Automation in Construction 94 (2018) 1–10

Contents lists available at ScienceDirect

Automation in Construction

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

Ergonomic posture recognition using 3D view-invariant features from single ordinary camera

Hong Zhang[a], Xuzhong Yan[a,b,*], Heng Li[b]

[a] Institute of Construction Management, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China
[b] Department of Building and Real Estate, Faculty of Construction and Environment, the Hong Kong Polytechnic University, Hong Kong, China

ARTICLE INFO

Keywords:
Ergonomics
Person posture recognition
3D view-invariant
Relative 3D joint position
Joint angle
Convolutional neural network
Construction worker
Safety and health

ABSTRACT

Manual construction tasks are physically demanding, requiring prolonged awkward postures that can cause pain and injury. Person posture recognition (PPR) is essential in postural ergonomic hazard assessment. This paper proposed an ergonomic posture recognition method using 3D view-invariant features from a single 2D camera that is non-intrusive and widely installed on construction sites. Based on the detected 2D skeletons, view-invariant relative 3D joint position (R3DJP) and joint angle are extracted as classification features by employing a multi-stage convolutional nerual network (CNN) architecture, so that the learned classifier is not sensitive to camera viewpoints. Three posture classifiers regarding arms, back, and legs are trained, so that they can be simultaneously classified in one video frame. The posture recognition accuracies of three body parts are 98.6%, 99.5%, 99.8%, respectively. For generalization ability, the relevant accuracies are 94.9%, 93.9%, 94.6%, respectively. Both the classification accuracy and generalization ability of the method outperform previous vision-based methods in construction. The proposed method enables reliable and accurate postural ergonomic assessment for improving construction workers' safety and healthy.

1. Introduction

Work-related musculoskeletal disorders (WMSDs) are common occupational hazards in the manually demanding construction industry. In Hong Kong, the Pilot Medical Examination Scheme (PMES) for Construction Workers revealed that 41% of registered workers have musculoskeletal pain [1]. In the United States, the median days away from work due to WMSDs increased from 8 days in 1992 to 13 days in 2014; the proportion of WMSDs among workers aged 55 to 64 years doubled [2]. In an ergonomic perspective, it is suggested the frequency and duration of awkward postures regarding trunk, upper and lower limbs be controlled within acceptable ranges [3–6]. Traditional observational methods require safety personnel to collect posture data through site observations and questionnaires, which may be inaccuracy and inefficient due to subjective bias [7]. Such limitations are significant on construction sites due to insufficient capable workforce and continuously changing environments [8].

To address the problem of manual observation methods, there are currently two technical streams in person posture recognition (PPR), one is wearable sensor-based methods, and the other is computer vision-based methods. Compared with the wearable sensor system (e.g. YEI 3-Space Sensor), optical system (e.g. Vicon) and depth cameras (e.g., Microsoft Kinect), the ordinary surveillance camera (common camera) is more practical for ergonomic posture capture in construction, because it is non-intrusive and has been widely installed in the construction industry for surveillance [9]. In this study, the videos captured by a single 2D ordinary camera are depended for ergonomic posture recognition.

One challenge in PPR from a single ordinary camera is huge variances in the projection of intra-class postures, and similar projection of inter-class postures while being viewed from different 2D camera viewpoints [10]. Therefore, the objective of this study is to propose an ergonomic posture recognition method using view-invariant features, i.e. relative 3D joint positions (R3DJPs) [11] and joint angles that are estimated from 2D video frames captured by a single ordinary camera for field surveillance. To achieve the research objective, view-invariant 3D skeletons are estimated by lifting 2D coordinates into 3D using a multi-stage convolutional neural network (CNN). The captured body is divided into 16 3D skeletons with 12 joints and 5 end points. Then discriminative R3DJPs and joint angle features are extracted for arms, back and legs respectively for ergonomic posture classification based on quantitatively defined ergonomic postures according to a classical ergonomic rule, the Ovako Working Posture Analysis System (OWAS) [4]. Using the view-invariant features extracted from video frame samples,

* Corresponding author at: Institute of Construction Management, College of Civil Engineering and Architecture, Zhejiang University, Hangzhou, China.
E-mail address: 11512075@zju.edu.cn (X. Yan).

https://doi.org/10.1016/j.autcon.2018.05.033
Received 6 March 2018; Received in revised form 30 May 2018; Accepted 31 May 2018
Available online 06 June 2018
0926-5805/

this study tested different machine learning methods in terms of classification performances. According to the average classification accuracy in 5-fold cross-validations, three optimal classifiers are selected for ergonomic posture classification regarding arms, trunk and legs respectively.

The contributions of this study are threefold. First, the paper proposes a view-invariant ergonomic posture recognition approach using 3D body skeletons and joints estimated from single 2D camera. Second, view-invariant R3DJPs and joint angle features are used as the discriminative representation of ergonomic postures in each 2D video frame. Third, trained by 3D view-invariant features, the three classifiers regarding arms, back and legs are tested to be effective in various camera viewpoints, also outperform previous vision-based ergonomic posture recognition methods in terms of average accuracy and generalization ability [12–15].

2. Background

2.1. Postural ergonomic assessment rules

WMSDs in construction are mainly caused by compression, shear, tensile stress and muscle force repeatedly acting on load bearing tissues [16]. Most of these loadings in the construction industry can be attributed to repetitive works in prolonged awkward working postures, including overhead reaching, stooping and squatting [17, 18]. From a perspective of ergonomic assessment, the frequency and duration of non-ergonomic postures of targeted body parts should be monitored and controlled so as to identify the hazardous working pattern and job site layout [3–6]. Accordingly, many ergonomic assessment rules have been proposed for postural ergonomic hazards monitoring and control. Representative research on awkward posture assessment rules include the “Rapid Upper Limb Assessment” (RULA), an ergonomic assessment tool focusing on upper limbs [6]; the “Ovako Working Posture Analyzing System” (OWAS) for identifying and evaluating working postures [4]; the ISO 11226:2000 for determining the acceptable angles and holding times of working postures [5], and the EN 1005-4 as a guidance when designing machinery component parts in assessing machine-related postures and body movements, i.e. during assembly, installation, operation, maintenance, repair and dismantlement [19].

The objective of these ergonomic assessment rules is to provide a quantitative and systematic criterion to identify and control postural ergonomic hazards in workplace. For example, the OWAS classified the posture combinations of arms, back and legs and relative proportions of certain postures during work time into four action categories based on the risk assessment of WMSDs [4, 20]. In the OWAS, the action categories range from 1 that requires no corrective actions to 4 that needs corrective measures immediately. When the proportion of a certain posture during the observation period is larger than the frequency threshold defined by the OWAS, the action category changes from lower to higher, which indicates the urgency of corrective actions is increasing. Similar to monitoring non-ergonomic posture frequency, some ergonomic rules provide acceptable thresholds of angles and holding times of targeted body parts that are prone to WMSDs. For example, the ISO 11226:2000 provides ergonomic guidelines for workplace design or redesign of jobs and products with the basic concepts of ergonomics in general and working postures in particular. The international standard provides recommended limits for working postures without any, or only with minimal external force exertion, while considering joint angles and holding time aspects.

The ergonomic assessment rules require posture data as input for ergonomic hazards analysis and control in workplace. At the time when the ergonomic assessment rules were developed, posture data were mainly collected through observation and questionnaires [8]. These manual data collection methods are labor-intensive and time-consuming, inconsistent and unreliable due to subjective bias and inactive workers' participation [8, 21, 22], which impose practical constraints on ergonomic assessment in construction. With the fast development and iterations of motion data acquisition technology, the traditional manual observation and questionnaire methods can be replaced by many advanced motion capture technologies. However, postural ergonomic assessment rules do not fade and still play an essential role in safety and health management in construction. For example, Xinming, et al. [23] created a 3D model to imitate and animate manual construction tasks in a virtual environment based on the RULA. They analyzed body joint angles from 3D visualization based on the traditional ergonomic assessment rule to identify postural ergonomic hazards. Some researchers also applied the RULA to establish a virtual 3D workplace for proactive ergonomic design of construction workplace [24]. Based on the ISO 11226:2000, Yan, et al. [25] developed a real-time motion warning system that enables workers' self-management of ergonomic hazards in operational pattern using wearable Inertial Measurement Units (WIMUs). Compared with traditional work-related postural ergonomic assessment methods that focus on the design of ergonomic rules in workplace, some studies in construction focus more on the combination of the well-developed ergonomic assessment rules and the advanced motion capture technology considering specific industrial contexts [12, 25, 26], based on which automated, accurate and reliable ergonomic assessment can be performed for the monitoring and control of work-related ergonomic hazards.

