for automated construction workers' motion recognition using a single in-pocket smartphone

Guohao Wang[1], Waleed Umer[2], Yantao Yu[3], and Heng Li[1]

[1]Department of Building and Real Estate, Faculty of Construction and Environment, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China.

[2]Department of Mechanical and Construction Engineering, Northumbria University, Newcastle, The United Kindom.

[3]Department of Civil and Environmental Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China. (Corresponding author) Email: ceyantao@ust.hk

ABSTRACT

The construction industry around the globe is afflicted with poor productivity and safety records. Understanding the time and frequency of unproductive construction activities and awkward postures contributes to the assessment of individual workers' productivity and risk of injury. Wearable kinematic sensing (WKS) techniques have been extensively investigated to determine worker movements during construction tasks due to their robustness in the presence of obstacles, weather, and on-site noise. However, most existing WKS-based motion recognition methods require multiple devices to be attached to construction workers' bodies in a fixed manner, which is physically invasive making workers unwilling to accept them. This study proposes an automated construction motion recognition method using a single in-pocket smartphone to provide a non-invasive approach to understanding worker's situation in construction workplaces. A smartphone instead of multiple devices was placed loosely in a worker's pocket, with its built-in sensors capturing motion data.

A residual attention-convolutional Long short-term memory network (RA-CLN) was presented by embedding the residual and temporal attention modules on top of the classical CLN to extract **"robust to device's offsets**" features. Finally, the proposed method was verified by a field test on construction sites. The results show that the proposed method can effectively identify five construction workers' basic motions, with an average macro F1-score of up to 91.1%. The superiority of RA-CLN was proved by comparing it with baseline models and existing models. This research will equip the construction industry with a low-cost and non-invasive motion recognition technique for construction workers, contributing to individual productivity and safety management.

INTRODUCTION

Low productivity and high risks of injury to workers are two key issues that the global construction industry is committed to improving (Arditi and Mochtar 2000; Entzel et al. 2007). According to the UK's national statistics, productivity growth has been slow in construction industries compared with the economy over the past two decades (Office for National Statistics 2021). Specifically for labor productivity, the construction industry lags since the annual increase in global labor productivity rate for construction was only one-third of that in manufacturing in the past twenty years (F. Barbosa 2017). In addition to low productivity, the persistence and prevalence of unsafe construction practices and health-related issues among frontline construction workers have attracted widespread attention. According to statistics, construction workers suffer twice more work-related injury risks when compared with other industries (Yu et al. 2019). Work-related musculoskeletal disorders (WMSDs) are one of the most prevalent types of nonfatal injuries in the construction sector (CPWR 2018). Almost 46% of construction workers self-reported they had one or more WMSDs-related symptoms (Dong et al. 2019). WMSDs have brought financial burdens to workers' families, and society with loss of income and productivity increased medical expenses, and compensation. For example, construction workers' median days away from work due to WMSDs increased from eight days in 1992 to 13 days in 2017 (Dong et al. 2019). This injury-induced prolonged absence can exacerbate labor shortages and affect the execution of pre-designed construction schedules, thus delaying production. In addition, WMSDs, such as overexertion, accounted for

21% of workers' compensation claim costs in construction in Ohio and Washington states in 2015 (CPWR 2018). WMSDs-related symptoms in construction workers often result from their intensive, repetitive, and awkward working postures. In order to prevent such injury, it is necessary to monitor workers' real-time working postures so as to grasp the frequency and duration of workers falling into awkward postures and thus infer the risk of injury (Luo et al. 2019; Nath et al. 2018).

Knowledge of construction workers' motion information not only facilitates expected production rates but also helps to reduce the risk of worker injury (Sherafat et al. 2020). Traditionally, such information has been measured manually using stopwatches and visual observation or by studying pre-recorded video footage of the operation (Rashid and Louis 2020). However, these manual approaches are time-consuming, laborious, and error-prone (Golparvar-Fard et al. 2013; Ryu et al. 2019), which generates a growing demand for alternative automated solutions. An extensive body of work related to automatically recognizing the motions of construction workers has been proposed. From the perspective of sensing technology, these works could be mainly divided into two categories: physical contactless sensing (PCS)-based methods (Cho et al. 2017; Gong et al. 2011; Han and Lee 2013; Khosrowpour et al. 2014; Memarzadeh et al. 2013; Teizer and Vela 2009; Yu et al. 2019; Zhu et al. 2017; Zhang et al. 2018; Rashid and Louis 2020), and wearable kinematic sensing-based methods (Bangaru et al. 2021; Kim and Cho 2021; Kim and Yong 2020; Ogunseiju et al. 2021; Yang et al. 2020; Zhao and Obonyo 2021; Sanhudo et al. 2021). One of the most significant superiorities of WKS-based methods over PCS-based methods is the ability to operate continuously in poor visual conditions and noisy construction environments, which are common characteristics in the construction site. However, most of the existing WKS-based motion recognition methods for construction workers have issues of low willingness to use brought by the fixed wearing style and high deployment cost. These issues largely hamper the practical application of the WKS-based monitoring system. In addition, the validation of most of the WKS-based methods relies on datasets from a single type of worker executing pre-designed workflows in a laboratory environment and still needs further testing on multi-trade datasets from real construction scenarios.

In terms of motion classifiers, deep learning (DL) models have replaced shallow machine learning algorithms as the dominant model for human activity recognition due to their end-to-end training strategy and powerful feature representation capabilities (Chen et al. 2021). Among various DL models, the convolutional long and short-term memory network (CLN) has become a representative model for the human activity recognition (HAR) problem because it effectively combines the advantages of convolutional neural network (CNN) and long short-term memory networks (LSTM) in capturing local spatial dependencies and temporal dependencies. However, the classical CLN model needs to be further enhanced to better identify the target motions from raw signal that contain noise data from device offset caused by intense construction activities.

To solve the above issues, a non-invasive and low-cost WKS-based automated motion recognition method for construction workers was proposed in this study. Firstly, a smartphone embedded with an accelerometer, gyroscope, and magnetic sensor was placed freely in the construction worker's pocket to collect motion data. Secondly, the Synthetic Minority Oversampling Technique (SMOTE) data augmentation method was used to solve the problem of poor recognition resulting from severe data imbalance. Subsequently, incorporating the residual module and the temporal attention module, a novel motion classification model, i.e., RA-CLN, was designed to improve the learning ability of the classical CLN model. The stratified K-fold cross-validation methods and Average macro F1-score were employed to evaluate the model accurately. Finally, the effectiveness of the proposed method was validated by a self-collected dataset consisting of 3 rebar workers, 2 carpenters, and 2 masonry workers from a real building construction site.

The structure of the paper is as follows. Section 2 reviewed previous studies regarding motion capture techniques and motion classifiers. In Section 3, the proposed method in this study was presented in detail. Section 4 demonstrated the details of the experiment. Results were analyzed and discussed in section 5. Finally, Section 6 and 7, respectively, presents the conclusion and limitations of this research and future works.

LITERATURE REVIEW

The choice of motion capture techniques and classifiers is crucial in determining the effective-

ness of HAR methods. This section systematically reviews commonly used construction workers' motion capture techniques and classifiers and summarizes current research gaps.

Motion capture techniques for construction workers

The latest efforts in monitoring workers' motions are more about leveraging advanced sensing technology to capture construction workers' motion data than conventional manual observation, which is time-consuming, error-prone, and costly. Two types of motion sensing techniques were reviewed: physical contactless sensing and wearable kinematic sensing techniques.

Physical contactless sensing techniques

Physical contactless sensing refers to the technology of acquiring images, videos, or audio by remote sensing through air mediums such as light and sound waves. Recently, the advancement of deep learning algorithms and the explosion of computing power contributed to the full exploitation of image and sound information in life, spawning two hot research fields (computer vision (CV) and audio recognition). Achieving success in these sectors affords researchers new management tools for construction personnel.

Diverse CV-based approaches have been proposed to automatically detect construction workers' postures, activities, and status. Khosrowpour et al. (Khosrowpour et al. 2014) proposed a vision-based automatic workforce assessment method using RGB-D sensors. In addition to flat information. RGB sensors might gather depth information that leads to workers' skeleton feature extraction. However, this approach is inappropriate for outdoor applications since RGB-D sensors are more susceptible to direct sunlight. Instead of RGB-D sensors, Han and Lee (Han and Lee 2013) adopt two view cameras to obtain depth information about workers. Yu et al. (Yu et al. 2021) enhanced 3D posture estimation by reducing two cameras to a monocular camera. The 3D pose of construction workers was utilized to conduct the joint-level ergonomic assessment (Yu et al. 2019). Other than pose-related information, Luo et al. (Luo et al. 2018) utilized individual cameras and presented an improved convolutional neural network to capture activity-related information from construction workers automatically. Despite the use of varied camera setups, all of the investigations above employed close-up cameras. The disadvantage of close-up cameras is the narrow field

of view which makes these methods have only a small number of objects to observe (Khosrowpour et al. 2014; Yu et al. 2021) and makes it easy for workers to disappear from the field of view (Han and Lee 2013; Khosrowpour et al. 2014) or occlude (Luo et al. 2018; Yu et al. 2019).

Other academics chose on-site surveillance cameras, which often offer a larger field of vision. Luo et al. (Luo et al. 2019) proposed a hierarchical statistical method for recognizing workers' activities in far-field surveillance videos. This study introduced a new fusion approach for temporal segment networks to offset the detrimental effects of short clip size and extreme motion camera blur. They further introduced a discriminative model by combining the spatial relevance of a worker group and deep activity features (Luo et al. 2020). These works show site surveillance video containing richer semantic information than close-up cameras, such as relationships within a working group and the environment. However, at the same time, they also claimed that information extraction from surveillance videos requires a relatively high-profile computer which may raise operating expenses (Luo et al. 2020). This issue was also claimed by Zhu et al. (Zhu et al. 2017). Instead of online processing, they performed the detection and tracking offline and individually for each video frame. Another typical issue faced by surveillance video-based methods is the recognition failure under poor environment illumination, such as night, cloudy and rainy, and occlusion conditions (Sherafat et al. 2020; Zhu et al. 2017).

