IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 7, 2019, accepted May 21, 2019, date of publication June 10, 2019, date of current version June 21, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922104

Efficient Human Activity Recognition Solving the Confusing Activities Via Deep Ensemble Learning

RAN ZHU[1], ZHUOLING XIAO[1], YING LI[1], MINGKUN YANG[1], YAWEN TAN[2], LIANG ZHOU[1], SHUISHENG LIN[1], AND HONGKAI WEN[3]

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Glasgow College, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]Department of Computer Science, University of Warwick, Coventry CV4 7AL, U.K.

Corresponding author: Zhuoling Xiao (zhuolingxiao@uestc.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61703076, and in part by the Funds for the Central Universities under Grant ZYGX2016J008 and Grant ZYGX2016KYQD125.

ABSTRACT The ubiquity of smartphones and their rich set of on-board sensors has created many exciting new opportunities, where smartphones are used as powerful computing platforms to sense and analyze pervasive data. One important application of mobile sensing is activity recognition based on smartphone inertial sensors, which is a fundamental building block for a variety of scenarios, such as indoor pedestrian tracking, mobile health care, and smart cities. Although many approaches have been proposed to address the human activity recognition problem, several challenges are still present: 1) people's motion modes are very different for different individuals; 2) there is only a very limited amount of training data; 3) human activities can be arbitrary and complex, thus handcrafted feature engineering often fails to work; and 4) the recognition accuracy tends to be limited due to confusing activities. To tackle those challenges, in this paper, we propose a human activity recognition framework based on convolutional neural networks (CNNs) with two convolutional layers using the smartphone-based accelerometer, gyroscope, and magnetometer. To solve the confusion between highly similar activities like going upstairs and walking, this paper presents a novel ensemble model of CNN to further improve the identification accuracy. The extensive experiments have been conducted using 235 977 sensory samples from a total of 100 subjects. The results have shown that the classification accuracy of the proposed model can be up to 96.11%, which proves the effectiveness of the proposed model.

INDEX TERMS Convolutional neural network, human activity recognition, sensor data, smartphone.

I. INTRODUCTION

Human activity recognition (HAR) aiming to identify the actions carried out by a person given a set of observations of subject, has attracted much attention from both academia and industry with widely application requirements appearing in the indoor pedestrian tracking [1], [2], healthcare [3], and smart cities [4]. Currently, HAR methods can be mainly summarized as two categories: vision-based and sensor-based. Vision-based mainly relies on various high-frame-rate video devices [5], [6]. External factors such as lighting condition, clothing color, and image background have a great impact on recognition accuracy. The sensor-based approach, by contrast, is more robust in complex environments, which makes the system convenient and portable. Also, it can identify confusing human activities with the mathematical model by directly measuring the motion from human activities without infringement of personal privacy [7].

With the advent of miniaturized sensors and powerful computing resources in smartphones, the concept of efficient and ubiquitous HAR on smartphones is ready to fulfill soon. Among recent studies focusing on smartphone-based HAR, most researchers chose waist as the position to carry smartphones [8], [9]. However, the requirement for rigid attachment and specified placement is incompatible with the way in which people use mobile devices. For example, over a period of a few minutes, a smartphone could be carried in the backpack and then shifted to a pocket, before being taken

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman.

out and being used to send a text message [10]. This may be one of the main reasons why it is so hard to conduct HAR using smartphone sensors.

Existing studies of sensor-based activity recognition often rely on supervised machine learning approaches such as Hidden Markov Model (HMM) [11], K-Nearest-Neighbors (KNN) [12], eXtreme Gradient Boosting (XGBoost) [13], Random Forest (RF) [14] and Support Vector Machine (SVM) [15], [16] using motion data collected from various types and quantities of motion sensors placed in different parts of body. However, these approaches are limited to three aspects: Firstly, due to the diversity and complexity of human activities, handcrafted feature extraction requires experience and expertise of the field. For the same reason, some extracted features show excellent performance in recognizing some activities, but rather bad at others [17]. Secondly, even for the same activity, the waveforms of motion sensors are quite different in different smartphone placements. This makes it difficult to recognize various different activities with high precision. Thirdly, because of the differences in behavioral habits, gender, and age, the movement patterns of different people vary greatly, which enhances the difficulty of dividing the boundaries of different activities. The recognition accuracy tends to be limited due to confusing activities which generate similar motion signals.