In this research, typical awkward postures are defined based on the OWAS that have been validated in many jobs in different industrial contexts [4], as shown in Table 1. The frequency of each posture during work time is defined as different action categories ranging from 1 (no actions required) to 4 (corrective measures needed immediately). Once the frequency value of a recognized posture during a working period exceeds its limit, the corresponding action category will change from lower to higher, indicating the urgency of corrective ergonomic interventions.

2.2. Ergonomic posture capture systems

Ergonomic posture capture is a process used to detect, track and record an object's postures that involve ergonomic hazards, based on which automated postural ergonomic assessment and control can be conducted. Various motion capture technologies have been actively developed that can serve as the ergonomic posture capture system, including wearable sensor-based system, depth sensor-based system and ordinary surveillance camera-based system.

A wearable sensor-based system captures motion data by using a set of portable sensors attached on the targeted body parts of a wearer. The most popular wearable sensor in ergonomic assessment is the wearable Inertial Measurement Unit (WIMU) sensor. The WIMUs have satisfactory performance in accuracy [25, 27–29]. However, they need to be attached tightly to the wearer's body to prevent output noise caused by unstable adherence. It has been revealed by the front-line construction

Table 1
Action category considering frequency of postures [4].

% of working time		10	20	30	40	50	60	70	80	90	100
Arms	A. Both arms below shoulder level	1	1	1	1	1	1	1	1	1	1
	B. One arm at or above shoulder level	1	1	1	2	2	2	2	2	3	3
	C. Both arms at or above shoulder level	1	1	2	2	2	2	2	3	3	3
Back	A. Straight back	1	1	1	1	1	1	1	1	1	1
	B. Back bent	1	1	1	2	2	2	2	2	3	3
	C. Back bent heavily	1	2	2	3	3	3	3	4	4	4
Legs	A. Standing with one or both straight legs	1	1	1	2	2	2	2	2	3	3
	B. Knees bent	1	2	2	3	3	3	4	4	4	4
	C. Squatting	1	1	2	2	2	3	3	3	3	3

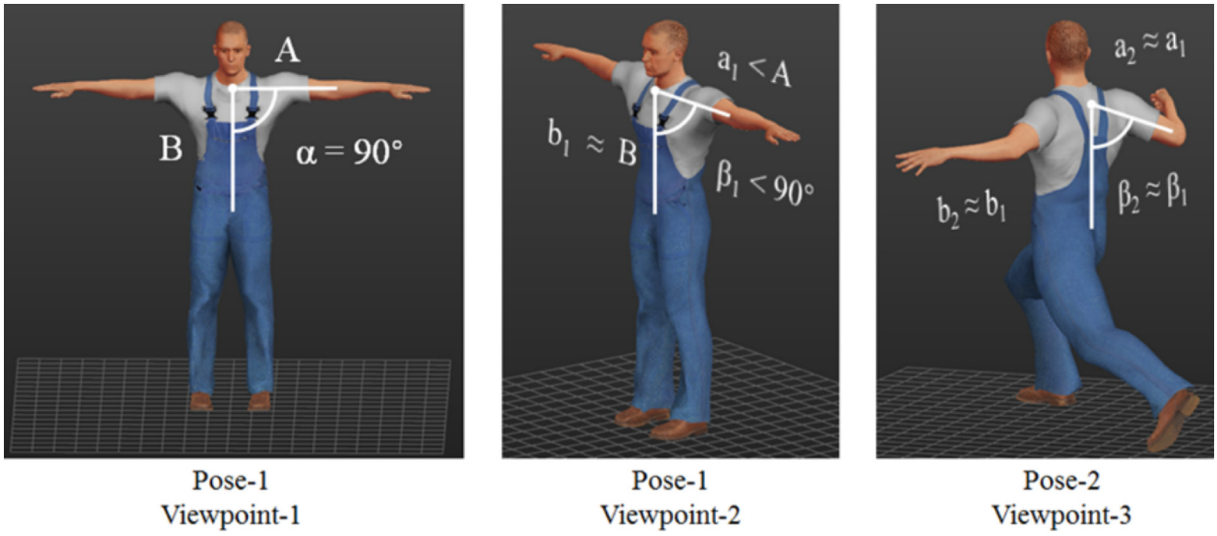

Fig. 1. View-invariance: Intra-class variability and inter-class similarity.

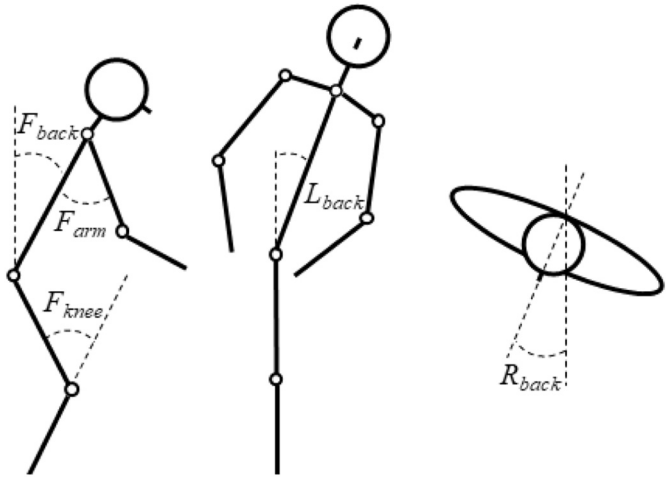

Fig. 2. Illustration of quantitative ergonomic posture.

Table 2
Quantitative ergonomic posture definition [12].

Body part	Posture description	Quantitative definition
Arms	A. Both below shoulder level	Both $\boldsymbol{F}_{\mathbf{arm}} \in [-90^{\circ}, +90^{\circ}]$
	B. One at or above shoulder level	One $\boldsymbol{F}_{\mathbf{arm}} \in (+90^{\circ}, +180^{\circ})$
	C. Both at or above shoulder level	Both $\boldsymbol{F}_{\mathbf{arm}} \in (+90^{\circ}, +180^{\circ})$
Back	A. Straight back	$\boldsymbol{F}_{\mathbf{back}} \in [-20^{\circ}, +20^{\circ}]$ and $\boldsymbol{L}_{\mathbf{back}}, \boldsymbol{R}_{\mathbf{back}} \in [-10^{\circ}, +10^{\circ}]$
	B. Back bent	$\boldsymbol{F}_{\mathbf{back}} \in (-60^{\circ}, -20^{\circ})$ or $(+20^{\circ}, +60^{\circ})$
	C. Back bent heavily	$\boldsymbol{F}_{\mathbf{back}} \in (-90^{\circ}, -60^{\circ})$ or $(+60^{\circ}, +180^{\circ})$
Legs	A. Standing	One or Both $\boldsymbol{F}_{\mathbf{knee}} \in (0^{\circ}, +20^{\circ})$
	B. Knees bent	One or Both $\boldsymbol{F}_{\mathbf{knee}} \in (+20^{\circ}, +90^{\circ})$
	C. Squatting	Both $\boldsymbol{F}_{\mathbf{knee}} \in (+90^{\circ}, +180^{\circ})$

workers that wearing tight sensor-based equipment is a physical burden in a prolonged working hours, which may interfere with normal manual operation and reduce productivity [30].

