

Xuhong Zhou [a], Shuai Li [a,*], Jiepeng Liu [a], Zhou Wu [b,*], Yohchia Frank Chen [c]

[a] School of Civil Engineering, Chongqing University, Chongqing 400044, China
[b] School of Automation, Chongqing University, Chongqing 400044, China
[c] Department of Civil Engineering, The Pennsylvania State University, Middletown, PA 17057, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Pose estimation
Activity analysis
Object tracking
Construction workers
Automatic systems

ABSTRACT

Identifying workers' construction activities or behaviors can enable managers to better monitor labor efficiency and construction progress. However, current activity analysis methods for construction workers rely solely on manual observations and recordings, which consumes considerable time and has high labor costs. Researchers have focused on monitoring on-site construction activities of workers. However, when multiple workers are working together, current research cannot accurately and automatically identify the construction activity. This research proposes a deep learning framework for the automated analysis of the construction activities of multiple workers. In this framework, multiple deep neural network models are designed and used to complete worker key point extraction, worker tracking, and worker construction activity analysis. The designed framework was tested at an actual construction site, and activity recognition for multiple workers was performed, indicating the feasibility of the framework for the automated monitoring of work efficiency.

1. Introduction

In the construction industry, workers are one of the most important resources, and their actions have a direct impact on the project's schedule and cost. The inefficiency of construction employees ultimately leads to low production, which wastes time and resources. Therefore, workers must be monitored on-site, and tracking their movements is an efficient method of gauging their productivity. Managers in the construction industry can obtain up-to-date information on the status of workers for reference purposes and adjust their strategies accordingly.

The majority of traditional monitoring systems rely on manual monitoring performed by on-site foremen, who oversee the status of workers. Some workers may not perform duties efficiently when not actively monitored, resulting in decreased productivity. This type of monitoring has considerable limits, and many foremen are required to monitor workers, which will raise expenses. The reliability of the results cannot be guaranteed and is somewhat subjective. Therefore, a system that automates the analysis of the actions of construction workers is required to guarantee that the workforce performs efficiently.

Researchers have developed various technologies to extract and analyze field information automatically and have designed corresponding systems to monitor construction sites. Currently, there are two types of methods: nonvisual and visual. Nonvisual methods include the global positioning system (GPS), inertial measurement unit (IMU) system, radio-frequency identification (RFID) system, and ultrawideband (UWB) system [1–4]. These methods are based on electronic sensors that continuously collect construction information (e.g., speed, acceleration, and direction of an object) to classify workers' activities. Construction workers, however, are required to wear sensor equipment for data acquisition, which is inconvenient. In addition, it is difficult to classify detailed worker activities using nonvisual methods. GPS, for example, is used to record changes in location information, but it is not capable of recognizing fine activities.

In contrast, visual methods have no disadvantages compared to nonvisual methods. Without direct touch, these systems can analyze the behaviors of construction workers using pictures and video data. Currently, a considerable amount of visual data is collected for monitoring construction sites [5]. Numerous studies

* Corresponding authors.
E-mail addresses: li_shuai@cqu.edu.cn (S. Li), wuzhsky@gmail.com (Z. Wu).

https://doi.org/10.1016/j.eng.2023.10.004
2095-8099/

Please cite this article as: X. Zhou, S. Li, J. Liu et al., Construction Activity Analysis of Workers Based on Human Posture Estimation Information, Engineering, https://doi.org/10.1016/j.eng.2023.10.004

have used noncontact ordinary red–green–blue (RGB) cameras or red–blue–green–depth (RGB-D) cameras with computer algorithms to identify the activities of workers [6,7]. RGB-D cameras (such as the Microsoft Kinect camera) are sensing systems that capture RGB images along with per-pixel depth information [8]. In addition, the depth information is used to extract the contour map of the humans and to obtain key point data about the workers. In related research, extracted critical point data are used to identify unsafe worker behaviors in real time [9]. Compared to an original image, key point data have a smaller data volume and can be processed more quickly. However, due to the limited shooting distance of an RGB-D depth camera and the influence of light, it is generally only suitable for indoor scenes. A site scene that is outdoors does not include depth information or key points.

Few studies correlate activities to specific workers because the majority of approaches only assess static video frame images and do not establish associations between workers with front and rear frames, thus further limiting the evaluation of the labor consumption of each worker [10]. This is a disadvantage of computer vision-based noncontact approaches, although some nonvision-based systems perform this task more effectively. In addition, the majority of worker action identification systems rely on raw images, which are computationally intensive and difficult to guarantee in real time compared to methods based on key point data. In addition, construction sites typically have several cameras, and the volume of processed video data is substantial. Compared to approaches based on key point data, the majority of existing vision-based systems rely on the original image to complete worker action detection, which is computationally demanding and difficult to perform in real time.

In this paper, a new vision-based activity recognition framework is introduced, where a video recorded by an ordinary camera is used as the input to automatically obtain the behavior of each worker. In this framework, the lightweight pose estimation network is used to obtain human key point information from the video taken by an ordinary camera. Then, a multiperson tracking algorithm is adopted, and the boundary frame of workers is extracted by using the key points to complete the extraction of motion and appearance information, thus effectively tracking workers and key points. Finally, multilayer fully connected (FC) neural networks and stacked long short-term memory (LSTM) are designed to classify the key point information of each worker, complete action recognition, and analyze the construction efficiency. The feasibility of the proposed framework is tested and verified using videos collected from actual construction sites. The contribution of the article is as follows:

(1) A human body pose extraction algorithm is used to extract worker key points from ordinary camera videos, and a neural network model is combined with the algorithm to analyze the activities of workers;

(2) The impact of the combination of time and space on the accuracy of worker activity analysis and actual site monitoring videos are considered;

(3) Appearance and motion information is used to establish the connection between workers in the front and back frames of the video to complete multiplayer action recognition and efficiency statistics.

2. Related work

This section introduces the existing nonvisual and visual construction activity recognition and an analysis of related work.

2.1. Nonvisual system

Research work on nonvision systems is mainly based on the Internet of Things (IoT). An IoT-based system takes a target object, which wears electronic sensors, as the analysis object and relies on electronic sensors to analyze the activity or working states of the object by continuously collecting the information of sensors, including speed, acceleration, and direction. The research on sensor-based contact monitoring focuses on remote positioning and tracking technology. Kelm et al. [2] designed a mobile RFID portal to check if the workers' personal protective equipment meets the corresponding specifications. Pradhananga et al. [11] further integrated the results of GPS to complete the measurement of productivity in earthmoving operations. Montaser et al. [12] developed a method for location identification and material tracking using RFID technology that can be used to obtain information required for near-real-time decision-making. Akhavian et al. [13] used mobile sensors (accelerometers, gyroscopes, and GPS) and machine learning classifiers to identify earthwork equipment activities by analyzing the activity of equipment. Cheng et al. [3] obtained worker movement and site distribution information through UWB and completed the identification of construction worker activities. Zhao et al. [14] and Sanhudo et al. [15] used deep neural network (DNN) models to recognize construction workers' activity from motion data captured by wearable IMU sensors. Some research has focused on the study of human body motions that could cause work-related musculoskeletal disorders (WMSDs) in construction-related activities based on the measurement of motion data from wearable wireless IMUs [16–18].

One of the main advantages of sensors is that they can effectively provide the identity information of the carrier or identification (ID) number and analyze the activities of sensors worn by different workers [19]. However, the application of the IoT system requires every construction worker to wear IoT sensors, and the wearing of sensors may affect normal construction activities and is generally disliked by construction workers [20].

2.2. Visual system

Researchers have used noncontact ordinary or depth cameras with computer algorithms to identify worker activities. In recent years, this research topic has drawn increasing interest. With the emergence of depth cameras, researchers have used them to obtain key points to complete worker activity recognition. For example, a Kinect depth camera was used to extract the human body contour and capture the human figure, and fast skeletonization was used to obtain human posture descriptors. In view of the problem that construction workers often suffer from various musculoskeletal diseases, incorrect activities are detected using depth camera posture acquisition [21]. Khosrowpour et al. [6] used a Microsoft Kinect sensor to estimate the human pose and the personnel key point information obtained to analyze the activity of workers. Research has confirmed that pose-estimated data can be informative enough to understand and classify human actions [22], but the current key point-based activity recognition relies on depth camera extraction with high costs and limited measurement distances [23].