Recent years have witnessed fast development and unparalleled performance in many areas (i.e. image recognition [18], natural language processing [19]) of deep learning. There is a growing trend of discovering meaningful representations of raw data by Convolutional Neural Network (CNN). It has shown great performance in different domains for avoiding handcrafted features. Therefore, we present the ensemble framework based on CNN to recognize human activities. Without tiring data preprocessing and feature extraction and selection, we put raw data that is partitioned by the sliding window into our network. By fully mining the information carried by the signal, it can achieve more accurate recognition on the combination of arbitrary activities and devices placement.

This paper presents a framework and performance analysis of smartphone-sensor based HAR. Sensor data from accelerometer, gyroscope and magnetometer were collected when participants performed some typical and daily human activities: going upstairs, going downstairs, running, walking, standing, bicycling and swinging. We then used the ensemble of CNN to recognize human activities, especially those easily confused. The experiments have demonstrated the improvement on recognition accuracy with the approach proposed in this paper. In summary, the key contributions of this paper are:

- A novel approach based on the ensemble of CNN has been proposed to solve the confusion between highly similar activities such as going upstairs and walking, which outperforms the single CNN model and achieves 96.11% accuracy.
- Based on the collected data, we compare our model with the commonly used classifiers. The fact proves that the approach proposed in this paper outperforms other existing models in feasibility and efficiency.
- A huge amount of motion data including 235 977 data samples from various types of motion sensors and sports scenes with different participants and postures are collected to validate the effectiveness of the proposed method.

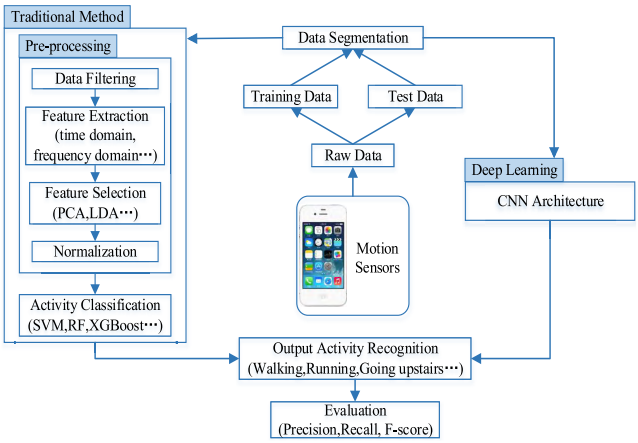


FIGURE 1. The human activity recognition system models.

The remainder of this paper is organized as follows: Section II introduces the background and related works. Section III provides details of the data acquisition and preprocessing procedure. Section IV describes the CNN-based framework of the ensemble model and some traditional classification algorithms. Section V presents our experimental results and improvements. Section VI concludes the paper and discusses ideas about future work.

II. BACKGROUND AND RELATED WORKS

HAR can be seen as a classification problem to discover human physical activity patterns by analyzing motion data. The input data is the motion signals collected from smartphone's motion sensors and the output is the activity class label. Fig. 1 shows two typical activity recognition models, both traditional methods and deep learning model.

A. TRADITIONAL METHODS FOR HAR

The most generally used traditional algorithms are KNN, HMM, SVM, RF, XGBoost, etc. These algorithms take three steps including raw data preprocessing, feature extraction and feature selection before recognition. In most related works [20], filtering techniques, like mean filter, low-pass filter, Gaussian filter and Kalman filter, are used to mitigate the effect of noise in obtained data. This is due to the fact that raw sensor data are always noise-corrupted, which makes it hard to measure and reflect the true motion change of smartphones accurately. After preprocessing the raw data, traditional methods extract a large amount of features and select some principal features [21] representing the essential difference between different activities. Features extracted from the time domain, frequency domain, wavelet energy

and interquartile range are extensively used. PCA or LDA is widely implemented to select the dominating features. In addition, normalization of the feature vector can control the number of features within a certain range.

Many HAR researches [22], [23] place sensors in various parts of body ignoring the practicability of the solution. Smartphones which has become daily supplies for most people have drawn researchers for their plenty of computing power and multiple sensors. Lee and Cho [24] utilized the tri-axial accelerometer from a handheld smartphone to identify five activities with hierarchical hidden Markov models. Motion data from four participants were collected. The result showed difficulty in distinguishing upstairs and downstairs movements. Kwapisz et al. [25] collected the acceleration data of 29 users from smartphones placed in the subjects' front trousers pocket. They extracted six features and built up to four classifiers to achieve an accuracy of over 90% for most activities. Sun et al. [26] proposed an activity recognition approach using an accelerometer to recognize seven physical activities based on six pocket positions. They extracted features from the collected data of seven subjects, including time domain and frequency domain features. With the prior knowledge of known pocket position, the overall F-score can reach 94.8% of the trained SVM classifier.