A depth sensor-based system can capture depth information of a scene by emitting a structured infrared laser in a grid form, analyzing the distortion of the infrared image and computing the depth based on disparity retrieval [31]. As a commercial depth sensor-based system, the Microsoft Kinect with an OpenNI software development kit (SDK) also allows for extraction of human body skeleton models from depth video frames. The Kinect can detect and track human body skeletons based on a deep randomized decision forest classifier [32], based on which PPR can be achieved using posture features. Many previous studies used the depth sensor-based system (e.g. Microsoft Kinect) as the posture capture device for ergonomic posture recognition in construction [15, 33–35]. For example, Ray and Jochen [13] rescaled the images captured by the Kinect to grayscale image and reshape each of them into a large size row vector as features for ergonomic posture classification. However, the Kinect is vulnerable to outdoor conditions where a certain amount of solar IR and ferromagnetic radiation can cause significant noises [36], even wash out the scene generated by the Kinect. Besides, the Kinect has higher power consumption, lower resolution, and is not as widely and cheaply available as ordinary cameras [37].

An ordinary camera-based system only uses a common color camera to capture 2D images without depth information. Compared with the depth sensor-based motion capture system, ordinary cameras are cheaper and more stable in an outdoor construction environment, and have already been widely used in construction so that such a camera is not device-intensive for most construction sites. To this end, the ordinary camera is used in this study as posture capture device for ergonomic PPR in construction.

Body skeleton estimation from a single camera is a severely underconstrained and much more challenging problem compared with posture estimation from wearable sensors or a depth camera. Monocular RGB body skeleton estimation in 2D has been widely studied. There are many representations of human posture in 2D video frames captured by the ordinary color camera [38], such as appearance-based [39], volume-based [40] and interest-point-based [41]. Since body joint positions are rich representations for PPR in computer vision [42], they are captured to represent human posture in 2D video frames using the Convolutional Pose Machines [43] in this study thanks to its performance and efficiency as a benchmark in body skeleton estimation. The estimated body skeletons and joint positions in 2D video frames captured by a single ordinary camera can be further used for ergonomic PPR in construction.

3. View-invariant PPR feature extraction from 2D video frames

The ordinary camera cannot capture depth information. As a result, it suffers from view-invariance, i.e., inter-class similarity and intra-class variability of 2D skeletons and joints due to view variance [10], as

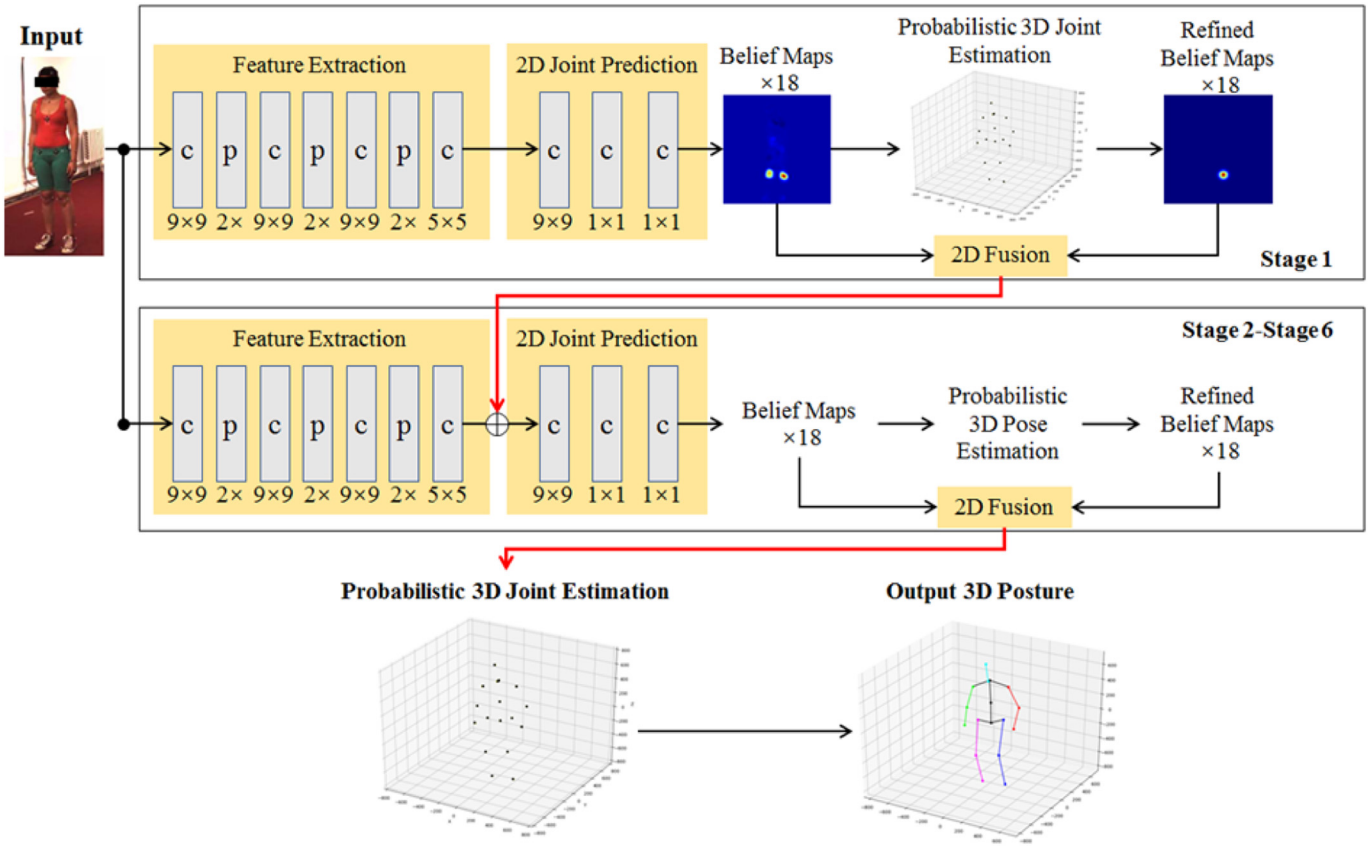

Fig. 3. The multi-stage CNN for 3D joint locations lifting from 2D imagery ('c' denotes convolutional layer, 'p' denotes pooling layer).

shown in Fig. 1. View variance produced three challenges to PPR from 2D video frames. The first challenge is intra-class variability of postures in the same category. Homogeneous postures may be perceived as heterogeneous geometric structure as perceived in different directions and distances with respect to the camera (see Viewpoint-1 and Viewpoint-2 in Fig. 1). The second challenge is inter-class similarity of postures from distinct categories. Distinctive characteristics of body segments movements may be only distinguished by subtle details (see Viewpoint-2 and Viewpoint-3 in Fig. 1). Third, the number of viewpoints is theoretically infinite. It is impossible to cover all viewpoints in real-world training samples. The recognition performance of static postures in a 2D image is more vulnerable to these challenges due to a lack of additional temporal or spatial information.

Geometrically, posture features are a projected set of points and line segments. The utility of a feature of the projection is a function of how it varies with different viewpoints. Though it has been proved by [44] that general-case view-invariants do not exist for a given number of points given true perspective, weak perspective or orthographic projection models, there are approximate view-invariant representations of postures captured by a single ordinary color camera, based on which PPR can be conducted in a principled manner. To extract view-invariant features for action recognition, Kong, et al. [45] used a sample-affinity matrix method to measure the similarities between different video frames in multiple views. Another work represented an action as triplets of points by decomposing posture into a set of point-triplets [46]. In their study, view invariant features are defined by triplets across two video frames. In our previous work, probability density of 2D angle and length ratio feature are extracted as view-invariant features for ergonomic posture recognition [12]. Compared with 2D view-invariant features, 3D features are types of features whose view variations are sufficiently lower to make discrimination training more feasible [10]. Hence, in order to improve the accuracy and ability of generalization in vision-based ergonomic posture recognition, this paper aims to extract 3D view-invariant features from 2D video frames captured by a single surveillance camera to learn ergonomic posture classifiers and test their performance in ergonomic PPR in construction.

4. Feature extraction and classifier training experiments

This section illustrated the proposed methodologies regarding view-invariant ergonomic posture recognition in a single ordinary camera, including view-invariant feature extraction for classification, classifiers training and selection of the optimal classifiers.