In the field of ordinary camera-based intelligent surveillance, with significant advances in computer vision technology, computer vision and pattern recognition have shown that deep learning methods are superior to traditional machine learning methods, including image classification [24], object detection [25], and action recognition tasks [26]. Fang et al. [27] and Kim et al. [28]

detected construction workers and equipment from RGB data based on the regional detection convolutional neural network (CNN). Fang et al. [29] used a faster CNN based on region nomination (Faster-RCNN) to determine if wearing a hardhat is necessary. Zhang et al. [30] presented a real-time deep learning approach for the detection of cracks on bridge decks. The above research is mainly based on images to accomplish the recognition of unsafe activities or objects. The above methods only process video frames or pictures separately, without establishing a connection between the front and back frame objects, so the processing results between the front and back frames exist independently, resulting in scenes where multiple workers exist and the corresponding detection cannot be achieved. Considering the continuous nature of activities, the researchers have based their studies on continuous video frame data. Yang et al. [5] studied vision-based worker action recognition using the bag-of-feature framework. Ding et al. [31] developed a hybrid deep learning model that integrates a CNN and LSTM [32] that automatically recognizes unsafe worker actions. Luo et al. [10,33] used images and optical flow images to obtain spatial and temporal information and used the temporal segment network (TSN) [34] method to solve the problem of identifying multiple workers' activities at a construction site through double-flow analysis. Chen et al. [35] used a three-dimensional (3D) CNN to precisely recognize excavator activities and classify them into detailed types (e.g., digging, loading, and swinging). However, in cluttered construction scenarios, there are many objects other than workers. The above method establishes a link between the front and back frames and combines the data from the front and back frames for action recognition. However, when there are many workers in the front and back frames, the above algorithm has difficulty distinguishing each worker. The above methods are based on double stream or 3D convolution, which is computationally intensive, and it is difficult to guarantee real-time performance. The drawback of deep learning approaches is that the learned representations may not be specifically focused on human actions because the entire areas of the video frames are provided with the learning representations. Moreover, few methods can automatically track workers or identify and track new workers without requiring manual intervention [10].

Compared with ordinary cameras, a depth camera can obtain the key point information of workers with faster information processing speeds, and the key point information eliminates excess data [36]. At present, there is a vision-based method to obtain the key point information of workers [37]. Roberts et al. [38] used the key point extraction algorithm, alphapose, for pose estimation, and 3D convolution was used to process dual stream and key point information. Since alphapose is a top-down estimation method, it becomes computationally intensive when the number of workers increases [39], and the 3D convolutional processing of dual-flow data does not guarantee real-time performance. In addition, the oriented fast and rotated brief (ORB) features in the extracted bounding boxes may not be the workers' own features, and the algorithm switches between the background and workers to find feature points, which makes it difficult to ensure the association of front and back frame features to achieve tracking. The key point extraction method has been shown to be feasible for activity recognition tasks [22,40,41]. In contrast, skeleton features provide quantified information about human joints and bones. Compared to RGB flows, skeleton features can provide more compact and useful information in dynamic situations with complicated backgrounds.

3. Proposed framework

The framework of multiple worker construction actions and working efficiency recognition is shown in Fig. 1, which includes four main modules: key point extraction of construction workers, location extraction and tracking of workers, action recognition, and working efficiency analysis. First, the pose estimation detector processes all workers in the video to provide the worker key point data. Second, the key point data are used to obtain the position information of each worker's current frame, and the tracking algorithm based on the CNN is used to identify each individual worker's location and movement trajectory. The tracking result returns the ID number of each worker. Through the worker's ID number, the key point information of the detection is mapped to different workers in time; that is, the key point information corresponding to each worker is obtained continuously. Then, a neural network model with temporal and spatial information processing capabilities is designed to recognize the actions of workers. Finally, the results of action recognition are summarized to calculate the work efficiency of each worker for a period and the construction efficiency of all workers.

3.1. Extraction of key points

The amount of key point data is a relatively small; hence, the influence of complex backgrounds can be eliminated. The purpose of this module is to extract key point information of the human body from a video of an ordinary camera rather than relying on a depth camera for key point extraction. To quickly extract key point information of multiple workers continuously at a construction site from video data, a lightweight posture extraction method is proposed. The model of the improved network structure, OpenPose [40], was used in the framework to extract key points and detect multiple workers. The OpenPose algorithm is used to extract the data of workers' key points and analyze the abnormal use of helmets by combining geometric relations [42].

The OpenPose model is an efficient method for multiperson pose estimation with competitive performance on multiple public benchmarks. The model takes an image with a size of $w \times h$ (width $\times$ height) as the input. The model outputs two-dimensional (2D) key points for each person after processing. In the original OpenPose network, an image was processed by a CNN whose first ten layers were composed of Visual Geometry Group 19-layer network (VGG-19) [43], generating a set of feature maps ($\boldsymbol{F}$). To further improve the estimation speed, a lightweight architecture was employed in this study, using MobileNetV2 [44] as the feature extraction network instead of VGG-19. The feature extraction network structure is shown in Fig. 2 and is composed mainly of the bottleneck residual block (BRB) structure. This structure takes as an input a low-dimensional compressed representation that is first expanded to a high dimension and filtered with a lightweight depthwise convolution [44]. Subsequently, features are projected back to a low-dimensional representation with a linear convolution. The remaining structures include standard convolution (SC), depthwise separable convolution (DSC) [45], and an upsampling layer (UPS-L). A set of feature maps $\boldsymbol{F}$ are generated by analyzing the raw image using the network.

Then, the network consists of a multistage CNN (Fig. 2) and is divided into multiple similar stages, with each stage consisting of two branches: one for confidence maps to obtain body part locations, shown in the orange dotted box, and another for predicting the part affinity fields (PAFs) to encode the degree of body part-to-part association, shown in the blue dotted box. Finally, the confidence maps and PAFs are parsed by greedy algorithm inference to output the 2D key points for all workers.

The first stage takes the feature map $\boldsymbol{F}$ as the input and generates a set of PAFs $\boldsymbol{L}^1$.

$$\boldsymbol{L}^1 = \varphi^1(\boldsymbol{F}) \tag{1}$$

where φ^1 is the multistage CNN in the orange dotted box for inference at stage 1.

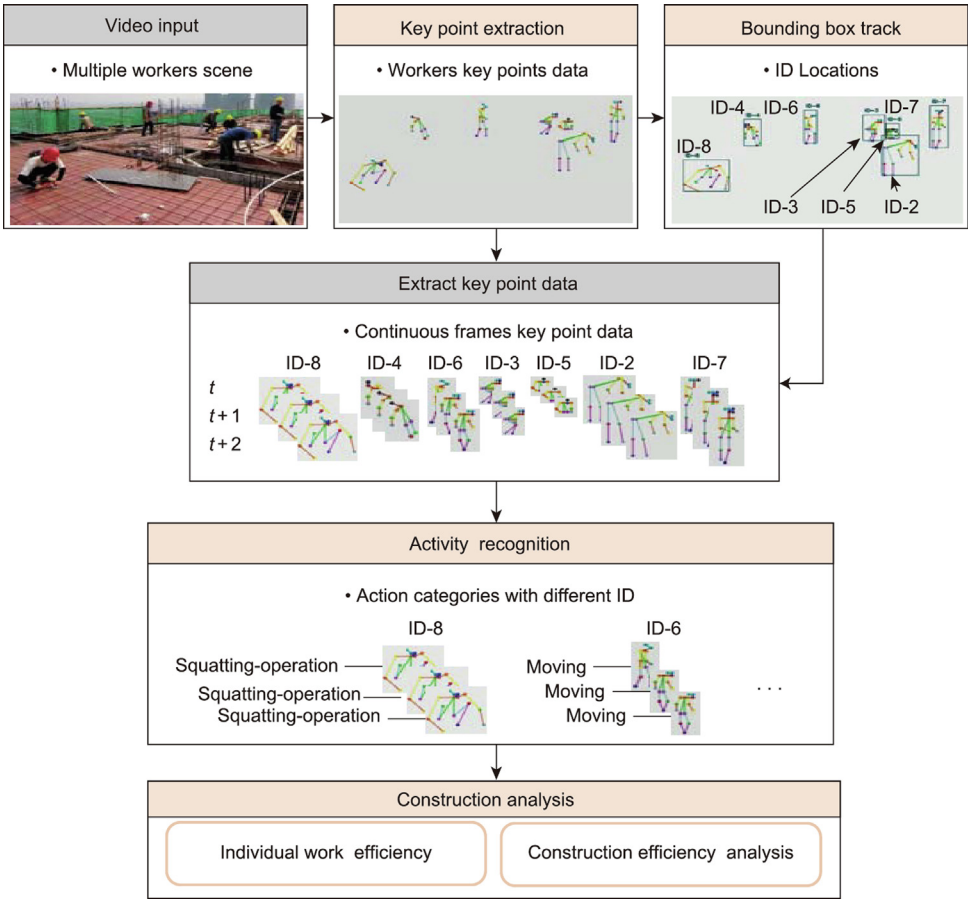


Fig. 1. The framework of multiple worker construction actions and working efficiency recognition.