B. DEEP LEARNING FOR HAR

The works described above heavily rely on heuristic handcrafted feature extraction, which is usually limited by empirical knowledge of the researchers. Furthermore, approaches using handcrafted features make it very difficult to compare between different algorithms due to different experimental grounds and encounter difficulty in discriminating very similar activities. As a result of those limitations, the performance of traditional pattern recognition algorithms is very restricted in terms of classification accuracy and model generalization. Different from traditional methods, deep learning can greatly relieve the effort on designing features and easily learns more meaningful high-level features by training the end-to-end neural network. Therefore, we reckon that deep learning has the capacity to do HAR which has been widely proved in the existing work [27]–[31].

CNN, a deep learning method, has established itself as a powerful technique because representations learned by CNNs can efficiently capture local dependency and scale invariance of a signal. The authors in [29] built an end-to-end CNN model to predict three arm movements performed in the daily activity. Motion data was collected from four different subjects using a wrist-worn tri-axial accelerometer sensor. The results achieved an average recognition rate up to 99.8%. Ming Et al. [30] proposed an approach based on CNN to recognize activities in various application domains. A modified weight sharing technique, called partial weight sharing, was proposed and applied to acceleration signals to achieve further improvements. The experimental results on three public datasets: Skoda, Opportunity, Actitracker, indicated that their novel CNN-based approach can achieve higher accuracy than existing state-of-the-art methods. Chen and Xue [31] collected acceleration data from eight typical activities of 100 subjects to achieve better performance (an accuracy of 93.8%) than SVM and Deep Belief Network (DBN).

In this paper, we propose a novel ensemble model based on CNN which effectively solves the confusion of highly similar activities such as going upstairs and walking. To evaluate the performance of the ensemble model, extensive comparative experiments are conducted using traditional methods including XGBoost and RF. The experiment results show our approach outperforms traditional methods and achieved higher accuracy up to 96.11%.

III. DATA ACQUISITION AND PREPROCESSING

Generally speaking, for a multi-class classification problem, a large amount of training data are required especially with the presence of a high dimension of the feature vector. In addition, rich features from a large amount of training data can effectively prevent overfitting and make the model robust. The data of this paper come from various sports scenes with different participants and device placements. The data were collected in a way to ensure the data amount of each activity is nearly the same. In this section, we will describe considerable details regarding data collection and preprocessing.

A. DATA COLLECTION

Many open source databases focusing on sensor-based activity recognition mainly provide a single accelerometer data collected from the smartphone in participants' trouser pocket or on the waist at a low sampling rate. To make things worse, these data have poor quantity and unbalanced distribution in various activities, which makes it difficult to construct a highly accurate classification model. To improve this situation, it is more than necessary to have a large number of participants contributing to the activity data set.

The experiment data of this paper are collected from accelerometer, gyroscope and magnetometer in ordinary smartphones at a sampling rate of 50Hz. A group of 100 participants ageing from 12 to 51 years was invited to finish the whole experiment. Each participant was asked to complete seven human activities including Going Upstairs (GU), Going Downstairs (GD), Standing (SD), Running (RU), Walking (WK), Bicycling (BY) and Swinging (SW) (the smartphone is possibly periodically shaken or tapped while not moving), as shown in Fig 2. SW is implemented to detect true pace accurately while disregarding motion signals such as shaking or tapping which could be mistaken for walking. We selected these activities because they are performed regularly by many people in their daily routines. This study takes into account four smartphones placement settings:

- Texting: The smartphone is held in front of the user while the carrier is performing a certain everyday activity;
- Handheld: The smartphone is held in a swinging hand while the carrier is performing a certain everyday activity;

FIGURE 2. Typical daily human activities.

TABLE 1. Motion sensor features representation.