Since this study defines ergonomic postures according to the OWAS [4, 47] (please refer to Table 1 and Fig. 2 [12]). Nine postures that are frequently encountered by construction workers are quantitatively defined in Table 2: $-\boldsymbol{F}_{\mathbf{back}}$ indicates backward extension; $+\boldsymbol{F}_{\mathbf{back}}$ indicates forward flexion; $-\boldsymbol{L}_{\mathbf{back}}$ indicates left lateral bending; $+\boldsymbol{L}_{\mathbf{back}}$ indicates right lateral bending; $-\boldsymbol{R}_{\mathbf{back}}$ indicates right rotation; $+\boldsymbol{R}_{\mathbf{back}}$ indicates left rotation; $-\boldsymbol{F}_{\mathbf{arm}}$ indicates backward stretch; $+\boldsymbol{F}_{\mathbf{arm}}$ indicates forward stretch; $\boldsymbol{F}_{\mathbf{knee}}$ indicates knee bending.

To guarantee the comparability of classifier performance between the developed ergonomic posture classification method based on 3D view-invariant features and the method based on 2D view-invariant features [12], the classifiers must be trained by the same posture dataset. Therefore, the posture dataset that were collected in the previous study for training is employed in this study. The posture dataset were collected and annotated based on the quantitative posture definition as shown in Table 2. In the data collection procedure of this dataset, three heights of camera lens above the ground level were set. For each camera height, eight camera viewpoints were set by dividing the circle that has a center point of the participant into eight equal areas. The reason that only several viewpoints were selected for sampling is to leave room for testing the generalization ability of the trained classifiers

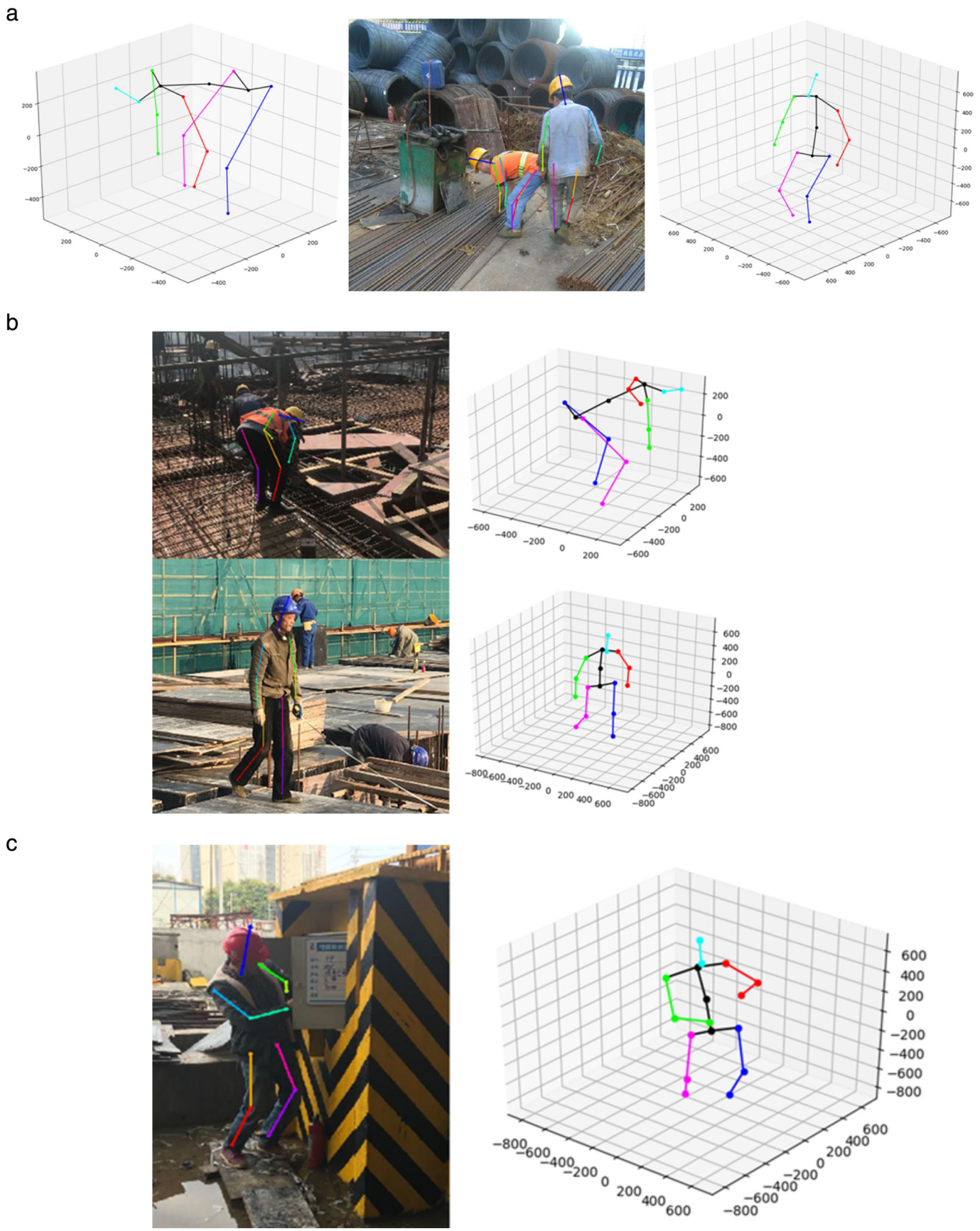

Fig. 4. Some tests of the multi-stage CNN architecture on a construction site. (a) Two workers' 3D skeletons are captured simultaneously in one frame. (b) One worker's 3D skeletons are captured in one frame, while workers in further distance to the camera are not estimated. (c) One worker's 3D skeletons are captured in one frame, while the worker's left arm is occluded.

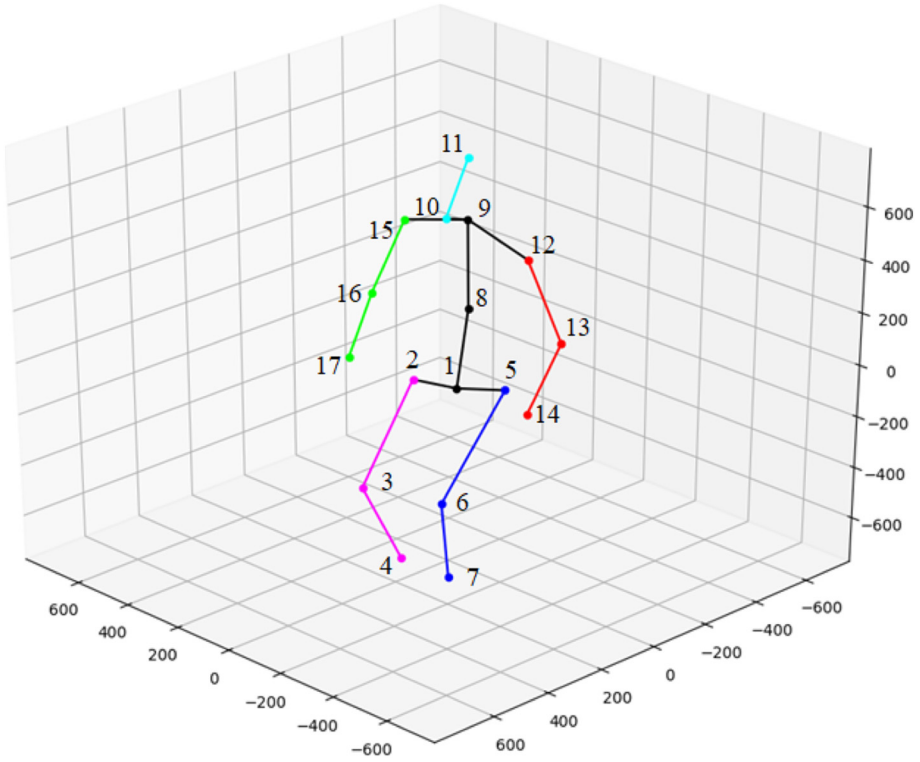

Fig. 5. Numbered joints for 3D view-invariant feature extraction.

Table 3
R3DJP and joint angle features of three body parts.

Body part	R3DJP	Joint angle
Arms	8, 9, 10, 12, 13, 14, 15, 16, 17.	∠1516-1509, ∠1516-0908, ∠0915-0908, ∠0908-0912, ∠0908-1213, ∠1209-1213.
Back	1, 2, 3, 5, 6, 8, 9, 10, 11.	∠0809-0801, ∠0108-0203, ∠0108-0506, ∠0809-0203, ∠0809-0506, ∠0910-0801, ∠0910-0203, ∠0910-0506.
Legs	1, 2, 3, 4, 5, 6, 7.	∠0302-0304, ∠0304-0201 ∠0203-0201, ∠0605-0607, ∠0607-0501, ∠0506-0501.