Audio recognition is yet another growing physical contactless sensing approach in the construction industry. Since each worker or equipment activity generates distinct sound, its identification provides imperative information regarding work processes and safety-relevant concerns (Zhang et al. 2018). The advantages of the audio-based recognition method are that it can gather 360-degree data, is less susceptible to physical obstruction and poor visual conditions, and occupies relatively minimal computational and storage resources (Sherafat et al. 2020). However, most audio-based studies (Sabillon et al. 2020; Sherafat et al. 2019) aim to identify construction machine operations other than those of construction workers. This is due to the fact that construction workers generally emit low-volume, insignificant sounds, which are challenging to record and further classify. Khandaker M. Rashid and Joseph Louis (Rashid and Louis 2020) proposed an automated

audio-based activity identification framework based on which five activities (hammering, nailing, sawing, drilling, and idle) can be recognized with 96.6% F-1 score. In spite of this, this method can only be applied to isolated and relatively low-noise sites (Lee et al. 2020), such as modular construction factories (Rashid and Louis 2020). Construction sites often have multiple processes operating simultaneously, meaning that the sounds generated by different activities can overlap. De-noising multi-source fused sound data to get the required sound domain remains a significant technological obstacle. In addition, activities that do not generate sound do not apply to this method.

Wearable kinematic sensing techniques

This technique relies on wearable kinematic sensors, such as accelerometers, gyroscopes, inertial measurement units (IMUs), and sensor-embedded devices like smartphones to capture different motion patterns of construction workers (Sherafat et al. 2020). In recent years, WKS-based human activity recognition methods have progressed because of ease of use, reliable and continuous data, and high recognition accuracy (Rashid and Louis 2020). Common application Scenarios include fall detection (Dai et al. 2010), fitness tracking, home and work automation (Tapia et al. 2004), etc. There already exist several open-source datasets corresponding to daily activities, such as WISDM (Kwapisz et al. 2011), UNIMIB SHAR (Micucci et al. 2017), and PAMAP2 (Reiss and Stricker 2012). However, these datasets are unavailable to obtain a recognition model for construction activities as they are quite different from and more complex than daily activities (Ryu et al. 2019).

Table 1 compares kinematic sensors' type and number, wearing position, and style used in previous works that used WKS techniques to check construction workers' status, activities, and motions. Most of these studies reported that their methods achieved over 80% recognition accuracy or F1-score in respective test environments demonstrating their effectiveness. Nevertheless, they still face some difficulties from the application in the actual construction management.

Firstly, most studies use one (Gondo and Miura 2020; Hosseinian et al. 2019) or even multiple accelerometers (Joshua and Varghese 2014) or IMUs (Kim and Cho 2021; Sanhudo et al. 2021; Valero et al. 2017; Zhao and Obonyo 2020) to collect workers' motion data. However, deploying,

operating, and maintaining (O&M) these devices in large-scale projects can be costly and time-consuming. In addition, the widely used accelerometers and IMUs usually do not have information processing units such as central processing units (CPU) or graphic processing units (GPU), which means that recorded data need to be uploaded to a central server for postprocessing. This may bring some delay in the activity's detection. The prevalence of smartphones offers us a new solution to overcome this limitation. Almost everyone is now equipped with a smartphone-embedded accelerometer, gyroscope, magnetic and Global Position System (GPS), etc. Using the personal smartphone on construction sites has huge potential to reduce the cost of sensor deployment and O&M. Additionally, computing chips (Cheng and Wang 2011; Francisco and Daniel 2016) integrated into smartphones enable them performing edge computing which largely contributes to real-time monitoring of worker activity (Francisco and Daniel 2016). Another benefit of edge computing is that workers do not need to upload source data but only the results, thus providing good privacy protection.

Another concern is the physical invasiveness of the WKS-based methods, as this largely determines the worker's willingness to wear them. As shown in Table 1, most of the studies used a fixed strap-on sensor-wearing style (Akhavian and Behzadan 2016; Gondo and Miura 2020; Hosseinian et al. 2019; Joshua and Varghese 2014; Nath et al. 2018; Sanhudo et al. 2021; Valero et al. 2017; Yang et al. 2019; Zhang et al. 2019; Zhao and Obonyo 2020; Zhao and Obonyo 2021). This wearing style is able to capture the activity of body parts more accurately; however, it is difficult to be accepted in actual project management. Firstly, fixed wear is highly physically invasive, which may cause discomfort and even anger, especially when workers are engaged in prolonged, high-intensity jobs (**?**). On the other hand, assuming that the device will not change its position and orientation under continuous high-intensity construction activities is not appropriate. Normally devices will experience a cheap position, such as sliding from the upper arm to the lower arm which may cause misclassification of the recognition model (Shi et al. 2020). Therefore, a more effective model should have robustness that can counteract the effects of smartphone position shifts. In addition, models need to be trained and tested on a dataset of worker activities from actual projects, as the

problem of equipment deflection is difficult to expose in a laboratory simulation environment of construction activities used in most current studies (Akhavian and Behzadan 2016; Hosseinian et al. 2019; Lim et al. 2016; Nath et al. 2018; Sanhudo et al. 2021; Valero et al. 2017; Yang et al. 2019; Zhao and Obonyo 2020).

Classifiers for WKS-based motion recognition

WKS-based human activity recognition is a classical classification problem for multi-variate time series (Ronao and Cho 2016). Most current studies developed classifiers by training machine learning (ML) models in a supervised learning manner. Commonly used shallow ML classifiers involve support vector machine (SVM) (Akhavian and Behzadan 2016; Antos et al. 2014; Chen et al. 2017; Ronao and Cho 2016), K-Nearest Neighbours (KNN) (Ruan et al. 2016; Wannenburg and Malekian 2017), Decision Trees (DT) (Ryu et al. 2019; Sztyler et al. 2017), Random Forest (RF) (Zhao and Obonyo 2020) and artificial neural network (ANN) (Bangaru et al. 2021; Zhang et al. 2019). Despite promising results, these shallow ML classifiers require heuristic feature engineering based on expert domain knowledge, which is time-consuming (Slaton et al. 2020). This process may also result in subjective bias impairing models' classification performance, as excessive features add noise, affecting models' performance, while insufficient features make it more difficult to differentiate motions (Hall 1999). Another issue of this pipeline is that they separate feature engineering, feature selection, and model parameters training, making it difficult for shallow ML classifiers to extract knowledge from the original data, resulting in sub-optimal models.

In contrast, deep learning models process information end-to-end, i.e., receiving the initial data as input and outputting the classification results directly. This is beneficial for feature extraction because models trained in this way know how to adapt themselves (i.e., updating parameters) according to their performance (i.e., classification results). DL's feature representation capabilities are further enhanced through the deep stacking of non-linear layers. Two representative models are CNN and LSTM. CNNs, by the ability to capture the local relevance of information (Ramanujam et al. 2021), are become the dominant technique in the computer vision field and have recently been

introduced to WKS-based human activity recognition. The superiority of CNN over shallow ML models in the automated extraction of distinctive motion features has been proved in (Ronao and Cho 2016; Shi et al. 2020). Another characteristic of CNN revealed in (Shi et al. 2020) was that CNN could effectively reduce the negative effect on recognition brought by dynamically changing device placement and orientation. For LSTM, the significant difference from a fully connected network is that LSTM could model temporal dependencies from long-scale data sequences by connecting memory cells in different timesteps. This characteristic originally made it the dominant model for time series problems such as natural language processing and HAR (Ashry et al. 2020; Kim and Cho 2021; Wang and Liu 2020; Yang et al. 2020).

The combined CNNs and LSTMs in a unified framework has ever offered state-of-the-art (SOTA) results in the speech recognition domain. In WKS-based motion recognition field, Ordonez and Roggen (Ordóñez and Roggen 2016) initially proposed a deep convolutional and LSTM neural network for multimodal wearable activity recognition. Test results on several open-source datasets show that CLN outperforms CNN or LSTM independent models. Subsequently, hybrid CLN models were applied to recognize daily living activities (Li et al. 2017) and construction workers' postures (Zhao and Obonyo 2020). Although the CLN yielded state-of-the-art performance in the above scenarios, it struggled to work well when recognizing construction workers' activities based on a single in-pocket smartphone. Compared with daily activities such as running and cycling, construction activities are more intense, more irregular, and have more frequent motion transitions. Unlike (Zhao and Obonyo 2020), which uses a fixed-worn sensor, when a smartphone is freely placed in a pocket, it will produce a body-independent extra offset. The unforeseen and dynamic offset may impair the classifier's performance as it acts as noise to the target activity recognition.

To alleviate the adverse effects of the offset, several approaches have been proposed. Guo et al. used the rotation matrix to transform data simulating different orientations of smartphones (Guo et al. 2016). Yurtman and Bashan (Yurtman and Barshan 2017) removed the impact caused by changes in the sensor orientation by proposing heuristic orientation-independent algorithms. A rule-based framework was presented by Theekakul et al. (Theekakul et al. 2011). to realize orientation-

independent activity recognition. They processed the offset from the perspective of enriching or correcting the input, ignoring the device's dynamic and continuous deflection. Considering this problem, efforts were taken to extract features that are robust to the devices' offset by constructing advanced recognition models. For example, Shi et al. proposed a convolutional neural network-based model to extract features that are invariant with device placement and orientation (Shi et al. 2020).

To overcome the above limitations, this study aims to take full advantage of the CLN model in spatial and temporal dependency capture and enhance its ability to extract "robust to device's offsets" features to recognize workers' motions from in-pocket smartphone signals efficiently. This is achieved by proposing a new motion classifier (i.e., RA-CLN) that improves the classical CLN model by introducing residual and temporal attention modules. Specifically, the classical CLN model is deepened to enhance its feature representation capabilities. The residual connection replaces the sequence connection to solve the problem of gradient disappearance during the training process of the deep learning model. The temporal attention module is embedded after the LSTM layer to help the model understand which timesteps in a sample are most representative of the target activity.