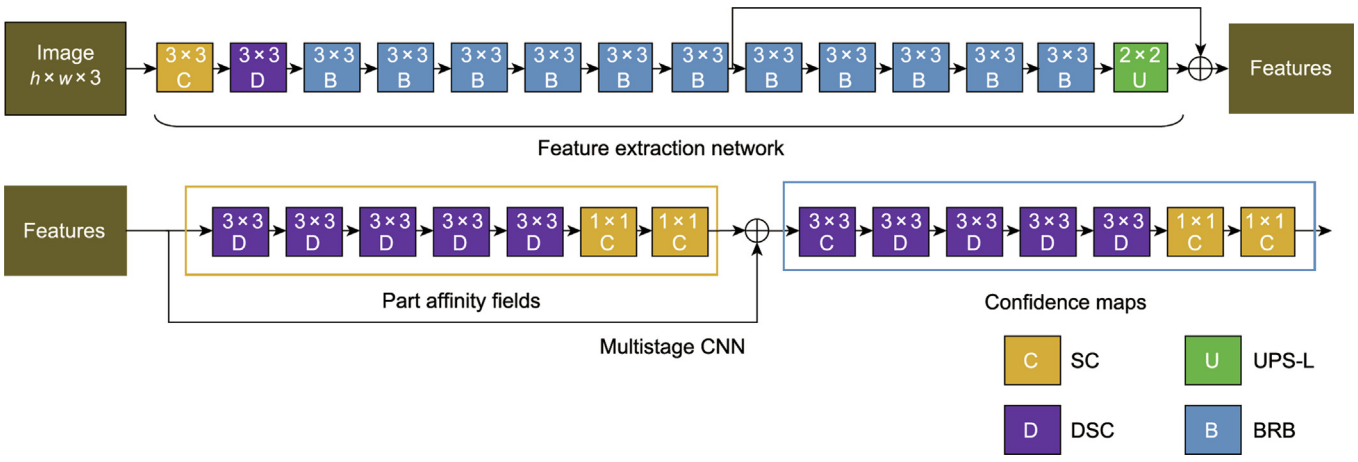


Fig. 2. Pose estimation network structure. SC: standard convolution; DSC: depthwise separable convolution; UPS-L: upsampling layer; BRB: bottleneck residual block.

At each subsequent stage, the feature map $\boldsymbol{F}$ and the predictions of PAFs $\boldsymbol{L}^{t-1}$ from the previous stage are concatenated as input for the multistage CNN, where t denotes the stage number and $\boldsymbol{L}^t$ denotes the prediction vector field of PAFs at stage t, and each element of $\boldsymbol{L}^t$ represents the direction and magnitude of a pixel belonging to a limb.

$$\boldsymbol{L}^t = \varphi^t\left(\boldsymbol{F}, \boldsymbol{L}^{t-1}\right), \forall 2 \le t \le T_{\mathrm{P}} \tag{2}$$

where φ^t is the multistage CNN in the orange dotted box for inference at stage t, and T_{P} is the number of total PAF stages.

After T_{P} iterations, starting with the most recently updated PAF prediction, a process similar to that in the PAF stages is repeated for the confidence map detection:

$$\boldsymbol{S}^{T_{\mathrm{P}}} = \rho^t\left(\boldsymbol{F}, \boldsymbol{L}^{T_{\mathrm{P}}}\right), \forall t = T_{\mathrm{P}} \tag{3}$$

$$\boldsymbol{S}^t = \rho^t\left(\boldsymbol{F}, \boldsymbol{L}^{T_{\mathrm{P}}}, \boldsymbol{S}^{t-1}\right), \forall T_{\mathrm{P}} < t \le T_{\mathrm{P}} + T_{\mathrm{C}} \tag{4}$$

where ρ^t is the multistage CNN in the blue dotted box for inference at stage t; T_{C} is the number of total confidence map stages; $\boldsymbol{S}$ is the

confidence maps of the corresponding stage; $\boldsymbol{S}^t$ is the output of the confidence map of the t-th stage, and each element of $\boldsymbol{S}^t$ represents the probability of a pixel belonging to a key point; $\boldsymbol{S}^{t-1}$ is the output of the confidence map of the $(t-1)$-th stage, which has the same shape and meaning as $\boldsymbol{S}^t$; $\boldsymbol{L}^{T_P}$ denotes the prediction vector field of PAFs at T_P stage; and $\boldsymbol{S}^{T_P}$ is the input of the confidence map of the T_P-th stage.

A loss function at the end of each stage is applied to keep the network iteratively detecting PAFs in the first branch and the confidence maps in the second branch. The loss function of the PAF branch at stage t_i and loss function of the confidence map branch at stage t_k are expressed by Eqs. (5) and (6), respectively:

$$f_{\boldsymbol{L}}^t = \sum_{c=1}^{C} \sum_{p} \boldsymbol{W}(p) \| \boldsymbol{L}_c^t(p) - \boldsymbol{L}_c^*(p) \|_2^2 \tag{5}$$

$$f_{\boldsymbol{S}}^t = \sum_{j=1}^{J} \sum_{p} \boldsymbol{W}(p) \| \boldsymbol{S}_j^t(p) - \boldsymbol{S}_j^*(p) \|_2^2 \tag{6}$$

where C and c are both indices that represent the connections between human body parts, C is the total number of connections, and c is a specific connection; J and j are both indices that represent human body parts, J is the total number of parts, and j is a specific part; $\boldsymbol{L}_c^*$ is the ground truth part affinity vector field of the c-th degree of body part-to-part association; $\boldsymbol{S}_j^*$ is the ground truth part confidence map of the j-th body part location; and $\boldsymbol{W}$ is a binary mask. For matrix $\boldsymbol{W}$, when an image pixel p lacks annotation, the element corresponding to p in $\boldsymbol{W}$ is 0. The overall loss function is:

$$f = \sum_{t=1}^{T_P} f_{\boldsymbol{L}}^t + \sum_{t=T_P+1}^{T_P+T_C} f_{\boldsymbol{S}}^t \tag{7}$$

In this study, the convolutions in multistage networks are replaced by separable convolutions. By using the DSC structure as a replacement, the pose estimation network structure is improved, as shown at the bottom of Fig. 2. For multistage networks, separable convolutions are used for replacement, which reduces the number of network calculations. If the size of the convolution kernel is dynamic kernel (DK) × DK, IN is the number of input channels, and ON is the number of output channels. The parameter quantity of SC can be formulated as SC = DK × DK × IN × ON. The depth separable volume integral performs the same processing operation and is formulated as DSC = DK × DK × IN + IN × ON.

The extraction of key points is shown in Fig. 3, with Fig. 3(a) showing the input original image and Fig. 3(b) showing the extracted key point information of workers. Each worker is composed of 18 key points, as shown in Fig. 3(c). The corresponding meaning or body part of each key point is indicated in Table 1.

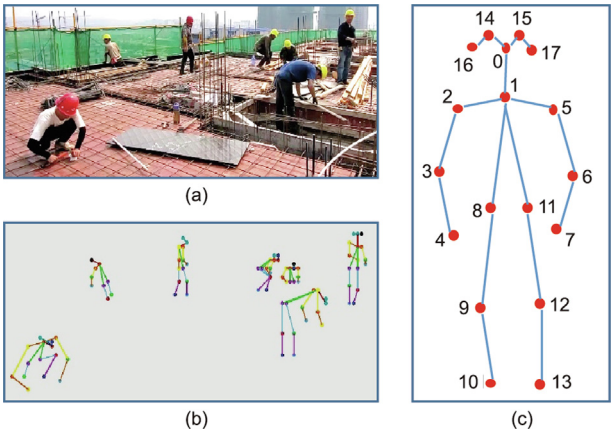


Fig. 3. Key point extraction. (a) Original image; (b) key point information; (c) key point composition.

Table 1
Key point location.

Number	Key point	Number	Key point
0	Nose	9	Left knee
1	Neck	10	Left foot
2	Left shoulder	11	Right hip
3	Left elbow	12	Right knee
4	Left wrist	13	Right foot
5	Right shoulder	14	Left eye
6	Right elbow	15	Right eye
7	Right wrist	16	Left ear
8	Left hip	17	Right ear

3.2. Extraction and tracking of worker locations

The purpose of this module is to process the key point information to obtain the worker's bounding box position information and then use the tracking algorithm to track and analyze the bounding box and the bounding box area image to obtain worker ID numbers. For only one worker, the method described in Section 3.1 can be used to extract the key point information from the video. Let $P = \{P^0, P^1, \ldots, P^K\}$ represent the key points of continuous frame number K, where $P^k = \{(x^0, y^0), (x^1, y^1), \ldots, (x^{17}, y^{17})\}$ represents the key points of the worker in frame k. When there are multiple workers in the video, it is necessary to obtain the key point information of all workers in each frame. The key point detection method can extract the key point information of all workers in each frame at the same time, but for a single worker, the data in the continuous frame cannot be correlated. Let N be the number of detected individuals in the obtained frame, and N is different in different frames. However, the ID numbers n corresponding to each worker in each frame are unchanged, where the key points in the k-th frame of worker ID = n are expressed as $P_n^k = \{(x_n^0, y_n^0), (x_n^1, y_n^1), \ldots, (x_n^{17}, y_n^{17})\}$. Fig. 4 is a schematic plot of the tracking results showing the process of continuously obtaining key points of workers.