Attribute	Description	Feature Identifier
Mean	Mean of sample($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	1∼2
Std	Standard deviation of sample($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	13∼24
Max	The maximal value of sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	25∼36
Min	The minimal value of sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	37∼48
Max−Min	The difference between the maximal value and minimal value of sample	49∼60
Median	The median value of sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	61∼72
RMS	Root mean square of sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	73∼84
TQ	tri-quartileof sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	85∼96
IQR	Interquartile range of sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	97∼108
Corr	Correlation of $(a_x, a_y), (a_x, a_z), (a_y, a_z), (g_x, g_y), (g_x, g_z), (g_y, g_z), (m_x, m_y), (m_x, m_z), (m_y, m_z)$	109∼117
FFT	First $1^{st} \sim 5^{st}$ amplitude of FFT of sample ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, \|a\|, \|g\|, \|m\|$)	118∼178

- Trouser Pocket: The smartphone is put in a trouser pocket(front) while the carrier is performing a certain everyday activity;
- Backpack: The smartphone is put in a backpack while the carrier is performing a certain everyday activity.

In our experiments, the sensor data of each participant was collected three times for each activity and smartphone placement setting within the duration of one minute. For going upstairs and downstairs, a 6-floor building with stairs was used. Bicycling dataset only contains two smartphone placements including backpack or trouser pocket. The swinging dataset is collected while the smartphone is handheld. As a result, the size of these activities is smaller than other five activities. Table 2 describes the detailed class distribution of the experiment data.

B. DATA PREPROCESSING

1) DATA PREPROCESSING OF TRADITIONAL CLASSIFIER

Based on the collected data, we compare the performance of some traditional methods with the proposed model in this paper. For each sample data, time domain features including mean, variance, root mean square, the maximum and minimum values of the axis, range, interquartile distance, correlation coefficients and frequency domain feature of amplitude of FFT are extracted to form a 178-dimensional feature vector as shown in Table 1.

TABLE 2. Class distribution of data.

Activity	Distribution
Going Upstairs (GU)	41371 (17.53%)
Going Downstairs (GD)	38497 (16.31%)
Standing (SD)	39467 (16.73%)
Running (RU)	438699 (18.59%)
Walking (WK)	43660 (18.50%)
Bicycling (BY)	20337 (8.62%)
Swinging (SW)	8778 (3.72%)

2) DATA PREPROCESSING OF DEEP CLASSIFIER

In the end-to-end deep architecture proposed in this paper, there is no need to perform additional processing on the data and the raw signal is directly used. To meet the format requirement of the proposed CNN model, a sliding window segmentation approach with fixed step size of fifty seconds is applied to each sensor data. In our work, the raw sensor data stream is cropped into the same size with an overlap of 25%. Every sample is a matrix with the size of 200 (data of four seconds)×3 (three motion sensors)×3 (X, Y, and Z axis data). After simple processing, we have 235 977 labeled samples.

IV. SYSTEM MODEL

In this section, we firstly introduce our ensemble model based on CNN. Fig. 3 shows a structure of CNN. Then, we give some traditional classifiers a brief illustration.

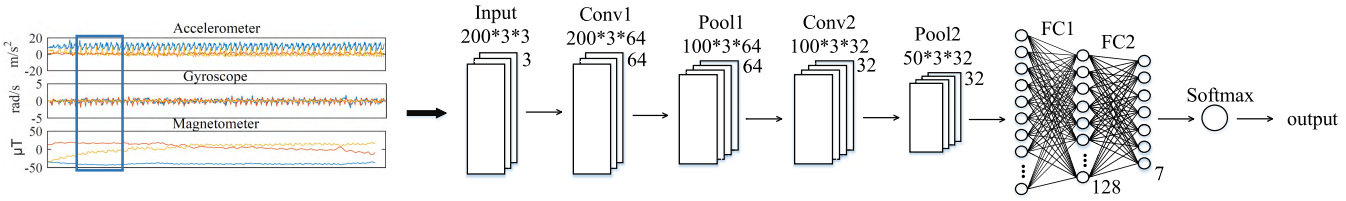


FIGURE 3. Structure of CNN-based human activity recognition model. The numbers of the first and second convolution kernels are 64 and 32 respectively.

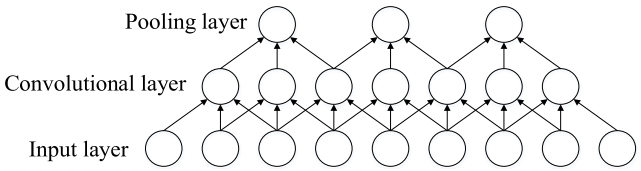


FIGURE 4. Key part of CNN.

A. DEEP ARCHITECTURE

The proposed CNN-based model has five kinds of layers: 1) an input layer, as described in Section III; 2) convolutional layers extract features from input data; 3) max-pooling layers reduce the size of extracted features and enhance the robustness of some detected features; 4) fully connected layers integrate all features extracted; 5) an output layer of the softmax function represents a categorical distribution over seven different activities.