Note: the numbers in the second column indicate the numbered joints for each body part in Fig. 5; the '∠' indicates joint angle between two skeletons that have directions. E.g. ∠1516-1509 indicates the joint angle between skeleton 15-16 and 15-09. Here, the direction of 1516 is from 15 to 16; the direction of 1509 is from 15 to 09.

in new posture data captured by new camera viewpoints. In addition, the recruited participants in this posture data collection procedure were asked to wear Inertial Measurement Unit (IMU) sensors on targeted body parts in the process of video recording. The outputted quaternions of each IMU sensor are translated into body joint angle according to [12]. Based on the translated body joint angles, all the collected posture samples were annotated as different combinations of arms, back and legs postures according to the posture definition in Table 2.

4.1. 3D joint location estimation

Compared with 2D view-invariant features, 3D features have sufficiently lower variations to make discrimination training more feasible [10]. Hence, to improve the accuracy and ability of generalization in vision-based ergonomic posture recognition, this study estimates 3D skeletons and joints from 2D video frames by using a multi-stage CNN architecture that combines the Convolutional Pose Machine [43] and a probabilistic 3D joint estimation model [48], as illustrated in Fig. 3.

The overall CNN architecture is trained end-to-end using back-propagation. In each stage of this deep architecture, belief maps of 2D joint locations of each input image are firstly detected by feature extraction layers and 2D joint prediction layers provided by the Convolutional Pose Machine [43]. Both feature extraction layers and 2D joint prediction layers consist of several convolutional layers and pooling layers. The Convolutional Pose Machine is a deep convolutional architecture for body joints estimation in 2D imagery. In this architecture, at each stage t and for each landmark p, the dense per pixel belief maps $b_t^p[u,v]$ can be calculated to show how confident it is that a body joint landmark occurs in any pixel (u,v) of a given image. The 2D joint estimation architecture in this study is initialized by using the weights with the parameters found in the Convolution Pose Machine model for all pre-existing layers with the new layers randomly initialized. During training on the Human 3.6 M dataset [11], the most confident pixels as the location of each landmark is selected, so that the belief maps can be transformed to locations for further 3D skeletons estimation:

$$Y_p = \arg\max_{(u,v)} b_p[u,v] \tag{1}$$

Then, each stage learns to combine the belief maps estimated by the Convolutional Pose Machine as well as the refined belief maps estimated by a probabilistic 3D joint estimation model that encodes geometric 3D skeletal information. For stages after the first one, the belief maps are a function of the information contained in the current stage and the information computed by the previous stage. Additional geometric 3D skeletal information is added because 3D joint location estimation from 2D joint location is an ill-posed problem due to the infinite space of possible 3D joint locations consistent with 2D joint locations. Thus, the probabilistic 3D joint estimation model is embedded in the multi-stage CNN as an additional layer that is responsible for lifting 2D joint coordinates into 3D considering additional geometric 3D skeletal information. The probabilistic 3D joint estimation model is trained on a dataset of 3D posture data [11], with more details of

Table 4
Comparison of different supervised classification learner (* denotes maximum classification accuracy).

Body part	Classification learner	Optimal parameters and functions	Accuracy
Arms	BP-ANN	I. 3 hidden layers (10-10-4 neurons) II. Transfer: tansig III. Training: Levenberg-Marquardt IV. Performance: MSE	97.9%
	DT	I. Maximum number of splits: 1000 II. Split criterion: Twoing rule III. Surrogate decision splits: On IV. Maximum number of surrogate: 5	97.3%
	SVM	I. Kernel: Gaussian II. Box constraint level: 4 III. Kernel scale: Heuristic subsampling IV. Multiclass method: one-vs-one	98.3%
	KNN	I. Number of neighbors: 10 II. Distance metric: City block III. Distance weight: Squared inverse	92.5%
	EC	I. Ensemble method: Bag II. Learner type: Decision trees III. Maximum number of splits: 5 IV. Number of learners: 30 V. Learning rate: 0.1 VI. Subspace dimension: 1	98.6%*
Back	BP-ANN	I. 3 hidden layers (10-10-10 neurons) II. Transfer: Tansig III. Training: Levenberg-Marquardt IV. Performance: MSE	99.3%
	DT	I. Maximum number of splits: 110 II. Split criterion: Maximum deviance reduction III. Surrogate decision splits: Off	98.4%
	SVM	I. Kernel: Gaussian II. Box constraint level: 2 III. Kernel scale: Heuristic subsampling IV. Multiclass method: one-vs-one	99.5%*
	KNN	I. Number of neighbors: 4 II. Distance metric: City block III. Distance weight: Squared inverse	99.4%
	EC	I. Ensemble method: Bag II. Learner type: Decision trees III. Maximum number of splits: 3 IV. Number of learners: 29 V. Learning rate: 0.1 VI. Subspace dimension: 1	99.4%
Legs	BP-ANN	I. 2 hidden layers (10-5 neurons) II. Transfer: Tansig III. Training: Levenberg-Marquardt IV. Performance: MSE	99.6%
	DT	I. Maximum number of splits: 60 II. Split criterion: Maximum deviance reduction III. Surrogate decision splits: Off	99.2%
	SVM	I. Kernel: Gaussian II. Box constraint level: 2 III. Kernel scale: Heuristic subsampling IV. Multiclass method: one-vs-one	99.8%*
	KNN	I. Number of neighbors: 4 II. Distance metric: Euclidean III. Distance weight: Squared inverse	99.5%
	EC	I. Ensemble method: Bag II. Learner type: Decision trees III. Maximum number of splits: 5 IV. Number of learners: 10 V. Learning rate: 0.1 VI. Subspace dimension: 1	99.7%

training given in [48]. This model first seeks the compact low-rank approximation of the 3D posture data in which all reconstructed poses appear to have the same orientation that allows for rotational invariance in the ground-plane. It then establishes a mixture of probabilistic PCA bases of the 3D posture data, based on which the most probable sample from the model that could give rise to a projected 2D image can be estimated [49]. The estimated 3D joint locations are then projected on to a new set of 2D belief maps to correct the beliefs of landmark locations at each stage. The newly projected belief maps $\hat{b}_t^p$ at each landmark p and stage t and the previous Convolutional Pose Machine-based belief maps b_t^p are fused according to the equation:

$$f_t^p = w_t \cdot b_t^p + (1 - w_t) \cdot \hat{b}_t^p \quad (2)$$

where $w_t \in [0,1]$ is the weight learned as a part of the end-to-end training. The fused belief landmarks f_t are then passed to the next stage and used as input to the refine the 2D joint estimation, as indicated by the red arrows. The 2D fusion of belief maps in the stage 6 are then lifted into 3D joint locations.

The advantage of this architecture is that the 2D joint location estimations are improved by guaranteeing that they satisfy with the anatomical 3D constraints encapsulated in the 3D posture structure. As a result, both 2D estimation and 3D estimation benefit from each other. Another advantage is that the 2D and 3D training data sources could be independent so that the posture dataset of both 2D and 3D can be augmented independently. This multi-stage CNN integrated with a probabilistic 3D joint estimation model outperform other methods [50, 51] by a 4.7 mm average improvement. For joint position estimation, the average error is 89.33 mm using *mean per joint position error* (MPJPE) metric; For joint angle estimation, the average error in 11.26° using *mean per joint angle error* (MPJAE) metric. This architecture is further validated on a construction site, as shown in Fig. 4. This paper uses the architecture to extract 3D joint position and angle features from 2D video frames for view-invariant PPR for ergonomic assessment in construction. It should be noted that multi-object detection can be achieved when the objects are captured within a certain distance to the camera (approximately 6 m) and without serious occlusion.