Considering the presence of new workers, missed detection of algorithms, disappearance in the field of vision, and continuous movement and occlusion of workers, it is difficult to keep the value of n per worker unchanged and integrate continuous worker frame key point data. To solve the above problems, this study uses a multitarget tracking algorithm to realize the continuous frame association between workers and key point data. In the tracking module, a deep simple online and real-time (SORT) tracker is applied to associate the same workers detected in the previous step across all the frames in the video [46]. The method is used to track the prefabricated wall and collect information on the location and time of the prefabricated wall from the surveillance video [47].

Let x_{max} and y_{max} represent the maximum values of the 18 key point coordinates, x_{min} and y_{min} represent the minimum values, and (w, h, x, y) represent the coordinates of the bounding box, where (x, y) are the coordinates of the center point of the bounding box, and w, h are the width and height of the bounding box, respectively. The coordinates of the bounding box can be found by a simple calculation using the maximum and minimum values. Fig. 5 shows the result of using the key point information to complete the worker's bounding box.

A worker's bounding box $b_n^k = \left(x_n^k, y_n^k, w_n^k, h_n^k\right)$ at frame k is extracted, where (x_n^k, y_n^k) is the location of the center of the ID = n worker's box in the k-th frame, and w_n^k and h_n^k are its width and

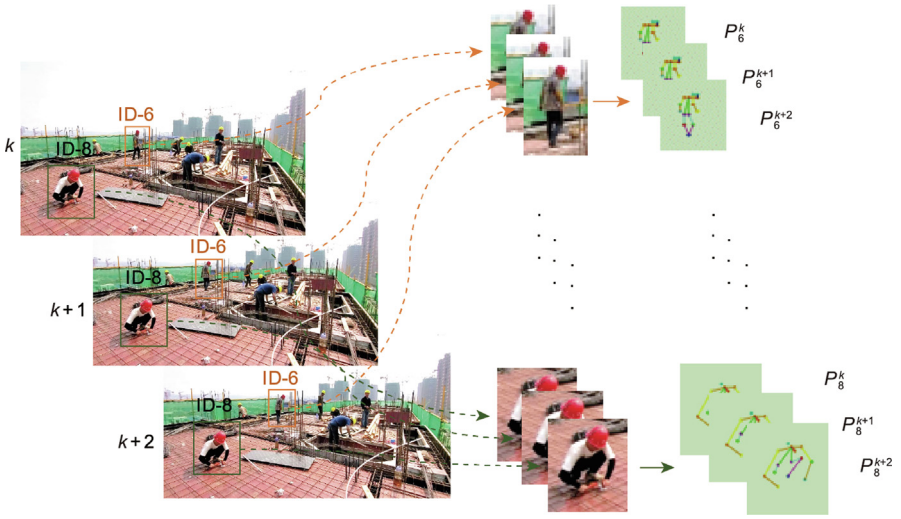


Fig. 4. Schematic plot of the tracking results.

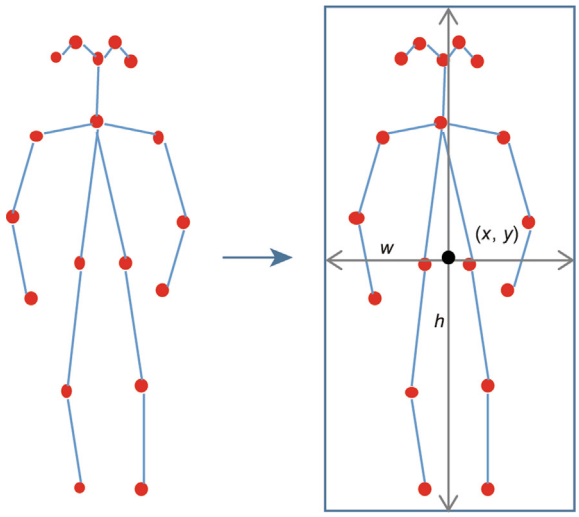


Fig. 5. Extraction of a bounding box.

height in pixels, respectively. In this study, the deep SORT algorithm is chosen as the object tracking method, which is a real-time tracker with competitive performance. This method defines the tracking scenario on an eight-dimensional vector $\boldsymbol{d}^k = [x, y, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h}]^{\mathrm{T}}$ that contains the center position (x, y) and height h of the bounding box, aspect ratio $\gamma = w / h$, and the respective velocities in image coordinates.

The updated trajectory is predicted using a standard Kalman filter with constant velocity motion and a linear observation model [44]. Then, the currently detected value is matched according to the predicted value. Based on the $\boldsymbol{d}^k$ prediction of the state at frame $(k + 1)$, $\boldsymbol{d}^{k+1}$ can be obtained.

The deep SORT algorithm uses two metrics of motion and appearance information for effective tracking. Appearance information (128 dimensional features) and motion information (the tracked position predicted by the Kalman filter) are used to determine the possibility of tracking and detection of the same person. For different scenes, the weights of the two distance-association degrees can be adjusted. The motion-information association speed is relatively fast and is suitable for situations where there is no occlusion in the short term. The appearance descriptors are closely associated with extracting the characteristics of workers and storing them for comparison, which can be applied to a camera in scenes with occlusion by workers. Through the tracking algorithm, the ID numbers n and the key point information of the detection are mapped to different workers in time.

3.3. Action recognition

The above two modules convert the video information taken by an ordinary camera into a sequence of key points corresponding to each worker in consecutive frames, representing the construction action information of different workers. This study selected a multilayer FC neural network to process spatial features such as relative distance and angles between the various key points in a frame. A stacked LSTM network can process temporal feature vectors for the extraction of key point information of continuous frames. The scores of the last nine construction actions are provided by the Softmax function to complete the action classification task.

The shape of the data input by the above network is $10 \times 18 \times 2$, which denotes 18 key points having x and y coordinates each, and 10 indicates ten consecutive frames of data (Fig. 6). The design of the spatial feature extraction network mainly relies on the four FC layers with the specific network parameters listed in Table 2. The dropout layers of the network within each layer are set to 0.1 to prevent overfitting. s^0, s^1, s^2,..., s^{31} form the feature $\boldsymbol{s}$, and the feature $\boldsymbol{s}$ output by the spatial feature extraction network is used as the input of the next stage of the network.

Through the spatial network feature of the extraction network, the spatial features can be obtained, and the corresponding results are input into the temporal feature extraction network. The temporal feature extraction network adopts LSTM to model the time series and is not limited to fixed-length input or output. Therefore, it can not only analyze single-frame feature information but also combine consecutive frames for feature analysis, which is more conducive to solving the problem of action classification. This study uses a two-layer stacked LSTM network with the specific network parameters listed in Table 2. The dimensions are denoted as (batch size, time step, input dimension) inside the parentheses. When the input dimension of LSTM1 is (32, 10, 32), the size of each input to the LSTM model in a batch size is time step $\times$ input dimension, and the number of executions is batch size $-$ 3 + 1 = 30. The output sequence information of the upper LSTM layer is used as the input information of the lower LSTM layer (Fig. 7). After the

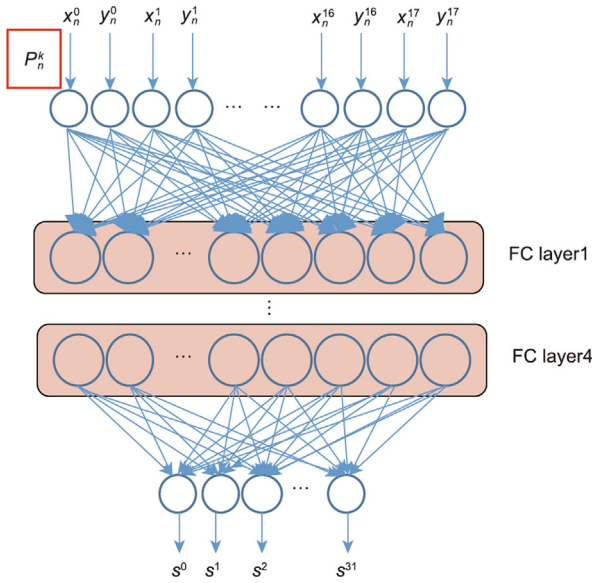


Fig. 6. Spatial feature extraction network. s^0, s^1, s^2,..., s^{31} form the feature ***s***, and the feature ***s*** output by the spatial feature extraction network is used as the input of the next stage of the network.

features extracted by the spatial network are processed by the LSTM layer, they are output to a FC layer with a Softmax activation function and nine neurons. Each of these nine outputs provides the probability of the corresponding action in the form of cross-entropy, and the largest probability among them is selected as the category of the current action. The action classification network has a total of seven layers, and a total of 36 169 parameters are trained. Compared with other neural network models, the number of calculations is small.