METHODOLOGY

This section presents the motion recognition method proposed in this study, which includes motion data collection and preprocessing, data augmentation method, RA-CLN model, and model performance evaluation method.

Motion data collection and preprocessing

Three smartphone-embedded sensors, i.e., accelerometer, gyroscope, and magnetic sensor, are used to capture motion information of construction workers. A free, easy-to-use, cross-platform motion data logging software named Sensor Logger (Hei 2021) is used to record, save, and transmit data from three sensors. Sensor Logger supports multiple sensors, sampling frequencies, and data storage formats. The sampling rate is set as 50hz. In addition, the whole process of data acquisition is recorded synchronously with a video recorder for data annotation.

The acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. Thus, the collected data is transformed from 9 into 12 dimensions. To train and evaluate the classifier, continuous data needs to be labeled for the ground-truth activity category and then segmented into samples according to data label and window size. Data labeling is done manually according to the synchronous video. Each data point will be assigned a specific label. The data segmentation is achieved by using the sliding window technique, which divides the sensor signals into smaller window-size segments. The window size is a crucial parameter for motion recognition which refers to the number of data points in one window. Previous studies have achieved promising results by choosing window sizes between 0.5 and 2s (Akhavian and Behzadan 2016; Bayat et al. 2014). Some dynamic motions often occur in a concise time. If the selected window size is larger than the time of occurrence, it will not be possible to use this part of the data, which may make the sample further unbalanced. Therefore, three window sizes, i.e., 0.6, 0.8, and 1s, corresponding to 30, 40, and 50 data points per sample, were selected in this study. Also, overlapping adjacent windows reduces the error caused by transition state noise (Su et al. 2014). Similar to previous studies (Antwi-Afari et al. 2020; Nath et al. 2018), a 50% overlap of the adjacent windows was adopted for this study.

Data augmentation

Human activity is often unbalanced (Chen et al. 2021), and so are the construction activities of workers, which means the time they engage in different activities varies considerably. For example, workers usually spend much more time standing and walking than in other postures, such as bending and squatting. The model trained on such an unbalanced dataset is more likely to predict the motion sample as the dominant class and less effective in identifying the minority class. However, identifying the minority classes, such as bending and squatting, is more important than the majority class. To overcome the imbalanced dataset problem for construction workers' motion classification, a classical oversampling method, namely the Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al. 2002), was applied to increase the minority class examples by creating new synthetic samples. SMOTE is based on a k-nearest neighbors algorithm and linear

interpolation. The test datasets are not involved in this procedure, so testing is based only on actual instances rather than synthetic ones.

The proposed RA-CLN network model

CLN model was first applied to the HAR problem by Ordonez and Roggen (Francisco and Daniel 2016) and achieved almost SOTA performance on several open-source datasets. This section proposes a novel CLN network model by embedding residual module and temporal attention layers on top of the classical CLN model to enhance the CLN's performance further. Firstly, residual blocks are built on top of CNN, and blocks are stacked by skip connections to increase the depth of the model and thus improve the capability of CNN feature representation. Skip-connections proposed in (He et al. 2016) have been proven that can effectively solve the problem of gradient disappearance in the process of deep model optimization. Secondly, considering that not all timesteps contribute equally to the recognition of the target motion category, temporal attention is proposed based on a self-attention mechanism to calculate the weighted average sum of all hidden states of the last LSTM layer. Figure 1 depicts the architecture of the proposed RA-CLN model. As shown in Figure 1, the input in the form of $S * T$ (sensor modalities * time steps) is first fed into the residual module for spatial feature extraction. Two LSTM layers are used to capture the temporal dependencies, and dropout was adopted after the layer to prevent overfitting of the model. Then each time step's output is assigned a weight and integrated by a matrix multiplying outputs and weights in the temporal attention module. Finally, the dense layer reasoned about the probability of the input samples belonging to each class, and the one with the highest probability was the recognized motion class. The software code for the model implementation is available at https://github.com/wangguohao-github/RA-CLN-for-smartphone-based-motion-recognition.git.

Residual module

Residual modules are designed to extract spatial representations of each motion class from raw sensor signals. As shown in Figure 1, each residual module consists of two 1D convolutional (Conv1D) layers, two batch normalization (BN) layers, and a relu activation layer. Suppose the input data shape is $S*T$, where S denotes the number of channels and T denotes the number of time

steps. The residual module firstly processed the input data by sweeping 1D convolutional kernels of shape $3 * S$ along the T axis. The same padding is applied, and the stride is set to 1 to maintain the shape of the T dimension. Thus, the output of the 1D convolutional layer x_t^f is given by

$$x_t^f = b^f + \sum_{i=1}^{3} \sum_{s=1}^{S} w_{i,s}^f \times x_{t-2+i,s} \quad (1)$$

where b^f and $w_{i,s}^f$ are bias and weight and bias of filter f, and $x_{t-2+i,s}$ is the input data at $(t-2+i)th$ time step of sth channel. The output shape thus is $F * T$. Through the manually parameter optimization, we found 256 filters generated better result, thus in this study, filter num is set to 256, i.e., $F = 256$.

After the convolutional layer, a batch normalization layer is added to speed up the training process and overcome the covariate shift issues. The relu activation function is then used to add the non-linear element. As for the 2^{nd}-4^{th} residual block, skip connection (He et al. 2016) is employed to add the previous block's output to the current block's output and use it as input for the next module.

Temporal attention module

In the classical CLN model (Francisco and Daniel 2016), only the final timestep's hidden state from the last LSTM layer is output to the dense layer for classification. The basic idea of temporal attention is that instead of learning a single vector representation in the last timestep, we have differentiated the information representations using the LSTM output at each time step.

The structure of the temporal attention module is shown in Figure 2. The temporal attention mechanism is designed to calculate H by weighting sum of LSTM in each step's output h_i, as follows:

$$H = \sum_{i=1}^{T} \alpha_i \times h_i \quad (2)$$

where the α_i are the attention weights. In the classical LSTM (Hochreiter and Schmidhuber 1997), the α^T is fixed to be 1, and $\alpha_i = 0$ when $i < T$. Here, the weight vector $\alpha_i = [\alpha_1, \alpha_2, ..., \alpha_T]$ are

calculated with a two full-connected feed-forward neural network:

$$D = tanh(W_1 \times h + b_1) \quad (3)$$

$$\alpha = softmax(W_2 \times D + b_2) \quad (4)$$

where $h = [h_1, h_2, ..., h_T]$ is the output vector of LSTM, D is the output of the first dense layer, W_1, W_2, b_1, and b_2 denote 2D matrices and biases, respectively, in the two linear layers, which consist of all trainable parameters of the temporal attention module. Softmax and tanh refer to two activation functions of two layers. In this way, the model is able to review the previous information and focus on more important parts to learn a better representation.

Performance evaluation

Stratified K-fold cross-validation for the model's evaluation

General evaluation of deep learning models can be done by splitting the collected experiment data into train, validation, and test set. The model's hyperparameters trained by the train set can be optimized based on its performance on the validation set. The final model's evaluation should be made on the test set. To avoid the subjectivity of data splitting, cross-validation techniques are commonly used to segment the dataset. Stratified K-fold cross-validation techniques were adopted in this study. Compared with the traditional K-fold, the Stratified K-fold could ensure that the proportion of each category's samples in the training, validation, and test sets is the same as the original dataset. This is a choice made by considering the imbalance in motion data.

The procedure of the Stratified K-fold cross-validation method is shown in Figure 3. Firstly, the whole dataset was divided into five subsets consisting of the same proportion's samples of each motion category. Then, four of them were selected as the training set and the remaining one as the test set. Subsequently, 80% of stratified samples from the train set were used to train the model, and the remaining 20% of samples acted as a validation set based on which the hyperparameters of the trained model were optimized. The optimized model was then evaluated on a test set. The above process was repeated five times, guaranteeing all subsets go through the test stage. The average

performance of models on five test sets was the final performance.

Performance metrics

Classification accuracy is the most used metric for classification tasks which is calculated as the number of correctly predicted outcomes to the total number of predictions. However, accuracy is insufficient to decide the robustness and reliability of classification results, especially for the task with an imbalanced dataset. In this study, the average $macro\ F1-score$ was selected as the performance metric. F1-score is the harmonic mean of precision and recall as follows:

$$F1\_score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

where precision and recall are also performance metrics given by $Precision = TP/(TP + FP)$, $Recall = TP/(TP + FN)$. TP is the number of correct positive predictions done by a positive model. FP is the number of classes predicted incorrectly where the model thinks predicted classes are positive (true), but it is not true. FN is the only misclassified metric where the model thinks the predicted activity is not positive (true), but it is true. Then, the $macroF1-score$ can be calculated as:

$$macro\ F1-score = \frac{(F1-score)_i \times n_i}{\sum_{i=1}^{k} (F1-score)_i \times n_i} \quad (6)$$

where n_i is the number of samples of the ith motion class, $(F1-score)_i$ is the ith motion class's $F1-score$ calculated by equation 5. Finally, the average $macro\ F1-score$ was used to evaluate the model, which is the mean of the model's $macro\ F1-score$ under five cross-validations as described in 3.

EXPERIMENTS

Data collection

Motion data were collected on a real building construction site, which involved three types of construction workers, namely masonry, carpenter, and rebar worker, performing their respective activities. Two smartphones were put respectively in their chest and trouser pockets to record

motion information. Workers were free to work according to their tasks without guidance and constraints throughout the process. A video camera covering the job area was used to record the process synchronously. After the recording was completed, data from smartphones' built-in motion sensors were exported in a CSV file format. The details information about workers and collected data are shown in Table 2.