4.2. Relative 3D joint position and angle feature extraction

This study tested the performance of posture recognition using joint position and angle as basic features of human posture as they are rich representations for PPR in computer vision [42, 52]. Specifically, relative 3D joint position (R3DJP) and joint angles are extracted as features to learn three classifiers regarding arms, back and legs. The parametrization is called relative when there is a root joint (No. 1 in Fig. 5), while the positions of other joints are estimated relative to it. This manipulation can make the 3D joint position features be representative and discriminative for training. Joint angles ware extracted between limbs for it is invariant to both scale and body proportions. To achieve the goal of recognizing postures of three body parts simutaneously per video frame, a specific set of features ware selected for each body part regarding arms, back and legs. All the extracted features prepared to be selected for training are listed in Table 3 as a feature pool. The extraction is based on the numbered 3D joints in

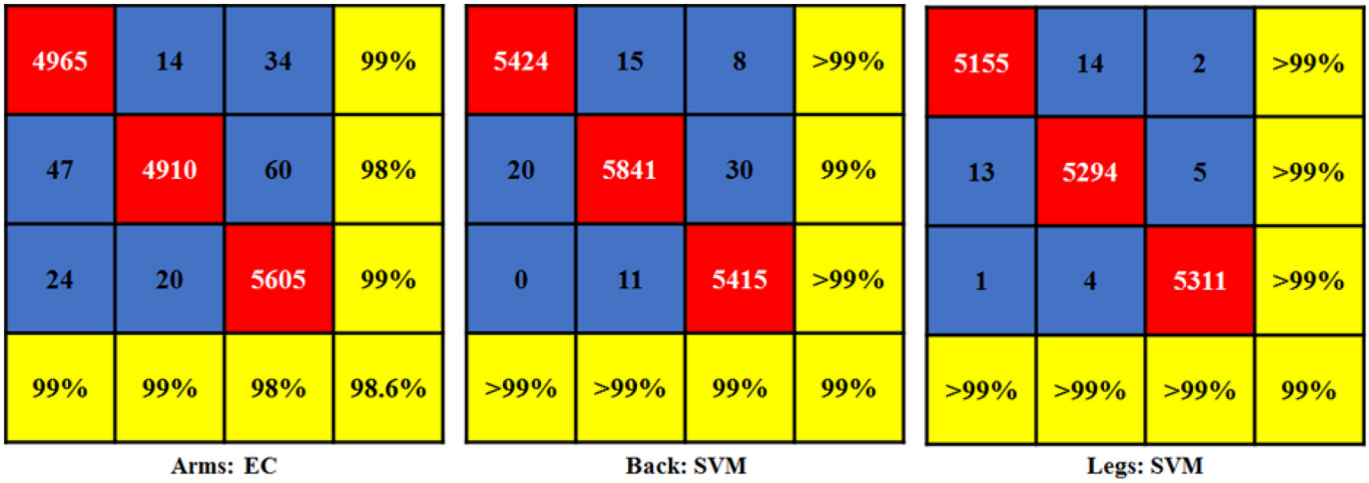

Fig. 6. Three confusion matrixes in 5-fold cross validation during training. Left column: True Positive Rate; Bottom row: Positive Predictive Value; Bottom right: F_1 score.

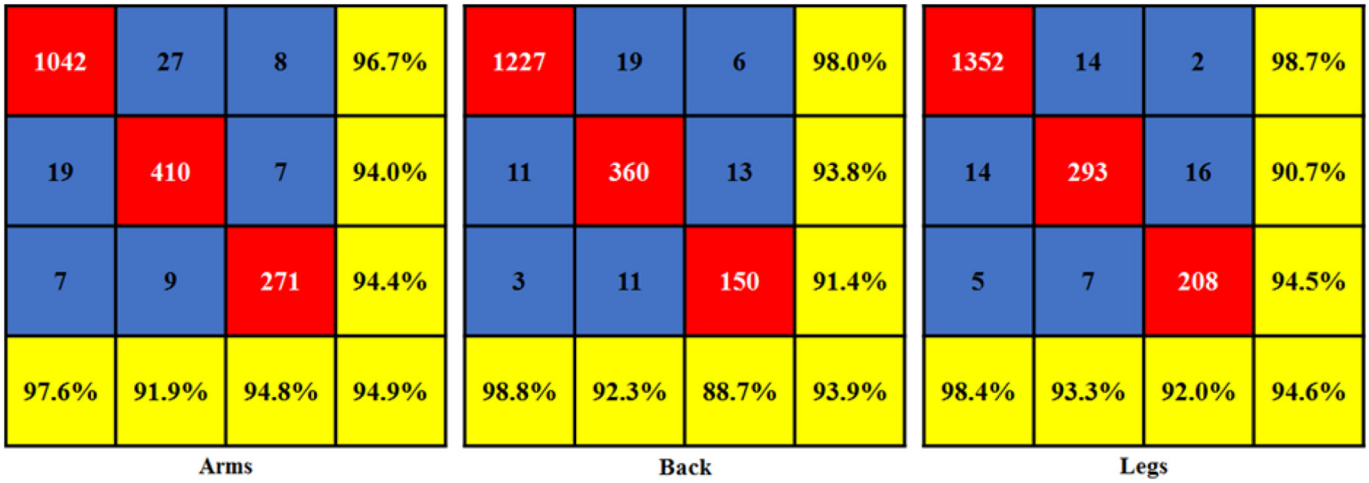

Fig. 7. Three confusion matrixes in 5-fold cross validation during testing. Left column: True Positive Rate; Bottom row: Positive Predictive Value; Bottom right: F_1 score.

Fig. 5. It should be noted that the each extracted feature is related only to one targeted body part, so that the recognition of three body parts would not interfere with each other.

It should also be noted that the study has also tested linear discriminant analysis (LDA) as dominant features extracted from 3D joint positions that are proposed by Lu, et al. [10], but the this additional data processing regarding feature extraction did not bring better performance in training the ergonomic posture classifier compared with the R3DJP and joint angle features. As a result, we only extract the three set of R3DJP and joint angle features for three body parts as the discriminative features for calssificaiton.

4.3. Supervised ergonomic posture classification

Since different supervised classification learners have various performances on the different data structure, this study learns the classifier by testing several types of mature machine learning models, including back-propagation artificial neural network (BP-ANN), decision trees (DT), support vector machines (SVM), K-nearest neighbor classifiers (KNN) and ensemble classifiers (EC). For each supervised learning method, different features, model parameters and functions are selected by comparing intragroup performance on the same feature set. The final accuracy of each classification learner is tested by a 5-fold cross validation (7:3). Then, the optimal classifier can be determined by comparing intergroup accuracy. The selected classification learner and corresponding features, parameters, functions and accuracy (> 90%) are listed in Table 4. According to this procedure, the optimal classifier for arms is the bagged decision trees with an average classification accuracy of 98.7%. For back, SVM outperforms others with an average classification accuracy of 99.5%. For legs, SVM outperforms others with an average classification accuracy of 99.8%. The corresponding total confusion matrixes are shown in Fig. 6. Each row of the confusion matrix represents the instances in an actual posture class, while each column represents the instances in a predictive posture class. Within a 6-meter detection range, the average accuracy of the proposed method outperforms the previous 2D view-invariant features-based method and other vision-based ergonomic posture classification methods.

5. Generalization and discussions

To guarantee the comparability of generalization ability of the classifiers between the developed 3D view-invariant features-based ergonomic posture classification method and the previous 2D view-invariant feature-based method [12], this study employed the same dataset for testing that were collected in the previous method. The procedure to collect the dataset for testing generalization ability is same as the procedure to collect dataset for training classifiers except for using new camera locations and participants. The confusion matrixes of all

the new testing data regarding three body parts are shown in Fig. 7.

The overall accuracy of the three classifiers is 94.9%, 93.9% and 94.6% respectively. This indicates that the 3D view-invariant features-based classifiers can be generalized to different camera viewpoints. The overall accuracy outperforms the previous 2D view-invariant feature-based method [12], indicating the potentials of using 3D view-invariant features for more accurate ergonomic posture classification in safety and health management in construction. In addition, this trained model can recognize ergonomic postures with a frame rate of 12 frames per second (fps) from a 2D video on a single GPU. Such frame rate can guarantee near real-time posture recognition.