4. Experiment and results

In this section, the proposed framework is employed and tested in actual situations. The lightweight posture detection network is used to process videos and complete the production of the workers' key point action dataset. The action recognition network is trained using the above dataset to obtain an action classification model.

4.1. Dataset production

The surveillance video was captured by a mobile camera at multiple construction sites with 30 frames per second and a resolution of 1920 × 1280. The camera was not a fixed installation, and it could be moved to record at different angles and observe multiple-worker construction actions. The video contained multiple workers, with the scene shown in Fig. 8.

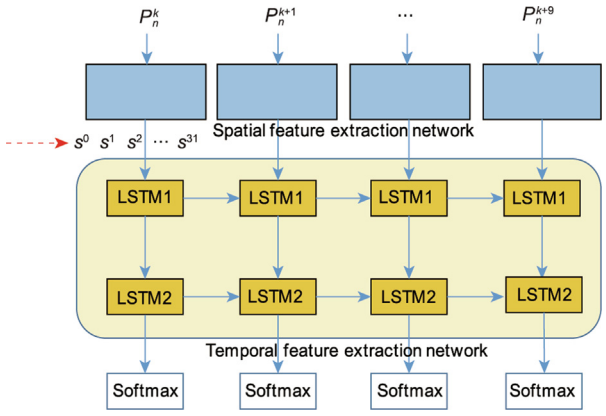


Fig. 7. Temporal feature extraction network.

Fig. 8. Video frames of the construction scene.

4.1.1. Production of the video dataset containing worker construction actions

In the key point dataset, to ensure that the action classification of each worker in each frame corresponding to the key points is correct, it is necessary to create an action video dataset so that each short video has only one worker and one action category.

To meet the requirements of the above video dataset, the collected video needs to be cut in space and time. In space, the image corresponding to each worker border box is cut frame by frame. In time, the same action time frame is cut for each worker in consecutive video frames. Short videos with nine actions are cut, and each video contains an action category for a worker. In this experiment, there are nine action categories, and this number can be adjusted according to different actual needs. The different categories are divided into the construction states (worker states = 1) and non-construction states (worker states = 0). The specific categories are listed in Table 3. The "other" category represents categories not related to construction activities.

Table 2
Parameters of the action recognition network.

Name	Input size	Output size	Number of neurons	Number of parameters	Activation function
FC layer 1	(32, None, 36)	(32, None, 128)	128	4736	ReLU
FC layer 2	(32, None, 128)	(32, None, 64)	64	8256	ReLU
FC layer 3	(32, None, 64)	(32, None, 64)	64	4160	ReLU
FC layer 4	(32, None, 64)	(32, None, 32)	32	2080	ReLU
LSTM1	(32, 10, 32)	(32, 10, 32)	32	8320	Tanh
LSTM2	(32, 10, 32)	(32, 10, 32)	32	8320	Tanh
FC layer 5	(32, None, 32)	(32, None, 9)	9	297	Softmax

The dimensions are denoted as (batch size, time step, input dimension) inside the parentheses in the input and output size columns. ReLU: rectified linear unit; Tanh: hyperbolic tangent function.

Table 3
Worker action category and number of frames.

Action name	Category number	Worker state	Number of frames
Shouldering materials	0	1	2160
Carrying materials	1	1	2160
Sitting-rest	2	0	2430
Standing-rest	3	0	2520
Squatting-operation	4	1	2430
Squatting-rest	5	0	2160
Standing-operation	6	1	2520
Moving	7	0	2520
Other	8	0	2400

4.1.2. Production of the key point dataset

After creating the short video dataset, the improved pose extraction algorithm is used to convert all short videos into key point data. The key point information contains the specific position information of workers in each picture. The unit of sampling is the recorded video frame rate, and the key points in the video are extracted frame by frame. Through data processing, a continuous sequence of key points is obtained for each short video clip. The action category of the short video clip is the category of the generated key point sequence. The key point sequences of the same category are placed together to form a long sequence. During network training, the overall dataset sequence is divided using the batch size. The duration of each video clip is different, and the duration of an action in the video dataset can be both long and short. Therefore, the length of each key point sequence is not fixed, and the length is approximately 90. Fig. 9 shows the key point dataset, and the data are normalized. Normalized by Eq. (8), the dataset contains 36 coordinate points and the action category number. The locations of the body joints in the original key point dataset are the pixel locations in the video, so they are normalized with respect to the size of the video. The raw data in this paper are normalized to the length and width of the video, so $x_{\min}$ and $y_{\min}$ are both 0, $x_{\max}$ is the length of video frame w, and $y_{\max}$ is the width of video frame h. Since the video size is 1920×1080, $x_{\max}$ is 1920 and $y_{\max}$ is 1080.

$$\begin{cases} \hat{x}_i = \dfrac{x_i - x_{\min}}{x_{\max} - x_{\min}} \\ \hat{y}_i = \dfrac{y_i - y_{\min}}{y_{\max} - y_{\min}} \end{cases} \tag{8}$$

where i is the index of the key point in the video, ranging from 0 to 17.

4.2. Training of the action recognition model

The key point dataset is used to train the action classification network. The time step of the recurrent network in this paper is 10, each sample sequence input to the network consists of 10 consecutive frames of key points, and the key point data of each frame are 36 dimensions. The total number of key points for the experiment is 63 900, and the total number of samples is 63 900, where the ratio of the training set, test set, and validation set are 8:1:1. The sample collection for this experiment is continuous; that is, if the key points between the key points at moment k and moment $(k + 9)$ are selected as one sample, the next sample consists of the key point data between $(k + 1)$ and $(k + 10)$.

The action classification network training environment is Python 3.6, the deep learning framework is Keras 2.2.0, the processor is an Intel i7 7700K, the graphics card is an Nvidia GTX 1050ti GPU, the system is Windows 10 64 bit. The training parameters are set as follows: batch size = 32, number of epochs = 200, and initial learning rate = 0.0001. The Adam optimizer is used to control the

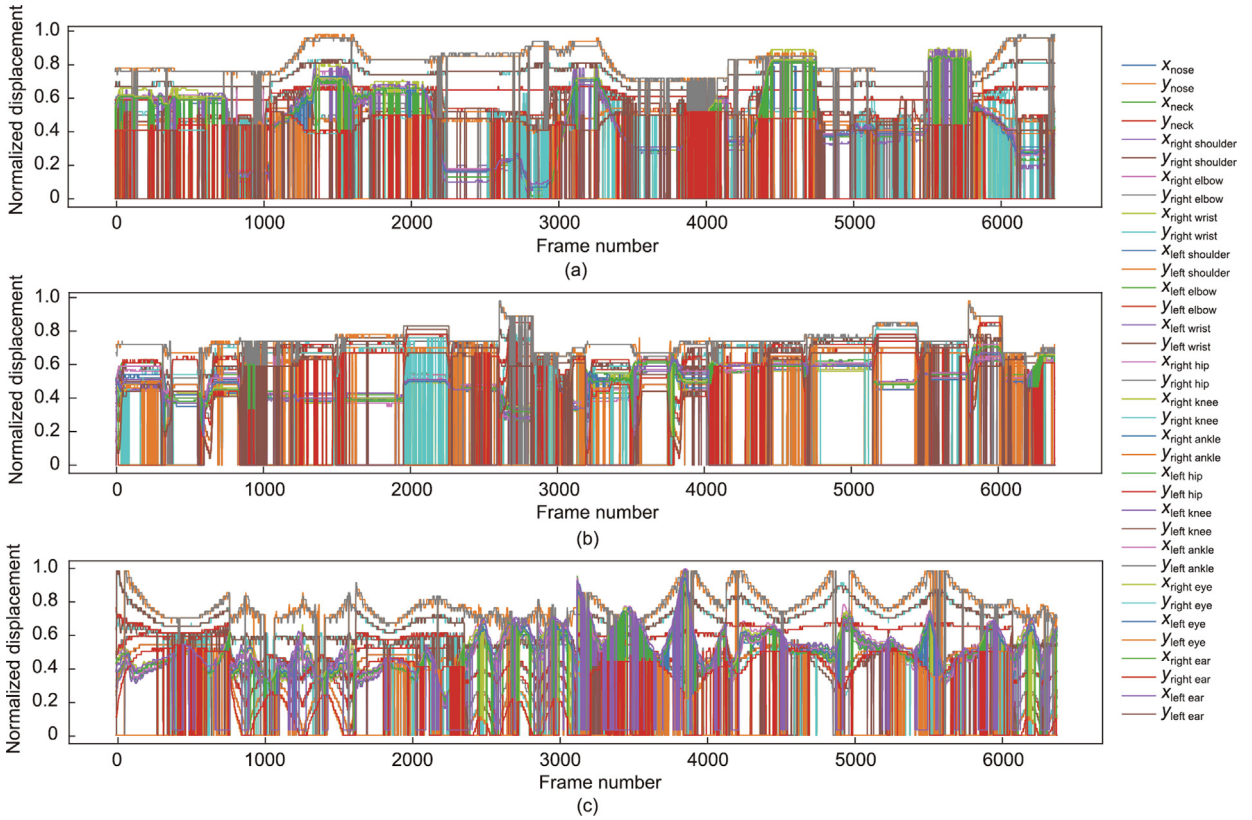


Fig. 9. Key point dataset. (a) Standing-operation key points data; (b) squatting-rest key points data; (c) shouldering materials key points data.