Dataset preparation

According to the recorded video, each data point was manually labeled with its corresponding motion category. The proportion of each motion class is shown in Table 2. Among classes, ST stands for straight standing or standing with minor shaking; WK stands for walking; BD stands for bending; SQ stands for squatting; TR stands for Transitional movements between other motions.

The sliding window technique was then used to divide the samples into consecutive data with the same labels according to windows of 0.6s, 0.8s, and 1s. A total of six datasets from two pockets with three-time windows were then divided into training, validation, and test sets by the stratified 5-fold cross-validation method as described in 3.

It can be seen from Table 2 that the collected data is very imbalanced. Precisely, workers engaged in ST or SQ motion more than 50% of the time, reaching a maximum of about 90% for R1. In order to obtain more reliable results, imbalanced data in the train sets should be augmented before constructing the model, i.e., expanding the minority class data. Smote techniques introduced in 3 was adopted here. Table 3 shows the sample distribution of the original (Ori) and augmented (Aug) training sets. ‘’— represents that data is not augmented.

Models building

Different models were proposed here to be compared with the proposed model, which helps to verify the effectiveness of the proposed method. In this study, the proposed RA-CLN adds the residual and temporal attention modules to the classical CLN. Therefore, CLN, residual CLN and CLN with attention were used as baseline models. Deep learning models proposed in previous studies on construction workers' motion recognition are also introduced for comparison. A brief description of each model is shown in Table 4. The model development environment is TensorFlow

2.9.0 (GPU version), CUDA-v11.3, and cudnn-v8.2.1. All models were implemented on a Windows 11 PC (Intel Core i7-11800H CPU@ 2.3 GHz, 16 GB RAM, NIVIDA GeForce GTX 3070 Laptop GPU@8 GB RAM).

The individual hyperparameters are optimized according to the performance of each model on the validation set. For a fair comparison, we selected an identical set of hyperparameters for training each model, i.e., a maximum training round of 300, a batch size of 256, and an initial learning rate of 0.00005. The callback function is used to dynamically evaluate whether the model training is under fitted or overfitted according to the training loss and validation loss. The early stopping strategy is employed to prevent model overfitting. The model will stop early at 20 epochs after the loss of validation set is not in decline. The best models were retained for subsequent model testing. In addition, Adam was chosen as the optimizer, and the loss function was cross-entropy.

RESULTS AND DISCUSSION

In this section, the performance of the proposed method was systematically evaluated. Firstly, the influence of the data augmentation method was analyzed. Then, model performance evaluation on the augmentation dataset was conducted. Finally, several baseline models and existing methods were compared to the proposed method to verify its effectiveness.

Evaluation of data augmentation method

This section analyzes the influence of the data augmentation method by comparing the performance of the model on the original dataset and the augmented dataset, as shown in Table 5. Details about two datasets have been introduced in section 4. Results show that the smote-based data augmentation method improved the model's performance on the unbalanced data set, as the average macro F1-score of the augmented model improved by more than 1% and reached 3.8%. In the evaluation and comparison later in this paper, the augmented dataset is used to train the model.

Figure 5 further depicts the proposed model's performance on each motion class of the original and augmentation dataset for each smartphone position and window size combination. From the results, the recognition effect of the model for the majority class remains basically the same before and after data augmentation and still exceeds the 95% F1 score. BD motion class recognition

improved by about 2-4% with data augmentation. The most significant boost occurred in the WK and TR categories, where the sample was the smallest, at about 4% and up to 10% (Trouser pocket with 0.8s window size). This indicates that adding minority class samples using SMOTE method does not affect the recognition performance of the model on the majority class. However, it can significantly improve the recognition performance of the minority class.

Performance evaluation of the RA-CLN model

This section analyzes the performance of the model trained on data with three-time windows (0.6s, 0.8s, and 1s) from two smartphone positions, i.e., chest pocket and trouser pocket. The performance of the proposed model under different combinations is shown in Figure 6. As can be seen from Figure 6, the proposed method achieves promising results for each combination of the time window and smartphone position. Specifically, all the average macro-F1 scores exceed 88%, with a maximum of 91.1% when the time window is 0.8s, and the smartphone is placed in the trouser pocket. Comparing the model's performance in different time windows, the worker motion data can be classified in a time window of 0.8s with better classification results than 0.6s and 1s. In terms of smartphone positions, we found that the trouser pocket's model achieves better than the chest pocket in each time window.

To further explore the impact of smartphone position in the chest pocket and trouser pocket on construction workers' motion recognition, the performance of both models under the 0.8s window size was compared and analyzed in the form of confusion matrices, as shown in Figure 7. Overall, both pockets' models effectively identify motions for each category. Similar results can be found from both pocket models, i.e., the recognition recall rate for both ST and SQ actions can exceed 95% with optimal results, followed by recognition of BD with around 90% recall rate, and TR and WK.

In addition, it can be seen from the normalized confusion matrix that both pockets' models misclassify about 10% of the WK and TR samples as ST, which directly reduces the recognition of WK and TR activities. One of the main reasons for such results is the imbalanced samples in each category. As shown in non-normalized confusion matrices which list how many samples in

test sets were correctly or incorrectly classified for each motion class, the number of ST and SQ motion samples is significantly more than BD, WK and TR. Although data augmentation methods have been employed to balance datasets and achieved some improvement, the model is still able to better capture the patterns of the majority of samples' motion classes, such as ST and SQ, thus yielding higher classification results.

Comparing the results of the two pockets' models, it can be found that the Chest pocket model outperforms the trouser pocket in recognition of WK and BD motions and vice versa for the recognition of TR and SQ motions. This result is understandable because the location of the smartphone determines the motion characteristics of the body part it captures. Trouser pocket smartphones capture lower limb motion like SQ and TR, while chest pocket ones are more responsive to upper limb motion like BD.

Comparison evaluation

The performance of various models was compared with the proposed RA-CLN model in this section. Model architectures have been introduced in 4. Comparison results are given in Table 6. Firstly, ablation studies were conducted by comparing three baseline models, R-CLN, A-CLN, and CLN models, to analyze the influence of the residual and temporal attention modules. As shown in Table 6, A-CLN slightly outperforms the basic CLN indicating that in addition to the output of the last step of the LSTM model, the output of the previous time step also contains valuable information for motion classification. The Temporal attention module effectively helps the model pay attention to the outputs of these informative time steps and integrates the outputs of each time step in a weighted manner to improve the model performance.

R-CLN is also improved on the basis of CLN. Compared with the classic CLN model, the R-CLN model stacks four residual blocks composed of two convolutional layers, hoping to enhance the representation ability of the model by deepening the model. However, models that are too deep may have problems with vanishing training gradients. The result of this study verifies the idea that skipping connections between layers can effectively alleviate this problem. Compared with the three baseline models, RA-CLN has the best performance, which shows that the method

successfully integrates the advantages of the residual and temporal attention modules.

Two existing deep learning models, i.e., CLN (Zhao and Obonyo 2020) and LSTM (Kim and Cho 2021), are also implemented for comparison. Results show that with almost the same parameters, the CLN model (Zhao and Obonyo 2020) slightly outperforms the LSTM model (Kim and Cho 2021), demonstrating convolutional neural networks' superiority in spatial feature extraction. Although the parameters of the RA-CLN model are three times higher than those of the CLN and LSTM models, the RA-CLN improves performance by 6% over the other two models. This improvement indicates that increasing the model depth by stacking CNN layers can significantly enhance the ability of model feature representation. However, this finding is not consistent with the study (Zhao and Obonyo 2020) , which claimed that the increase in the number of CNN layers did not significantly improve the model performance. One possible reason is that unlike the squared 2D convolution used in (Zhao and Obonyo 2020) , the rectangular 1D convolution was used in this study. The rectangular kernel makes it have a larger receptive field in the channel axis direction (Sensor modalities) and therefore global dependencies that can be captured among all sensor modalities. For this reason, 1D convolution has been widely used for time series type of classification and prediction problems.

CONCLUSION

This study proposed a less invasive and low-cost method to recognize construction workers' motions automatically. The overall methodology consists of data collection using a single smartphone that is placed freely in workers' pockets, SMOTE-based data augmentation, and a novel RA-CLN classifier that improves the classical CLN by introducing residual module and temporal attention module. An actual dataset containing a total of 7 workers and 2 pockets in three trades (2 carpenters, 3 rebar workers, and 2 masons) was used to validate the proposed method. Results indicate that the proposed method can effectively identify five types of motion (standing, walking, transitional movement, bending, and squatting) of construction workers with an average macro F1 score of up to 91.1%. The data augmentation method can significantly improve the recognition of minority motion classes without compromising the recognition of majority classes. Comparing

different time windows and pocket combinations reveals that the model has optimal performance with the trouser pocket and 0.8s window combination. Comparison experiment with baseline models, including R-CLN, A-CLN, and CLN, and existing models, including CLN (Zhao and Obonyo 2020) and LSTM (Kim and Cho 2021), demonstrates that deepening the model, i.e., increasing the number of CNN layers can enhance the model's ability to characterize features. Residual connection helps to alleviate the problem of gradient disappearance in the training process due to model deepening and thus can improve the model. The temporal attention module helps the model focus on the time step, which contains more important information about the target motions. The proposed method uses only a smartphone that is freely placed in the worker's pocket to recognize the worker's motions, which has the advantages of non-invasiveness and no additional cost, so it has great potential to be applied to the measurement of worker productivity in actual engineering construction or ergonomics risk assessment.

LIMITATIONS AND FUTURE WORKS

There are still some limitations in this study that need to be addressed in future studies. First, as discussed in sections 5, although the SMOTE data augmentation method used in this study can somewhat improve the model's recognition performance on minority class, about 10% of the samples are still misclassified as majority class. Addressing this issue will significantly improve the model's performance. Some possible solutions include designing weighted loss functions to force the model to optimize the classification of a few categories by increasing the loss of minority classes or using an ensemble learning strategy that integrates the advantages of different models in the recognition of different motion classes.