Based on comparison between the prediction and the ground truth, it is found that some frequently occurred reasons may lead to misclassification. For the leg classifier, the discrimination between leg posture "knees bent" and "squatting" is vulnerable when the participant squated with his back to the camera. This discrimination is a result of occlusion of legs, which is a major obstacle that affects the classification performance. Another limitation is that during the feature extraction procedure, the joint position and joint angle of occluded skeletons are only simplified as zero. This may also affect the accuracy of the posture classifier due to a lack of skeleton position information. To cope with these problems, future work would be conducted to collect more posture data samples for training, meanwhile improve the structure of the 3D joint location estimation model to estimate the occluded joint positions. In addition to posture, force is another very important variable for biomechanical and ergonomic assessment. Without using wearable sensors to directly measure force, it could be estimated by the speed and frequency of postures from consecutive video frames, which would also be a research direction in our future work.

6. Conclusions

This paper has presented a PPR method for assessment of postural ergonomic hazards in construction. To improve the accuracy and ability of generalization in vision-based ergonomic posture recognition, this study estimates 3D skeletons and joints from 2D video frames by using a multi-stage CNN architecture that combines the Convolutional Pose Machine [43], a probabilistic 3D joint estimation model [48], a relative 3D joint position and angle feature extraction component, and a posture classification component that is trained end-to-end. The technical advance of this study is to extract 3D view-invariant features for near real-time simultaneous non-ergonomic postures recognition of three body parts in single 2D camera. Considering view variance of projection during video recording, this study has extracted R3DJP and 3D joint angles as training features, whose view variations are sufficiently low to make discrimination training feasible. Then, three classifiers in terms of arms, back and legs are trained using different supervised machine learning methods. According to the average classification accuracy in 5 cross-validation, three optimal classifiers in terms of arms, back and legs that outperform other classifiers are selected. Finally, an experiment is conducted to test the generalization ability of the selected classifiers. The results show that the proposed posture classification method outperforms previous vision-based ergonomic posture recognition methods in terms of overall accuracy and generalization ability. Hence, the method has potentials in reliable postural ergonomic assessment to improve workers' safety and healthy in construction.

This study applies a deep CNN-based learning in 3D ergonomic posture extraction and recognition. There are three contributions in this study: First, view-invariant 3D ergonomic postures recognition in single 2D camera; Second, view-invariant R3DJP and joint angle features extraction in single 2D image; third, the average posture recognition accuracy of the proposed method is higher than previous vision-based ergonomic posture recognition methods. The contributions of this study could lead to automated postural ergonomic monitoring and assessment at outdoor construction sites where ordinary cameras are widely installed. The proposed method can also be employed in other worker behavior-based analysis by changing posture definitions and selecting corresponding machine learning classifiers.

Acknowledgements

The project is supported by the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130101110058). The authors would also like to acknowledge the funding support of Hong Kong Construction Industry Council through funding a research project titled "Waistband Enabled Construction Workers Low Back Health Monitoring System" (A/C K-ZB64).

References

[1] W. Yi, A. Chan, Health profile of construction workers in Hong Kong, Int. J. Environ. Res. Public Health 13 (12) (2016) 1–15, http://dx.doi.org/10.3390/ijerph13121232.
[2] X. Wang, X.S. Dong, S.D. Choi, J.M. Dement, Work-related musculoskeletal disorders among construction workers in the United States from 1992 to 2014, Occup. Environ. Med. 74 (5) (2017) 374–380, http://dx.doi.org/10.21275/v5i5.nov163238.
[3] K. Schaub, Ergonomics of manual handling-part 1: lifting and carrying, in: W. Karwowski (Ed.), Handbook on Standards and Guidelines in Ergonomics and Human Factors, CRC Press, 2005, pp. 255–269 (ISBN: 9781482289671).
[4] K. Osmo, P. Kansi, I. Kuorinka, Correcting working postures in industry: a practical method for analysis, Appl. Ergon. 8 (4) (1977) 199–201, http://dx.doi.org/10.1016/0003-6870(77)90164-8.
[5] N. Delleman, M. Boocock, B. Kapitaniak, P. Schaefer, K. Schaub, ISO/FDIS 11226: evaluation of static working postures, Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 44(35) 2000, pp.) 442–443, , http://dx.doi.org/10.1177/154193120004403512.
[6] L. McAtamney, E.N. Corlett, RULA: a survey method for the investigation of work-related upper limb disorders, Appl. Ergon. 24 (2) (1993) 91–99, http://dx.doi.org/10.1016/0003-6870(93)90080-s.
[7] L. Straker, A. Campbell, J. Coleman, M. Ciccarelli, W. Dankaerts, In vivo laboratory validation of the physiometer: a measurement system for long-term recording of posture and movements in the workplace, Ergonomics 53 (5) (2010) 672–684, http://dx.doi.org/10.1080/00140131003671975.
[8] L. Heikki, M. Marjamäki, K. Päivärinta, The validity of the TR safety observation method on building construction, Accid. Anal. Prev. 31 (5) (1999) 463–472, http://dx.doi.org/10.1016/s0001-4575(98)00084-0.
[9] Q. Fang, H. Li, X.C. Luo, L. Ding, H. Luo, T.M. Rose, W. An, Detecting non-hardhat-use by a deep learning method from far-field surveillance videos, Autom. Constr. 85 (2017) 1–9, http://dx.doi.org/10.1016/j.autcon.2017.09.018.
[10] X. Lu, C. Chen, J.K. Aggarwal, View invariant human action recognition using histograms of 3D joints, Computer Vision and Pattern Recognition Workshops in IEEE Computer Society Conference, 2012, pp. 20–27, , http://dx.doi.org/10.1109/cvprw.2012.6239233.
[11] C. Ionescul, D. Papaval, V. Olarul, C. Sminchisescu, Human 3. 6M: large scale datasets and predictive methods for 3D human sensing in natural environments, IEEE Trans. Pattern Anal. Mach. Intell. 36 (7) (2014) 1325–1339, http://dx.doi.org/10.1109/tpami.2013.248.
[12] X.Z. Yan, H. Li, C. Wang, J. Seo, H. Zhang, H. Wang, Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion, Adv. Eng. Inform. 34 (2017) 152–163, http://dx.doi.org/10.1016/j.aei.2017.11.001.
[13] S.J. Ray, T. Jochen, Real-time construction worker posture analysis for ergonomics training, Adv. Eng. Inform. 26 (2) (2012) 439–455, http://dx.doi.org/10.1016/j.aei.2012.02.011.
[14] J. Seo, K. Yin, S. Lee, Automated postural ergonomic assessment using a computer vision-based posture classification, ASCE Constr. Res. Congr. (2016) 809–818, http://dx.doi.org/10.1061/9780784479827.082.
[15] H. SangUk, L. SangHyun, P.M. Feniosky, Vision-based detection of unsafe actions of a construction worker: case study of ladder climbing, J. Comput. Civ. Eng. 27 (6) (2013) 635–644, http://dx.doi.org/10.1061/(asce)cp.1943-5487.0000279.
[16] J.H.V. Dieën, M.J.M. Hoozemans, H.M. Toussaint, Stoop or squat: a review of biomechanical studies on lifting technique, Clin. Biomech. 14 (10) (1999) 685–696, http://dx.doi.org/10.1016/s0268-0033(99)00031-5.
[17] W. Umer, H. Li, G.P.Y. Szeto, A.Y.L. Wong, Identification of biomechanical risk factors for the development of lower-back disorders during manual rebar tying, J. Constr. Eng. Manag. 143 (1) (2016) 04016080, , http://dx.doi.org/10.1061/(ASCE)CO.19437862.0001208.
[18] C. Jiayu, Q. Jun, A. Changbum, Construction worker's awkward posture recognition through supervised motion tensor decomposition, Autom. Constr. 77 (2017) 67–81, http://dx.doi.org/10.1016/j.autcon.2017.01.020.
[19] N.J. Delleman, J. Dul, International standards on working postures and movements ISO 11226 and EN 1005-4, Ergonomics 50 (11) (2007) 1809–1819, http://dx.doi.org/10.1080/00140130704167443O.
[20] M. Mattila, M. Vilkki, OWAS methods, in: W. Karwowski, W.S. Marras (Eds.), Occupational Ergonomics: Principles of Work Design, CRC Press, 2003, pp. 1–11 Chapter26. (ISBN: 13: 978-0-203-50792-6).