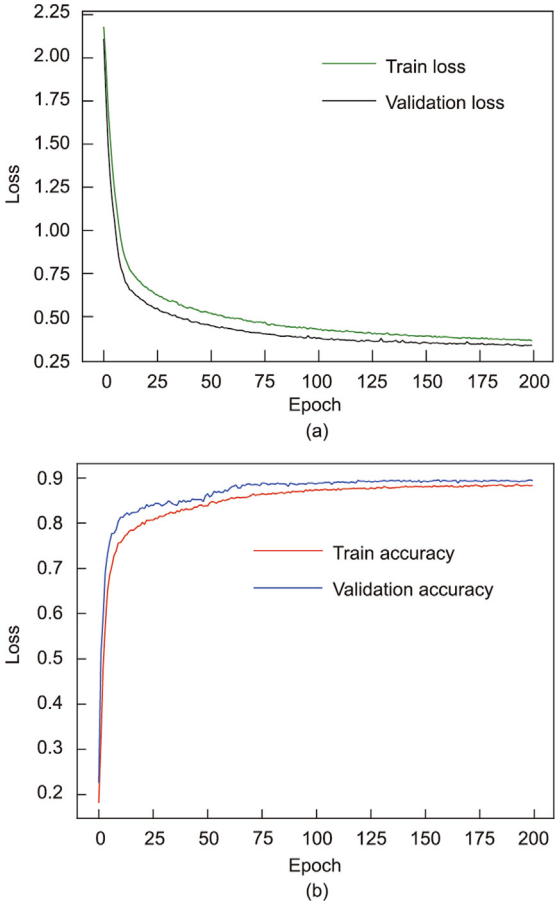


Fig. 10. Training and validation results of the action recognition model. (a) Loss curve; (b) accuracy curve.

learning rate. When the training reaches 200 epochs (Fig. 10), the training loss function becomes stable, and the time consumed is 690 s. Due to the relatively small number of network parameters, the training time is very short compared with that of the original image data. The loss curve during training is shown in Fig. 10(a), and the training loss is stable.

4.3. Test results of the action classification model

In this study, the accuracy and confusion matrix are used as the performance indicators. The accuracy gives the probability of a correct prediction and has been a widely used measurement method in activity recognition. The confusion matrix can not only completely show the numbers of correct and incorrect predictions but also completely show the specific prediction categories of the results of the corresponding activity prediction errors. Additionally, the accuracy of the prediction results for each category can be calculated.

The accuracy can also measure the probability of a classification network making correct predictions. It is a widely accepted evaluation index used in the field of action recognition. Fig. 10(b) shows the accuracy of training, which is approximately 89%.

The confusion matrix analyzes the accuracy of activity recognition on the test dataset, and it can be used to analyze the predictions of each category in detail. The confusion matrix obtained on the test dataset is shown in Fig. 11, and it displays the correct prediction number for each category and the specific recognition of other categories. The overall accuracy of the test dataset is approximately 90%. The accuracy of squatting-operation category

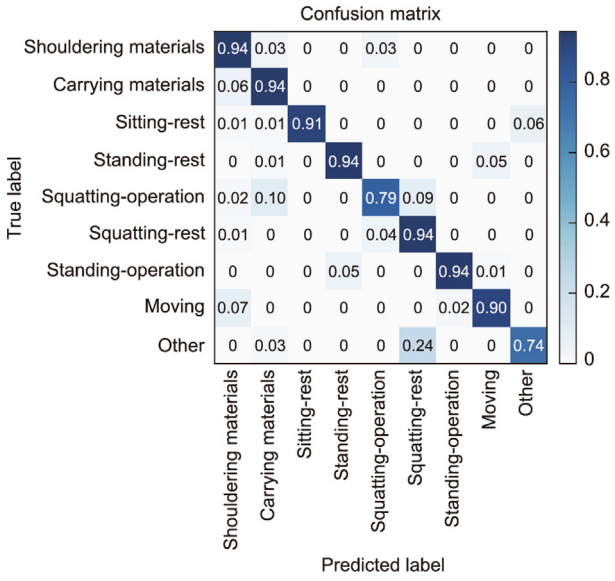


Fig. 11. Matrix diagram of action recognition confusion.

identification and other category identification in Fig. 11 is low. Squatting-operation is mainly misidentified as carrying materials and squatting-rest categories. The possible reasons for this are the similarity of the data itself and the presence of lower body occlusion when carrying materials leading to the loss of key points of the legs. The actions in the other category are partially recognized as squatting-operation due to the large differences in their own data, and the recognition accuracy of this category is the lowest. The accuracy of the remaining categories is higher than 90%, which is high recognition accuracy.

4.4. Activity identification and efficiency analysis results of the actual scene

In this section, the effectiveness of the proposed framework is demonstrated in real scenes, including activity recognition and efficiency analysis. First, tracking the location of a worker is essential to correlate the key point data with the individual worker's number. As mentioned in Section 3.2, the deep SORT algorithm is employed for tracking purposes, enabling the association of key points with the respective worker's number. Consequently, continuous key point data from each worker over a period is acquired for subsequent action recognition. As shown in Fig. 12, the movement trajectories of several workers in two scenes were recorded, with different colors used to mark the movement paths of each worker. The experimental results confirm that the aforementioned tracking method can pinpoint each worker's location while persistently and precisely tracing their movement trajectories throughout the scene.

Workers may perform many kinds of actions during the construction process. In this paper, a construction action is classified as state = 1 (Table 3). For a worker, the actual construction time (the sum of the number of frames with state = 1) can be used to measure worker performance. The actual construction time is compared to the total time (total number of frames) to determine the construction efficiency of workers over a period. Similarly, the construction progress statistics of all workers within the monitoring range of a certain construction stage can be calculated by the number of construction workers (the number of people in state = 1 in each frame) and the construction efficiency of all workers in each frame. The above information can be recorded over time to analyze

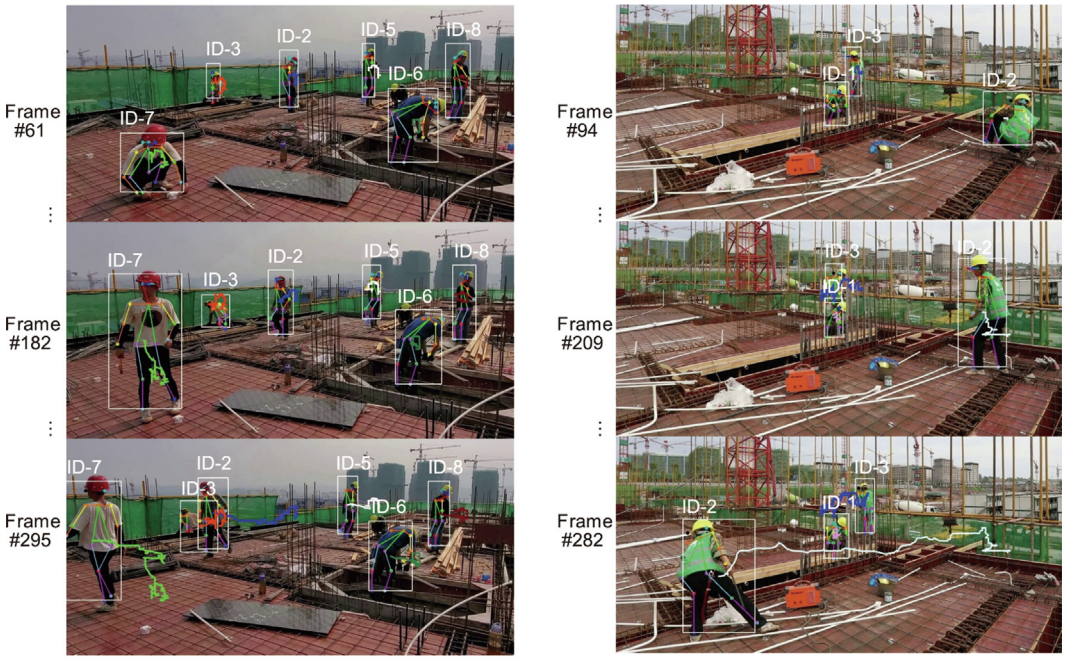


Fig. 12. Worker tracking results. (a) Six workers were tracked in Scene 1; (b) three workers were tracked in Scene 2.