The proposed method effectively identifies five motions of construction workers; However, this is not enough. In the future, workers' motions need to be further classified. For example, transitional motions should be further divided into squatting up, squatting down, etc. A more detailed classification would help to manage more effectively. However, a more detailed division would create a more severe sample imbalance problem.

Additionally, although the RA-CLN model proposed in this study outperforms the CLN and

LSTM models, the parameters of the RA-CLN model are three times larger than those of the other two models. This may hinder it from running in real-time on the smartphone terminal. So, the model needs to be deployed in smartphones for testing in the future. Also, the device's energy consumption while running the recognition model needs to be measured.

DATA AVAILABILITY STATEMENT

Some or all data, models, or code that support the findings of this study are available from the corresponding author upon reasonable request.

ACKNOWLEDGMENTS

REFERENCES

Akhavian, R. and Behzadan, A. H. (2016). "Smartphone-based construction workers' activity recognition and classification." *Automation in Construction*, 71, 198–209.

Antos, S. A., Albert, M. V., and Kording, K. P. (2014). "Hand, belt, pocket or bag: Practical activity tracking with mobile phones." *Journal of Neuroscience Methods*, 231, 22–30.

Antwi-Afari, M. F., Li, H., Umer, W., Yu, Y. T., and Xing, X. J. (2020). "Construction Activity Recognition and Ergonomic Risk Assessment Using a Wearable Insole Pressure System." *Journal of Construction Engineering and Management*, 146(7).

Arditi, D. and Mochtar, K. (2000). "Trends in productivity improvement in the US construction industry." *Construction Management & Economics*, 18(1), 15–27.

Ashry, S., Ogawa, T., and Gomaa, W. (2020). "CHARM-Deep: Continuous Human Activity Recognition Model Based on Deep Neural Network using IMU Sensors of Smartwatch." *IEEE Sensors Journal*, 20(15), 8757.

Bangaru, S. S., Wang, C., Busam, S. A., and Aghazadeh, F. (2021). "ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors." *Automation in Construction*, 126(1), 103653.

Bayat, A., Pomplun, M., and Tran, D. A. (2014). "A Study on Human Activity Recognition Using Accelerometer Data from Smartphones." *Procedia Computer Science*, 34, 450–457.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." *Journal of Artificial Intelligence Research*, 16(1), 321–357.

Chen, J., Qiu, J., and Ahn, C. (2017). "Construction worker's awkward posture recognition through supervised motion tensor decomposition." *Automation in Construction*, 77, 67–81.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z., and Liu, Y. (2021). "Deep Learning for Sensor-based Human Activity Recognition: Overview, Challenges and Opportunities." 54(4), 1–40.

Cheng, K. T. and Wang, Y. C. (2011). "Using mobile GPU for general-purpose computing – a case study of face recognition on smartphones." 1–4.

Cho, C., Lee, Y. C., and Zhang, T. Y. (2017). "Sound Recognition Techniques for Multi-Layered Construction Activities and Events." *Computing in Civil Engineering 2017: Smart Safety, Sustainability, and Resilience*, 326–334.

CPWR (2018). *The construction chart book: The US construction industry and its workers*. CPWR-The Center for Construction Research and Training, <https://www.cpwr.com/research/data-center/the-construction-chart-book/>.

Dai, J., Bai, X., Yang, Z., Shen, Z., and Dong, X. (2010). "PerFallD: A pervasive fall detection system using mobile phones." 292–297.

Dong, X. S., Betit, E., Dale, A. M., Barlet, G., and Wei, Q. (2019). "Trends of musculoskeletal disorders and interventions in the construction industry." *Report no.*, <https://stacks.cdc.gov/view/cdc/86273> (September). Place: Silver Spring, MD.

Entzel, P., Albers, J., and Welch, L. (2007). "Best practices for preventing musculoskeletal disorders in masonry: Stakeholder perspectives." *Applied Ergonomics*, 38(5), 557–566.

F. Barbosa, M. Parsons, J. M. (2017). "Improving Construction Productivity." *McKinsey*.

Francisco, O. and Daniel, R. (2016). "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition." *Sensors*, 16(1), 115.

Golparvar-Fard, M., Heydarian, A., and Niebles, J. C. (2013). "Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers." *Advanced Engineering Informatics*, 27(4), 652–663.

637 Gondo, T. and Miura, R. (2020). "Accelerometer-Based Activity Recognition of Workers at Con-
638 struction Sites." *Frontiers in Built Environment*, 6, 563353.
639 Gong, J., Caldas, C. H., and Gordon, C. (2011). "Learning and classifying actions of construction
640 workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models."
641 *Advanced engineering informatics*, 25(4), 771–782.
642 Guo, H., Chen, L., Chen, G., and Lv, M. (2016). "Smartphone-based activity recognition indepen-
643 dent of device orientation and placement." *International Journal of Communication Systems*,
644 29(16), 2403–2415.
645 Hall, M. A. (1999). "Correlation-based feature selection for machine learning." Ph.D. thesis, Ph.D.
646 thesis. Publication Title: The University of Waikato.
647 Han, S. U. and Lee, S. H. (2013). "A vision-based motion capture and recognition framework for
648 behavior-based safety management." *Automation in Construction*, 35, 131–141.
649 He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition."
650 770–778.
651 Hei, C. T. (2021). *Sensor Logeer Version 1.7*, <https://www.tszheichoi.com/sensorlogger>.
652 Hochreiter, S. and Schmidhuber, J. (1997). "Long Short-Term Memory." *Neural Computation*,
653 9(8), 1735–1780.
654 Hosseinian, S. M., Zhu, Y., Mehta, R. K., Erraguntla, M., and Lawley, M. A. (2019). "Static and
655 Dynamic Work Activity Classification from a Single Accelerometer: Implications for Ergonomic
656 Assessment of Manual Handling Tasks." *Iise Transactions on Occupational Ergonomics &*
657 *Human Factors*, 1–20.
658 Joshua, L. and Varghese, K. (2014). "Automated recognition of construction labour activity us-
659 ing accelerometers in field situations." *International Journal of Productivity and Performance*
660 *Management*, 63(7), 841–862(22).
661 Khosrowpour, A., Niebles, J. C., and Golparvar-Fard, M. (2014). "Vision-based workface assess-
662 ment using depth images for activity analysis of interior construction operations." *Automation in*
663 *Construction*, 48, 74–87.

664 Kim, K. and Cho, Y. K. (2021). "Automatic Recognition of Workers' Motions in Highway Construc-
665 tion by Using Motion Sensors and Long Short-Term Memory Networks." *Journal of Construction*
666 *Engineering and Management*, 147(3).
667 Kim, K. and Yong, K. C. (2020). "Effective inertial sensor quantity and locations on a body for
668 deep learning-based worker's motion recognition." *Automation in Construction*, 113, 103126.
669 Kwapisz, J. R., Weiss, G. M., and Moore, S. A. (2011). "Activity recognition using cell phone
670 accelerometers." *Acm Sigkdd Explorations Newsletter*, 12(2), 74–82.
671 Lee, Y. C., Scarpiniti, M., and Uncini, A. (2020). "Advanced Sound Classifiers and Performance
672 Analyses for Accurate Audio-Based Construction Project Monitoring." *Journal of Computing in*
673 *Civil Engineering*, 34(5), 04020030.
674 Li, X., Zhang, Y., Zhang, J., Chen, S., Marsic, I., Farneth, R. A., and Burd, R. S. (2017). "Concurrent
675 Activity Recognition with Multimodal CNN-LSTM Structure." arXiv:1702.01638.
676 Lim, T. K., Park, S. M., Lee, H. C., and Lee, D. E. (2016). "Artificial Neural Network–Based
677 Slip-Trip Classifier Using Smart Sensor for Construction Workplace." *Journal of Construction*
678 *Engineering & Management*, 142(2), 04015065.
679 Luo, H. B., Xiong, C. H., Fang, W. L., Love, P. E. D., Zhang, B. W., and Ouyang, X. (2018). "Convo-
680 lutional neural networks: Computer vision-based workforce activity assessment in construction."
681 *Automation in Construction*, 94, 282–289.
682 Luo, X., Li, H., Yu, Y., Zhou, C., and Cao, D. (2020). "Combining deep features and activity context
683 to improve recognition of activities of workers in groups." *Computerided Civil and Infrastructure*
684 *Engineering*, 35(9).
685 Luo, X. C., Li, H., Yang, X. C., Yu, Y. T., and Cao, D. P. (2019). "Capturing and Understanding
686 Workers' Activities in Far-Field Surveillance Videos with Deep Action Recognition and Bayesian
687 Nonparametric Learning." *Computer-Aided Civil and Infrastructure Engineering*, 34(4), 333–
688 351.
689 Memarzadeh, M., Golparvar-Fard, M., and Niebles, J. C. (2013). "Automated 2D detection of con-
690 struction equipment and workers from site video streams using histograms of oriented gradients

and colors." *Automation in Construction*, 32, 24–37.

Micucci, D., Mobilio, M., and Napoletano, P. (2017). "UniMiB SHAR: A Dataset for Human Activity Recognition Using Acceleration Data from Smartphones." *Applied Sciences*, 7(10), 1101.

Nath, N. D., Chaspari, T., and Behzadan, A. H. (2018). "Automated ergonomic risk monitoring using body-mounted sensors and machine learning." *Advanced Engineering Informatics*, 38, 514–526.

Office for National Statistics (2021). *Productivity in the construction industry, UK: 2021*.

Ogunseiju, O. R., Olayiwola, J., Akanmu, A. A., and Nnaji, C. (2021). "Recognition of workers' actions from time-series signal images using deep convolutional neural network." *Smart and Sustainable Built Environment*.