[21] R.E. Levitt, N.M. Samelson, Construction Safety Management, John Wiley & Sons, New York, 1993 (ISBN: 978-0-471-59933-3).
[22] S. Han, S. Lee, F. Peña-Mora, Application of dimension reduction techniques for motion recognition: construction worker behavior monitoring, Comput. Civil Eng. (2011) 102–109, http://dx.doi.org/10.1061/41182(416)13.
[23] L. Xinming, S. Han, M. Gül, M. Al-Hussein, M. El-Rich, 3D visualization-based ergonomic risk assessment and work modification framework and its validation for a lifting task, J. Constr. Eng. Manag. 144 (1) (2017) 04017093, , http://dx.doi.org/10.1061/(asce)co.1943-7862.0001412.
[24] A. Golabchi, S. Han, J. Seo, S. Han, S. Lee, M. Al-Hussein, An automated biomechanical simulation approach to ergonomic job analysis for workplace design, J. Constr. Eng. Manag. 141 (8) (2015) 04015020, , http://dx.doi.org/10.1061/(asce)co.1943-7862.0000998.
[25] X.Z. Yan, H. Li, A.R. Li, H. Zhang, Wearable IMU-based real-time motion warning system for construction workers' musculoskeletal disorders prevention, Autom. Constr. 74 (2017) 2–11, http://dx.doi.org/10.1016/j.autcon.2016.11.007.
[26] J. Seo, R. Starbuck, S. Han, S. Lee, T.J. Armstrong, Motion data-driven biomechanical analysis during construction tasks on sites, J. Comput. Civ. Eng. 29 (4) (2014), http://dx.doi.org/10.1061/(ASCE)CP.1943-5487.
[27] M.C.Jr. Schall, N.B. Fethke, H. Chen, S. Oyama, D.I. Douphrate, Accuracy and repeatability of an inertial measurement unit system for field-based occupational studies, Ergonomics 59 (4) (2016) 591–602, http://dx.doi.org/10.1080/00140139.2015.1079335.
[28] A. Alwasel, M. Nahangi, C. Haas, E. Abdel-Rahman, Level-of-expertise classification for identifying safe and productive masons, Comput. Civil Eng. (2017) 359–368, http://dx.doi.org/10.1061/9780784480823.043.
[29] W. Lee, E. Seto, K.Y. Lin, G.C. Migliaccio, An evaluation of wearable sensors and their placements for analyzing construction worker's trunk posture in laboratory conditions, Appl. Ergon. 65 (2017) 424–436, http://dx.doi.org/10.1016/j.apergo.2017.03.016.
[30] H.L. Guo, Y.T. Yu, M. Skitmore, Visualization technology-based construction safety management: a review, Autom. Constr. 73 (2017) 135–144, http://dx.doi.org/10.1016/j.autcon.2016.10.004.
[31] K. Khoshelham, Accuracy analysis of Kinect depth data, ISPRS-international archives of the photogrammetry, Remote Sens. Spatial Inf. Sci. 38 (2011) 133–138, http://dx.doi.org/10.5194/isprsarchives-xxxviii-5-w12-133-2011.
[32] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, A. Blake, Real-time human pose recognition in parts from single depth images, IEEE Conf. Comput. Vis. Pattern Recognit. (2011) 1297–1304, http://dx.doi.org/10.1109/cvpr.2011.5995316.
[33] H. SangUk, L. SangHyun, P.M. Feniosky, Comparative study of motion features for similarity-based modeling and classification of unsafe actions in construction, J. Comput. Civ. Eng. 28 (5) (2014) A4014005, , http://dx.doi.org/10.1061/(asce)cp.1943-5487.0000339.
[34] H. SangUk, M. Achar, S. Lee, F. Peña-Mora, Empirical assessment of a RGB-D sensor on motion capture and action recognition for construction worker monitoring, Vis. Eng. 1 (1) (2013) 1–13, http://dx.doi.org/10.1186/2213-7459-1-6.
[35] J. Yang, M.-W. Park, P.A. Vela, M. Golparvar-Fard, Construction performance monitoring via still images, time-lapse photos, and video streams: now, tomorrow, and the future, Adv. Eng. Inform. 29 (2) (2015) 211–224, http://dx.doi.org/10.1016/j.aei.2015.01.011.
[36] I.P.T. Weerasinghe, J.Y. Ruwanpura, J.E. Boyd, A.F. Habib, Application of Microsoft Kinect sensor for tracking construction workers, Construction Research Congress 2012: Construction Challenges in A Flat World, 2012, pp. 858–867, , http://dx.doi.org/10.1061/9780784412329.087.
[37] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.P. Seidel, W.P. Xu, C. Dan, C. Theobalt, VNect: real-time 3D human pose estimation with a single RGB camera, ACM Trans. Graph. 36 (4) (2017) (No. 40), https://doi.org/10.1145/3072959.3073596.
[38] A. Saad, M. Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 288–303, http://dx.doi.org/10.1109/tpami.2008.284.
[39] H. Jiang, M.S. Drew, Z.N. Li, Successive convex matching for action detection, IEEE Conf. Comput. Vis. Pattern Recognit. 2 (2006) 1646–1653, http://dx.doi.org/10.1109/cvpr.2006.297.
[40] L. Jingen, S. Ali, M. Shah, Recognizing human actions using multiple features, IEEE Conf. Comput. Vis. Pattern Recognit. (2008), http://dx.doi.org/10.1109/cvpr.2008.4587527.
[41] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, IEEE Conf. Comput. Vis. Pattern Recognit. (2008) 1–8, http://dx.doi.org/10.1109/cvpr.2008.4587756.
[42] P. Ronald, A survey on vision-based human action recognition, Image Vis. Comput. 28 (6) (2010) 976–990, http://dx.doi.org/10.1016/j.imavis.2009.11.014.
[43] S.E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016), http://dx.doi.org/10.1109/cvpr.2016.511.
[44] B.J. Brian, S. Richard, E.M. Riseman Weiss, View variation of point-set and line-segment features, IEEE Trans. Pattern Anal. Mach. Intell. 15 (1) (1993) 51–68, http://dx.doi.org/10.1109/34.184774.
[45] Y. Kong, Z. Ding, J. Li, Y. Fu, Deeply learned view-invariant features for cross-view action recognition, IEEE Trans. Image Process. 26 (6) (2017) 3028–3037, http://dx.doi.org/10.1109/tip.2017.2696786.
[46] S. Yuping, H. Foroosh, View-invariant action recognition from point triplets, IEEE Trans. Pattern Anal. Mach. Intell. 31 (10) (2009) 1898–1905, http://dx.doi.org/10.1109/tpami.2009.41.
[47] T.H. Lee, C.S. Han, Analysis of working postures at a construction site using the OWAS method, Int. J. Occup. Saf. Ergon. 19 (2) (2013) 245–250, http://dx.doi.org/10.1080/10803548.2013.11076983.
[48] D. Tome, C. Russell, L. Agapito, Lifting from the deep convolutional 3D pose estimation, IEEE Conf. Comput. Vis. Pattern Recognit. (2017) 2500–2509, http://dx.doi.org/10.1109/cvpr.2017.603.
[49] P. Nikolaos, C. Russell, L. Agapito, Learning a manifold as an Atlas, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2013) 1642–1649, http://dx.doi.org/10.1109/cvpr.2013.215.
[50] S. Marta, V. Ntouskos, F. Pirri, Bayesian image based 3D pose estimation, European Conference on Computer Vision, 2016, pp. 566–582, , http://dx.doi.org/10.1007/978-3-319-46484-8_34.
[51] B. Tekin, P. Márquez-Neila, M. Salzmann, P. Fua, Learning to fuse 2D and 3D image cues for monocular body pose estimation, IEEE Int. Conf. Comput. Vis. (2017), http://dx.doi.org/10.1109/iccv.2017.425.
[52] L. Thi-Lan, M.Q. Nguyen, T.T.M. Nguyen, Human posture recognition using human skeleton provided by Kinect, Computing, Management and Telecommunications, International Conference on IEEE, 2013, http://dx.doi.org/10.1109/commantel.2013.6482417.