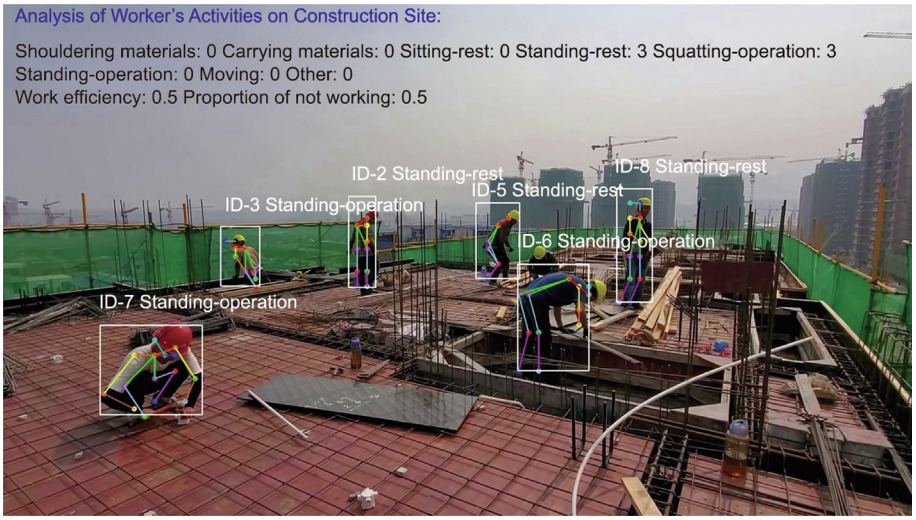


Fig. 13. Activity recognition results. The overall construction efficiency is marked in the frame.

the construction efficiency of each worker and the construction efficiency of all workers over the same period.

A section of the construction worker monitoring video is used to test the proposed framework to solve the problems of multiworker construction activity recognition and working efficiency analysis statistics, as described in this section. For construction activity recognition verification, the action categories of all workers in the video are identified. Seven workers are in the construction video. Fig. 13 shows an example of each worker's activity recognition results. A specific ID number is assigned to each worker automatically by the proposed framework, and the activity category of the worker's current frame is displayed next to each number. Moreover, the system completes the statistics of the number of personnel activity categories, which are used to determine the construction efficiency statistics.

Further statistics are obtained for the three-worker scene in Fig. 14(a) to obtain detailed information on the workers' construction efficiency. The length of this video is 20 s with 30 frames per second, and the video is fed to the network for frame-by-frame analysis. The vertical axis in Fig. 14(b) indicates the category of the action and is between 0 and 8, and the horizontal axis indicates the number of frames. The workers are assigned one of three numbers, and each number records the category of the worker's action. The broken line in the figure indicates that the key point is not detected because the worker is obscured. Using the obtained action information of each worker, the statistics of worker efficiency are obtained. The statistical results of worker efficiency are shown in Fig. 15, where green indicates that the key point of the worker was lost, orange indicates the worker's efficiency during this time, and blue indicates the time when the worker did not work. Work-

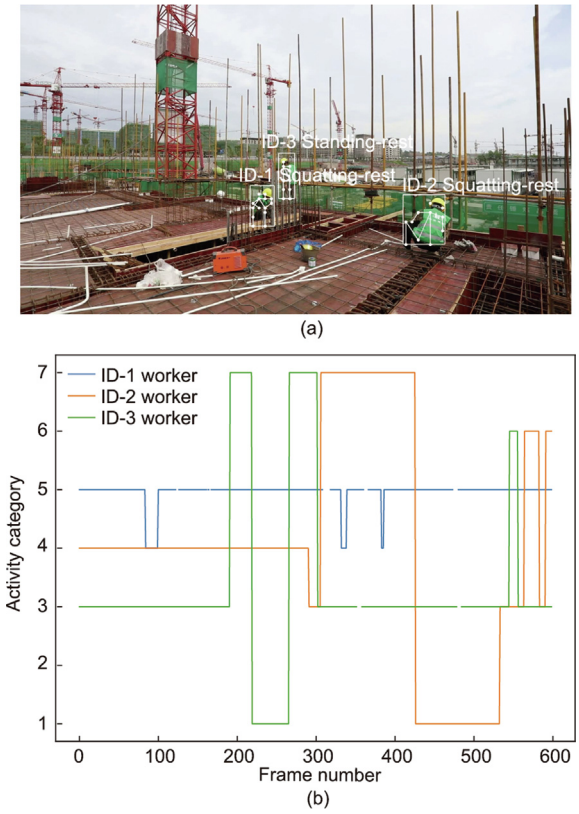


Fig. 14. Worker action recognition results. (a) Testing in a three-worker scene; (b) the distribution of action categories for each worker ID in different frames.

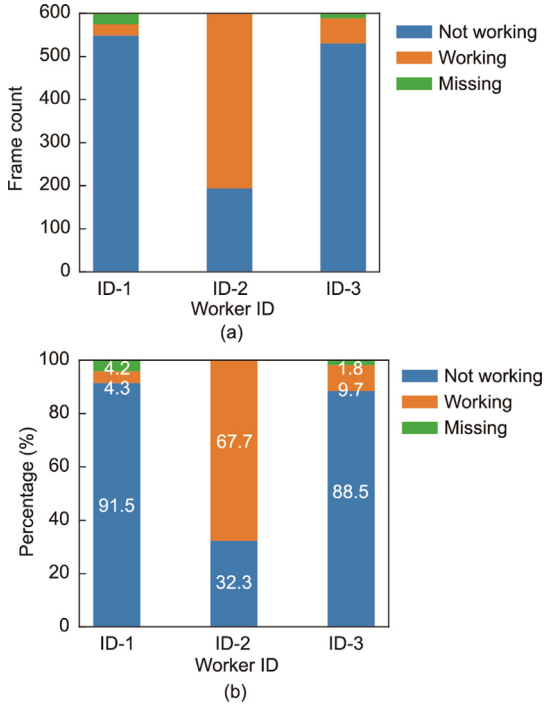


Fig. 15. The statistical results of worker efficiency. (a) The number of frames of working, non-working and missing for different worker IDs; (b) the efficiency of working for different worker IDs.

ers ID-1 and ID-3 spent most of the time resting. From Table 3, we can classify Action 3 and Action 5 as squatting-rest and standing-rest, respectively. Worker ID-2 spent more time carrying materials and squatting to perform operations, and his actual construction efficiency is higher.

4.5. Discussion

Compared with the use of target detection for different worker activity recognition and safety behavior detection, this study not only considers spatial information as the key point coordinate information but also analyzes the key point information of a continuous frame and is thus more in line with the characteristics of continuous changes in activity over time. The method used for activity recognition in this study can complete end-to-end training, facilitate the later adaptation of the number of datasets or the increase in activity categories. The key points of the human body are identified based on a depth camera, which is susceptible to light and is expensive. In this study, only an ordinary camera is used to obtain the key point information from a video stream; the depth sorting tracking algorithm is used to track each worker; and the initial frame does not need to be manually marked with each worker's detection box. In the process of action recognition, the key points cannot be recognized, which leads to the loss of key point data. Too much lost data leads to the inability to determine the construction status of workers and affects the statistics of each worker's construction efficiency.

This study is still somewhat limited. First, there are various types of actions at a construction site. This means that not all activity categories were considered, albeit a category ("other" label) unrelated to construction activities was considered. In addition, when workers leave the camera video range for a long period, resulting in object tracking loss, the proposed framework will assign a new ID number for action analysis.

5. Conclusions

A deep learning-based activity analysis framework is proposed to handle the very large amount of construction information involving multiple workers. The framework integrates key point extraction, tracking, activity recognition and efficiency analysis modules. The pose estimation detector processes 2D pose information obtained from RGB video feeds. The key point action dataset of a construction site is also proposed to complete the training of the activity classification network. The small amount of key point data can represent human actions and positions. The proposed framework can handle the identification of multiple worker activities recorded in videos and the statistics of construction efficiency. By validating the proposed procedure on construction monitoring videos collected from actual construction sites, the proposed framework is shown to be effective in monitoring the activities of construction workers. The experimental results show that the 2D pose information is useful for construction worker activity analysis and efficiency analysis.