Ordóñez, F. J. and Roggen, D. (2016). "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition." *Sensors*, 16(1), 115 Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Ramanujam, E., Perumal, T., and Padmavathi, S. (2021). "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review." *IEEE Sensors Journal*, 21(12), 13029–13040.

Rashid, K. M. and Louis, J. (2020). "Activity identifi cation in modular construction using audio signals and machine learning." *Automation in Construction*, 119, 103361.

Reiss, A. and Stricker, D. (2012). "Introducing a New Benchmarked Dataset for Activity Monitoring." 108–109.

Ronao, C. A. and Cho, S. B. (2016). "Human activity recognition with smartphone sensors using deep learning neural networks." *Expert Systems with Applications*, 59, 235–244.

Ruan, W., Chea, L., Sheng, Q. Z., and Yao, L. (2016). "Recognizing daily living activity using embedded sensors in smartphones:A data-driven approach." Springer, 250–265.

Ryu, J., Seo, J., Jebelli, H., and Lee, S. (2019). "Automated Action Recognition Using an Accelerometer-Embedded Wristband-Type Activity Tracker." *Journal of Construction Engi-*

718 *neering and Management*, 145(1), 04018114.
719 Sabillon, C., Rashidi, A., Samanta, B., Davenport, M. A., and Anderson, D. V. (2020). "Audio-
720 Based Bayesian Model for Productivity Estimation of Cyclic Construction Activities." *Journal*
721 *of Computing in Civil Engineering*, 34(1), 04019048.
722 Sanhudo, L., Calvetti, D., Martins, J. P., Ramos, N. M. M., Meda, P., Goncalves, M. C., and
723 Sousa, H. (2021). "Activity classification using accelerometers and machine learning for complex
724 construction worker activities." *Journal of Building Engineering*, 35, 102001.
725 Sherafat, B., Ahn, C. R., Akhavian, R., Behzadan, A. H., Golparvar-Fard, M., Kim, H., Lee,
726 Y. C., Rashidi, A., and Azar, E. R. (2020). "Automated Methods for Activity Recognition
727 of Construction Workers and Equipment: State-of-the-Art Review." *Journal of Construction*
728 *Engineering and Management*, 146(6).
729 Sherafat, B., Rashidi, A., Lee, Y. C., and Ahn, C. R. (2019). "Automated Activity Recognition
730 of Construction Equipment Using a Data Fusion Approach." *Computing in Civil Engineering*
731 *2019: Data, Sensing, and Analytics*, 1–8.
732 Shi, J. H., Zuo, D. C., Zhang, Z., and Luo, D. Y. (2020). "Sensor-based activity recognition inde-
733 pendent of device placement and orientation." *Transactions on Emerging Telecommunications*
734 *Technologies*, 31(4).
735 Slaton, T., Hernandez, C., and Akhavian, R. (2020). "Construction activity recognition with
736 convolutional recurrent networks." *Automation in Construction*, 113, 103138.
737 Su, X., Tong, H., Ji, P., Department, C. S., Center, G., York, t. C. U. o. N., and College, C.
738 (2014). "Activity Recognition with Smartphone Sensors." *Tsinghua Science and Technology*,
739 19(3), 235–249.
740 Sztyler, T., Stuckenschmidt, H., and Petrich, W. (2017). "Position-aware activity recognition with
741 wearable devices." *Pervasive & Mobile Computing*, 38, 281–295.
742 Tapia, E. M., Intille, S. S., and Larson, K. (2004). "Activity Recognition in the Home Using Simple
743 and Ubiquitous Sensors." 158–175.
744 Teizer, J. and Vela, P. A. (2009). "Personnel tracking on construction sites using video cameras."

Advanced Engineering Informatics, 23(4), 452–462.

Theekakul, P., Thiemjarus, S., Nantajeewarawat, E., Supnithi, T., and Hirota, K. (2011). "A rule-based approach to activity recognition." *Knowledge, Information, and Creativity Support Systems*, Springer, 204–215.

Valero, E., Sivanathan, A., Bosche, F., and Abdel-Wahab, M. (2017). "Analysis of construction trade worker body motions using a wearable and wireless motion sensor network." *Automation in Construction*, 83, 48–55.

Wang, L. K. and Liu, R. Y. (2020). "Human Activity Recognition Based on Wearable Sensor Using Hierarchical Deep LSTM Networks." *Circuits Systems and Signal Processing*, 39(2), 837–856.

Wannenburg, J. and Malekian, R. (2017). "Physical activity recognition from smartphone accelerometer data for user context awareness sensing." *IEEE Transactions on Systems Man & Cybernetics Systems*, 47(12), 3142–3149.

Yang, K., Ahn, C. R., and Kim, H. (2020). "Deep learning-based classification of work-related physical load levels in construction." *Advanced Engineering Informatics*, 45, 101104.

Yang, Z., Yuan, Y. B., Zhang, M. Y., Zhao, X. F., and Tian, B. Q. (2019). "Assessment of Construction Workers' Labor Intensity Based on Wearable Smartphone System." *Journal of Construction Engineering and Management*, 145(7).

Yu, Y., Yang, X., Li, H., Luo, X., Guo, H., and Fang, Q. (2019). "Joint-Level Vision-Based Ergonomic Assessment Tool for Construction Workers." *Journal of Construction Engineering and Management*, 145(5), 04019025.

Yu, Y. T., Li, H., Cao, J. N., and Luo, X. C. (2021). "Three-Dimensional Working Pose Estimation in Industrial Scenarios With Monocular Camera." *Ieee Internet of Things Journal*, 8(3), 1740–1748.

Yurtman, A. and Barshan, B. (2017). "Activity recognition invariant to sensor orientation with wearable motion sensors." *Sensors*, 17(8), 1838.

Zhang, M. Y., Cao, T. Z., and Zhao, X. F. (2019). "Using Smartphones to Detect and Identify Construction Workers' Near-Miss Falls Based on ANN." *Journal of Construction Engineering and Management*, 145(1), 04018120.

772 Zhang, T. Y., Lee, Y. C., Scarpiniti, M., and Uncini, A. (2018). "A Supervised Machine Learning-
773 Based Sound Identification for Construction Activity Monitoring and Performance Evaluation."
774 358–366, <://WOS:000541116600035>.
775 Zhao, J. and Obonyo, E. (2018). "Towards a Data-Driven Approach to Injury Prevention in Con-
776 struction." Springer, 385–411.
777 Zhao, J. and Obonyo, E. (2021). "Applying Incremental Deep Neural Networks-based Posture
778 Recognition Model for Injury Risk Assessment in Construction." *Advanced Engineering Infor-*
779 *matics*, 50, 101374.
780 Zhao, J. Q. and Obonyo, E. (2020). "Convolutional long short-term memory model for recog-
781 nizing construction workers' postures from wearable inertial measurement units." *Advanced*
782 *Engineering Informatics*, 46, 101177.
783 Zhu, Z., Ren, X., and Chen, Z. (2017). "Integrated detection and tracking of workforce and
784 equipment from construction jobsite videos." *Automation in Construction*, 81, 161–171.

List of Tables

TABLE 1. Summary of previous works on WKS-based activities recognition for construction workers.

References	Type and number of sensors	Position of sensor placement	Wearing style	Data acquisition
(Sanhudo et al. 2021)	3 IMUs	Waists and the dominant leg	Fixed by straps	10 volunteers in the laboratory environment
(Joshua and Varghese 2014)	3 ACCs	Right lower arm left lower arm, and waist for ironworkers and head and waist for the carpenters	Fixed by bands	10 ironworkers and 10 carpenters in the real construction site
(Hosseinian et al. 2019)	1 ACC	Chest	Fixed by a belt	27 volunteers performing designed activities in a laboratory setting
(Gondo and Miura 2020)	1 ACC	Waist	Fixed by belt	2 carpenters on the construction site of two real houses
(Valero et al. 2017)	8 IMUs	Arms, lower back, upper back, upper legs, and lower legs	Fixed by elastic straps	6 subjects in the laboratory experiment
(Zhao and Obonyo 2018)	5 IMUs	Thigh, calf, ankle, upper arm, and head	Fixed by straps	1 student performing predesigned activities in the laboratory experiment
(Zhao and Obonyo 2020)	5 IMUs	Forehead, chest, right upper arm, right thigh, and right crus	Stuck on workers' clothing or helmet	1 masonry, 1 demolition, and 2 electricians from a residential building construction project
(Kim and Cho 2021)	2 IMUs	The back of the neck and lower back	Placed in pockets of the safety vest	1, 2, and 4 workers in three different job sites, respectively
(Akhavian and Behzadan 2016)	1 Smartphone	The upper arm of the dominant hand	Secured by an armband	2 subjects under a test environment
(Yang et al. 2019)	2 Smartphones	Thigh and wrist	Fixed by elastic armbands	25 graduate students in the laboratory experiments
(Lim et al. 2016)	1 Smartphone	Left hip pocket	Placed in pocket	3 subjects performed a pre-designed task cycle
(Zhang et al. 2019)	1 Smartphone	Sacrum	Fixed by a band	5 subjects in the simulating experiment
(Nath et al. 2018)	2 Smartphones	The upper arm and waist	Mounted by armbands	2 workers performed the assigned task cycle in the laboratory

TABLE 2. Description of data collection.

Trade	NO.	Age (yr)	Weight (cm)	Height (kg)	Duration (min)	The proportion of each motion class				
						ST	WK	TR	BD	SQ
Rebar worker	R1	35	175	70	30.8	70.7%	6.5%	2.3%	1.6%	18.9%
	R2	42	176	72	30.2	75.8%	7.6%	3.3%	3.3%	9.9%
	R3	51	173	76	31.7	70.7%	12.9%	2.2%	3.3%	10.8%
Masonry	M1	48	170	70	51.4	38.0%	1.3%	8.1%	13.2%	39.3%
	M2	50	168	80	50.1	65.2%	1.2%	11.4%	17.3%	4.9%
Carpenter	C1	37	172	75	30	36.7%	4.9%	5.2%	18.6%	34.7%
	C2	40	175	80	25.2	39.4%	16.4%	5.6%	16.9%	21.7%

TABLE 3. Sample distribution of the original and augmented training sets.

Motion	Chest pocket						Trouser pocket					
	0.6s		0.8s		1s		0.6s		0.8s		1s	
	Ori	Aug	Ori	Aug	Ori	Aug	Ori	Aug	Ori	Aug	Ori	Aug
ST	20303	—	15112	—	11984	—	20285	—	15107	—	11979	—
WK	2173	7000	1571	5000	1219	4000	2195	7000	1585	5000	1231	4000
TR	2157	7000	1381	5000	900	4000	2160	7000	1382	5000	906	4000
BD	3847	7000	2829	5000	2218	4000	3818	7000	2810	5000	2197	4000
SQ	7235	—	5401	—	4303	—	7225	—	5394	—	4295	—

TABLE 4. Brief description of each model.

Module	Baseline model			The proposed model	Existing model	
	CLN	R-CLN	A-CLN		CLN (Zhao and Obonyo 2020)	LSTM (Kim and Cho 2021)
1	(Cv1D-256)x2	(Cv1D-256)x2	(Cv1D-256)x2	(Cv1D-256)x2	Cv2D-64	Dense
2	(Cv1D-256)x2	*(Cv1D-256)x2	(Cv1D-256)x2	*(Cv1D-256)x2	Flatten	
3	(Cv1D-256)x2	*(Cv1D-256)x2	(Cv1D-256)x2	*(Cv1D-256)x2		
4	(Cv1D-256)x2	*(Cv1D-256)x2	(Cv1D-256)x2	*(Cv1D-256)x2		
5	(LSTM-64)x2	(LSTM-64)x2	(LSTM-64)x2	(LSTM-64)x2	(LSTM-128)x2	(LSTM-180)x2
6	Dense	Dense	Tem_att-128	Tem_att-128	Dense	Dense
7			Dense	Dense		

TABLE 5. Performance of the RA-CLN on the original and augmentation dataset.

Performance	Dataset	Chest pocket			Trouser pocket		
		0.6	0.8	1	0.6	0.8	1
Average Macro F1-score	Original dataset	87.8%	88.1%	86.2%	88.3%	87.3%	88.5%
	Augmentation dataset	89.2%	89.2%	88.6%	89.7%	91.1%	89.7%
	Improvement	1.4%	1.1%	2.4%	1.4%	3.8%	1.2%

TABLE 6. Comparison results between the proposed method, baseline models and existing models.

Model	Parameters (million)	Chest pocket			Trouser pocket			Average
		0.6	0.8	1	0.6	0.8	1	
The proposed model	1.51	**89.2%**	**89.2%**	88.6%	**89.7%**	**91.1%**	**89.7%**	**89.58%**
R-CLN	1.50	88.7%	88.8%	**88.9%**	**89.7%**	88.2%	89.6%	88.98%
A-CLN	1.51	88.0%	**89.2%**	88.5%	89.3%	90.4%	88.3%	88.95%
CLN	1.50	87.6%	88.1%	88.0%	89.3%	89.0%	88.4%	88.40%
CLN (Zhao and Obonyo 2020)	0.59/ 0.74/0.85	85.4%	83.9%	82.9%	86.3%	83.7%	83.1%	84.22%
LSTM (Kim and Cho 2021)	0.57	82.8%	83.8%	82.5%	85.2%	84.9%	83.2%	83.73%

List of Figures

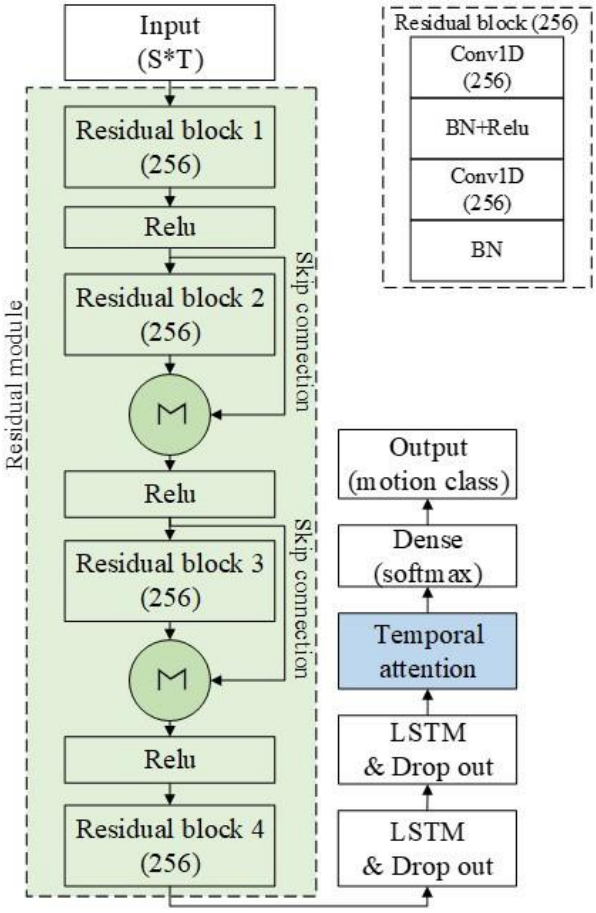


Fig. 1. The architecture of the proposed RA-CLN model.

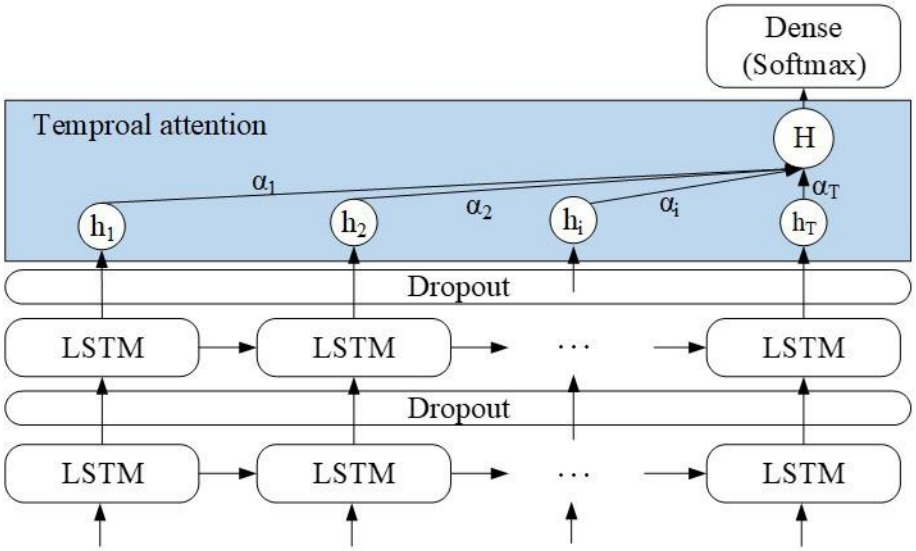


Fig. 2. The structure of the temporal attention module.

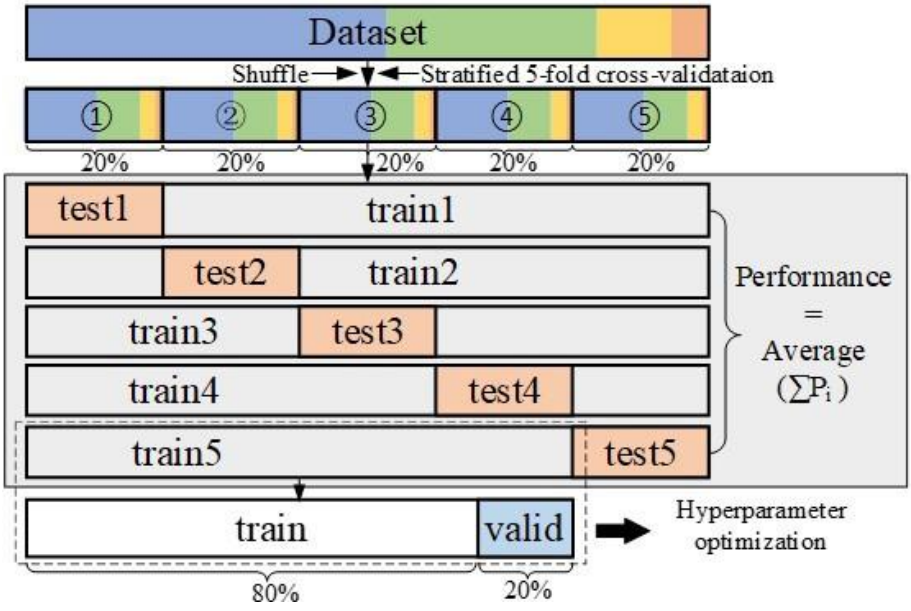


Fig. 3. Stratified 5-fold cross-validation method for model evaluation.

(a) Bending (b) Squatting (c) Standing (d) Walking (e) Transition

Fig. 4. Construction workers' motions involved in the dataset.