For future studies, the focus will be on the more accurate and elaborate identification of worker action categories with different working types by adding tool identification. In this way, worker action categories can be more detailed and more conducive to the statistics of construction progress. Furthermore, cross-camera worker tracking and activity recognition may be explored because the scene of a construction site is usually large, and one camera has

a limited range. The results of action recognition can also be used for safety analysis by introducing more action categories that include unsafe behavior, such as climbing, jumping and falling, with timely detection. Furthermore, subsequent research can consider the prediction of worker actions; predict the occurrence of dangerous behaviors, such as fall prediction when workers lose their balance; perform real-time analysis of video data; and conduct timely generation of early warnings to alert workers and managers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (52130801, U20A20312, 52178271, and 52077213), and the National Key Research and Development Program of China (2021YFF0500903).

Compliance with ethics guidelines

Xuhong Zhou, Shuai Li, Jiepeng Liu, Zhou Wu, and Yohchia Frank Chen declare that they have no conflict of interest or financial conflicts to disclose.

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.

References

[1] Slaton T, Hernandez C, Akhavian R. Construction activity recognition with convolutional recurrent networks. Autom Constr 2020;113:103138.
[2] Kelm A, Laußat L, Meins-Becker A, Platz D, Khazaee MJ, Costin AM, et al. Mobie passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites. Autom Constr 2013;36:38–52.
[3] Cheng T, Teizer J, Migliaccio GC, Gatti UC. Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data. Autom Constr 2013;29:24–39.
[4] Kim J, Ham Y, Chung Y, Chi S. Systematic camera placement framework for operation-level visual monitoring on construction jobsites. J Constr Eng Manage 2019;145(4):04019019.
[5] Yang J, Shi Z, Wu Z. Vision-based action recognition of construction workers using dense trajectories. Adv Eng Inf 2016;30(3):327–36.
[6] Khosrowpour A, Niebles JC, Golparvar-Fard M. Vision-based workface assessment using depth images for activity analysis of interior construction operations. Autom Constr 2014;48:74–87.
[7] Han SU, Lee SH. A vision-based motion capture and recognition framework for behavior-based safety management. Autom Constr 2013;35:131–41.
[8] Yu N, Wang S. Enhanced autonomous exploration and mapping of an unknown environment with the fusion of dual RGB-D sensors. Engineering 2019;5(1):164–72.
[9] Luo X, Li H, Cao D, Dai F, Seo JO, Lee SH. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. J Comput Civ Eng 2018;32(3):04018012.
[10] Luo X, Li H, Cao D, Yu Y, Yang X, Huang T. Towards efficient and objective work sampling: recognizing workers' activities in site surveillance videos with two-stream convolutional networks. Autom Constr 2018;94:360–70.
[11] Pradhananga N, Teizer J. Cell-based construction site simulation model for earthmoving operations using real-time equipment location data. Visualization in Eng 2015;3(1):12.
[12] Montaser A, Moselhi O. RFID indoor location identification for construction projects. Autom Constr 2014;39:167–79.
[13] Akhavian R, Behzadan AH. Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers. Adv Eng Inf 2015;29(4):867–77.
[14] Zhao J, Obonyo E. Convolutional long short-term memory model for recognizing construction workers' postures from wearable inertial measurement units. Adv Eng Inf 2020;46:101177.
[15] Sanhudo L, Calvetti D, Martins JP, Ramos NMM, Mêda P, Gonçalves MC, et al. Activity classification using accelerometers and machine learning for complex construction worker activities. J Build Eng 2021;35:102001.
[16] Valero E, Sivanathan A, Bosché F, Abdel-Wahab M. Musculoskeletal disorders in construction: a review and a novel system for activity tracking with body area network. Appl Ergon 2016;54:120–30.
[17] Yan X, Li H, Li AR, Zhang H. Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention. Autom Constr 2017;74:2–11.
[18] Golabchi A, Han SH, Seo JO, Han SU, Lee SH, Al-Hussein M. An automated biomechanical simulation approach to ergonomic job analysis for workplace design. J Constr Eng Manage 2015;141(8):04015020.
[19] Kanan R, Elhassan O, Bensalem R. An IoT-based autonomous system for workers' safety in construction sites with real-time alarming, monitoring, and positioning strategies. Autom Constr 2018;88:73–86.
[20] Chi S, Caldas CH. Automated object identification using optical video cameras on construction sites. Comput Aided Civ Infrastruct Eng 2011;26(5):368–80.
[21] Seo JO, Lee SH, Seo J. Simulation-based assessment of workers' muscle fatigue and its impact on construction operations. J Constr Eng Manage 2016;142(11):04016063.
[22] Gatt T., Seychell D., Dingli A. Detecting human abnormal behaviour through a video generated model. In: 2019 11th International Symposium on Image and Signal Processing and Analysis; 2019 Sep 23–25; Dubrovnik, Croatia. Piscataway: IEEE; 2019. p. 264–70.
[23] Wang D, Li W, Liu X, Li N, Zhang C. UAV environmental perception and autonomous obstacle avoidance: a deep learning and depth camera combined solution. Comput Electron Agric 2020;175:105523.
[24] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25:1097–105.
[25] Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49.
[26] Donahue J, Hendricks LA, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T, et al. Long-term recurrent convolutional networks for visual recognition and description. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA. Piscataway: IEEE; 2015. p. 2625–34.
[27] Fang W, Ding L, Zhong B, Love PED, Luo H. Automated detection of workers and heavy equipment on construction sites: a convolutional neural network approach. Adv Eng Inf 2018;37:139–49.
[28] Kim H, Bang S, Jeong H, Ham Y, Kim H. Analyzing context and productivity of tunnel earthmoving processes using imaging and simulation. Autom Constr 2018;92:188–98.
[29] Fang Q, Li H, Luo X, Ding L, Luo H, Rose TM, et al. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. Autom Constr 2018;85:1–9.
[30] Zhang Q, Barri K, Babanajad SK, Alavi AH. Real-time detection of cracks on concrete bridge decks using deep learning in the frequency domain. Engineering 2021;7(12):1786–96.
[31] Ding L, Fang W, Luo H, Love PED, Zhong B, Ouyang X. A deep hybrid learning model to detect unsafe behavior: integrating convolution neural networks and long short-term memory. Autom Constr 2018;86:118–24.
[32] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[33] Luo X, Li H, Yang X, Yu Y, Cao D. Capturing and understanding workers' activities in far-field surveillance videos with deep action recognition and Bayesian nonparametric learning. Comput-Aided Civ Infrastruct Eng 2019;34(4):333–51.
[34] Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, et al. Temporal segment networks for action recognition in videos. IEEE Trans Pattern Anal Mach Intell 2018;41(11):2740–55.
[35] Chen C, Zhu Z, Hammad A. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. Autom Constr 2020;110:103045.
[36] Silva V, Soares F, Leão CP, Esteves JS, Vercelli G. Skeleton driven action recognition using an image-based spatial-temporal representation and convolution neural network. Sensors 2021;21(13):4342.
[37] Toshev A., Szegedy C. DeepPose: human pose estimation via deep neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition; 2014 Jun 23–28; Columbus, OH, USA. Piscataway: IEEE; 2014. p. 1653–60.
[38] Roberts D, Calderon WT, Tang S, Golparvar-Fard M. Vision-based construction worker activity analysis informed by body posture. J Comput Civ Eng 2020;34(4):04020017.
[39] Cao Z., Simon T., Wei S.E., Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 July 21–26; Honolulu, HI, USA. Piscataway: IEEE; 2017. p. 1302–10.
[40] Chen B, Hua C, Li D, He Y, Han J. Intelligent human–UAV interaction system with joint cross-validation over action–gesture recognition and scene understanding. Appl Sci 2019;9(16):3277.
[41] Okumura T, Urabe S, Inoue K, Yoshioka M. Cooking activities recognition in egocentric videos using hand shape feature with OpenPose. In: CEA/MADiMa '18: Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management; 2018 Jul 15; Stockholm, Sweden. New York: Association for Computing Machinery; 2018. p. 42–5.
[42] Chen S, Demachi K. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric

relationships and hierarchical scene graph. Autom Constr 2021;125 103619.
[43] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
[44] Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.C. MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. Piscataway: IEEE; 2018. p. 4510–20.
[45] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. Piscataway: IEEE; 2017. p. 1800–7.Xception C.F. Deep learning with depthwise separable convolutions. Honolulu, HI, USA. Piscataway: IEEE; 2017. p. 1800–1807.
[46] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: 2017 IEEE International Conference on Image Processing; 2017 Sep 17–20: eijing, China. Piscataway: IEEE; 2017. p. 3645-9.
[47] Wang Z, Zhang Q, Yang B, Wu T, Lei K, Zhang B, et al. Vision-based framework for automatic progress monitoring of precast walls by using surveillance videos during the construction phase. J Comput Civ Eng 2021;35 (1):04020056.