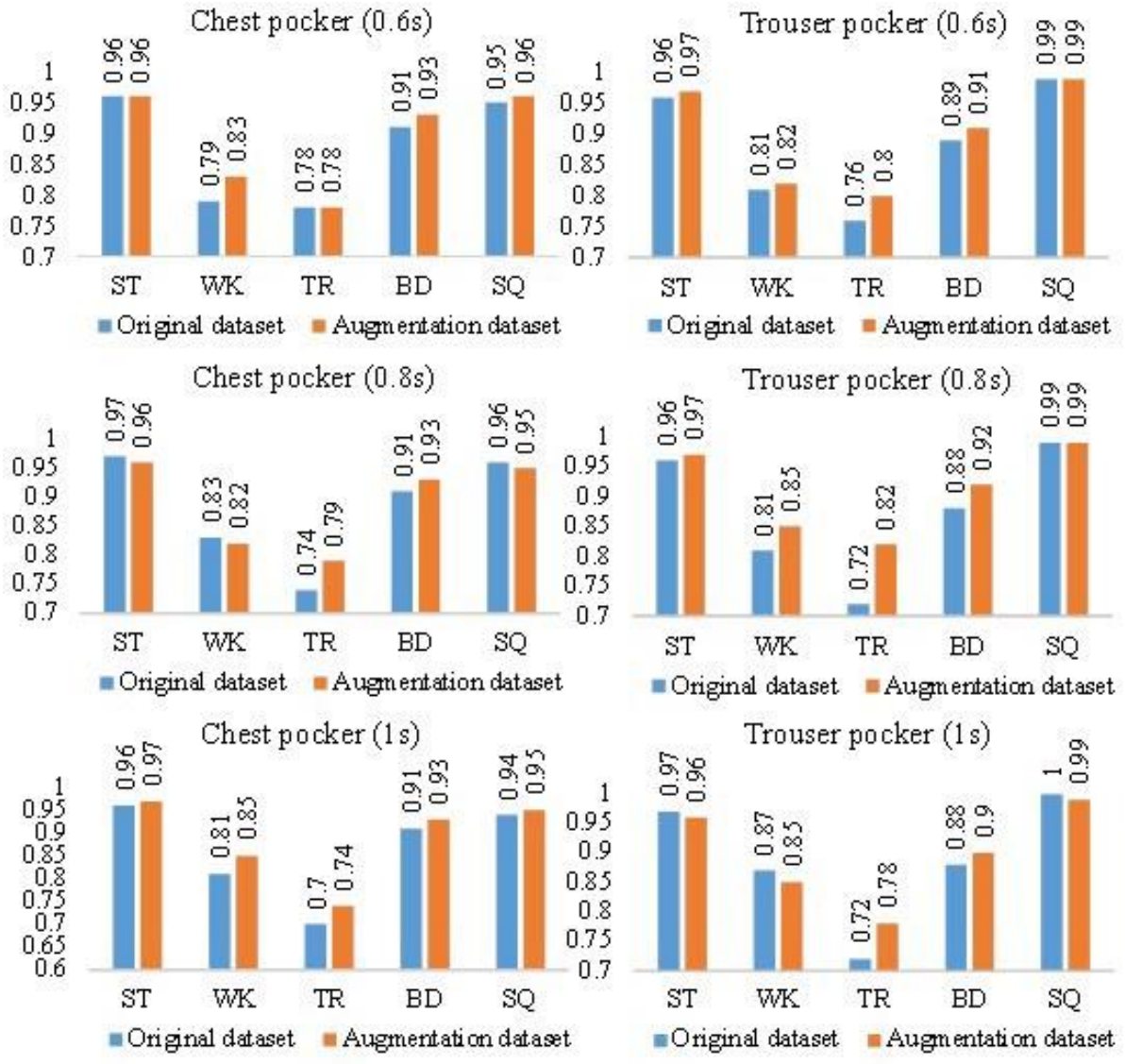


Fig. 5. Performance of the RA-CLN on the original and augmentation dataset.

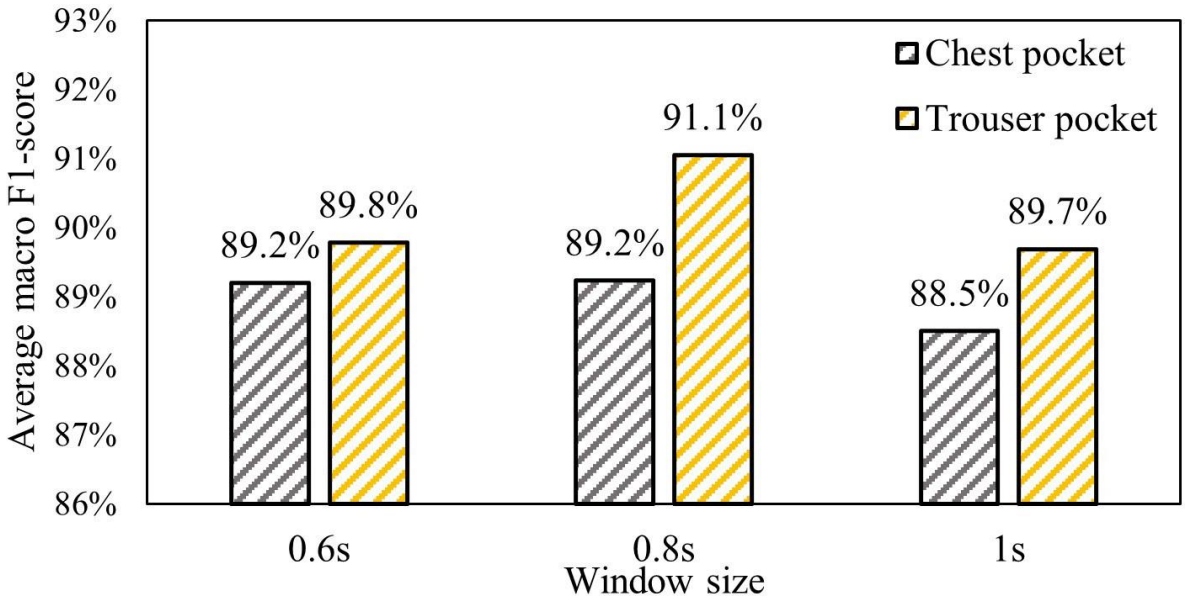
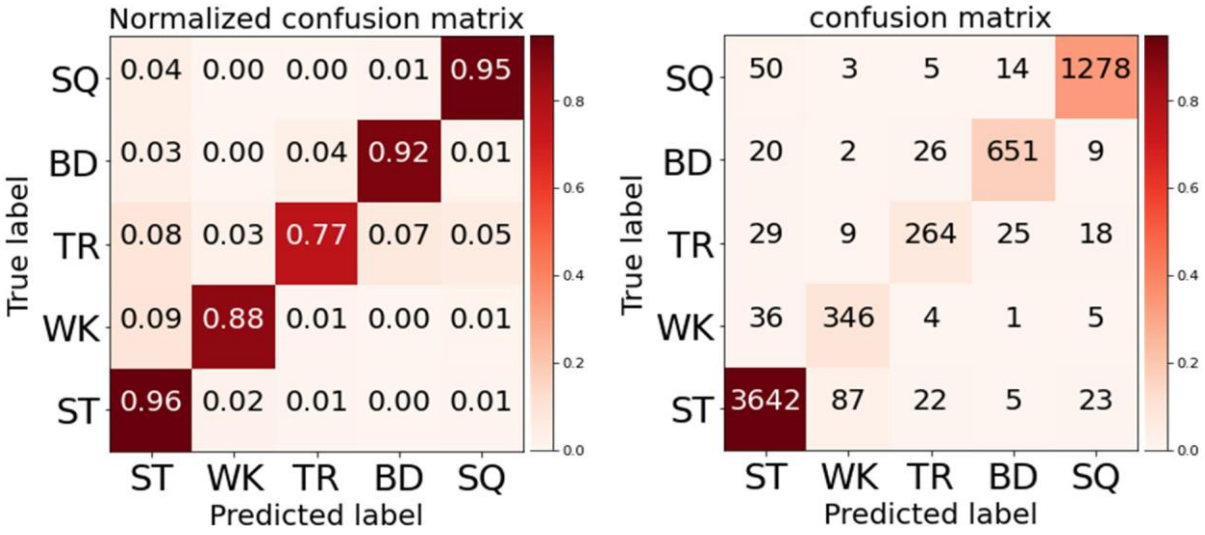
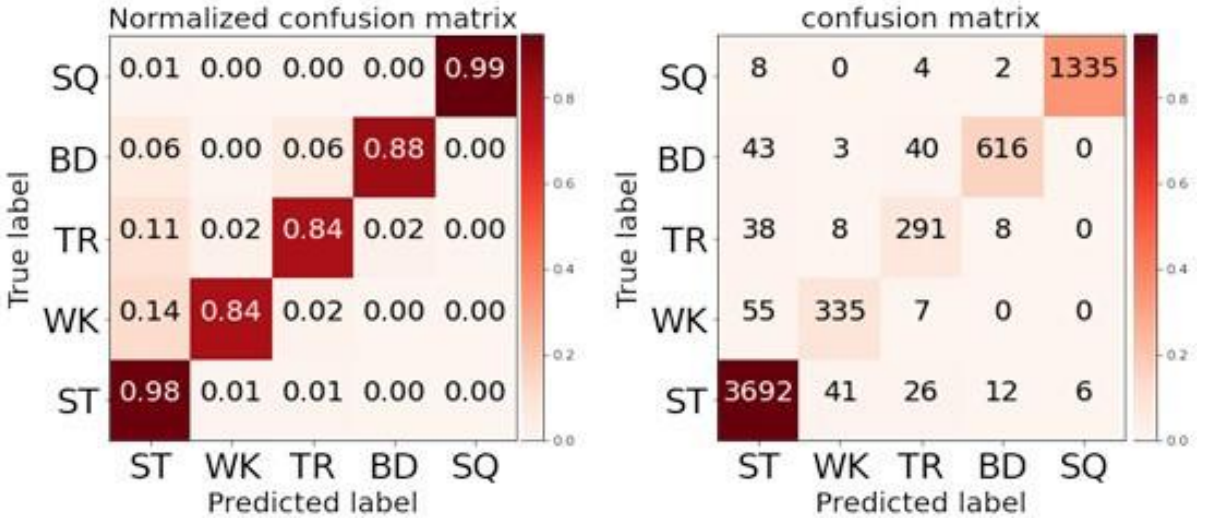


Fig. 6. Performance of the RA-CLN on each pocket and time window combination.



(a) Chest pocket's normalized and non-normalized confusion matrix.



(b) Trouser pocket's normalized and non-normalized confusion matrix.

Fig. 7. Confusion matrixes of two pockets' models with 0.8s window size.