1) CONVOLUTIONAL LAYER

CNN is different from other neural networks in terms of sparse connectivity between units of adjacent layers and parameter sharing in the same layer. For example, in Fig. 4, the units in the middle layer are only connected to a local subset of units in the input layer. CNN uses local filters in input space for feature extraction, which perform inner product operation of local filters and use the output result as the value of the corresponding dimension of the convolutional output matrix.

Suppose we have a N-units layer as the input followed by convolutional layer. If we use an m-size filter, the output will be $(N-m+1)$ units. The detailed calculation of convolutional layer l is as follows:

$$x_i^{l,j} = f(\Sigma_{a=1}^{m} w_a^j x_{i+a-1}^{l-1,j} + b_j), \tag{1}$$

where $x_i^{l,j}$ is the output of jth feature map on the ith unit of the convolutional layer l. w_a^j is the convolutional kernel matrix and b_j is the bias of convolutional feature maps. Weights are convoluted with previous layer output feature map before summed with the bias. Then the nonlinear mapping is performed through the activation function f. Our model uses the reluctant function $relu(\cdot)$. Take Fig. 4 as an example, the first hidden unit of the first local filter is:

$$x_1^{1,1} = relu(w_1^1 x_1^{0,1} + w_2^1 x_2^{0,1} + w_3^1 x_3^{0,1} + b_1) \tag{2}$$

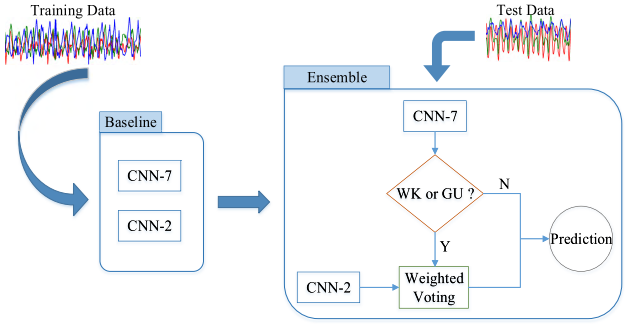


FIGURE 5. The ensemble of CNN for human activity recognition.

2) MAX-POOLING LAYER

This research uses max-pooling strategy to address the activity recognition problem. Once features have been detected in the convolutional layer, max-pooling layers, without breaking the internal relationship of data, reduce the size of extracted features and make some features more robust. The activation function in the max-pooling layer in CNN is given by:

$$x_i^{l,j} = max_{i,j=1}^{r}(x_{i,j}), \tag{3}$$

where $x_i^{l,j}$ represents a local output after the pooling process, and r is the size of pooling kernel.

In the max-poling layer, features extracted in the convolutional layer are split into several partitions. The maximum values are given as the output of each partition. The input data size of the first max-pooling layer is (200, 3, 64), and the output data size is (100, 3, 64). After the last max-pooling, the data size obtained is (50, 3, 32), indicating that both the data dimension and network parameters have been greatly reduced.

B. ENSEMBLE MODEL

In this section we propose a novel framework based on the ensemble of CNN to tackle the confusion in human activity recognition, which greatly improves the robustness of existed models. In our training procedure, there are two CNN models called seven-class network (CNN-7) and two-class network (CNN-2). In our test procedure, two CNN models perform weighted voting to recognize human activities. It shows that the ensemble learning approach is fairly efficient to distinguish the confusion between certain highly similar and thus

confusing activities like going upstairs and walking. The detailed process is described in Algorithm 1 below.

Algorithm 1 The Framework of Ensemble of CNN

Input:

Testing data:$X_{i=1}^{MaxSample}$

Output:

Human activity (GU, GD, RU, WK, SD, BY or SW)

for i to $MaxSample$ **do**
  CNN-7 network gives the predicted activity
  if activity is GU or WK **then**
    Normalize the probability of GU and WK P_u^1, P_w^1
    The CNN-2 network gives another prediction P_u^2, P_w^2
    Define weights:$\alpha = \frac{P_u^1}{P_u^1+P_u^2}$, $\beta = \frac{P_w^1}{P_w^1+P_w^2}$
    if activity is GU **then**
      $P_u = \beta \times P_u^1 + (1-\beta) \times P_u^2$
      $P_w = \beta \times P_w^1 + (1-\beta) \times P_w^2$
    else
      $P_u = \alpha \times P_u^1 + (1-\alpha) \times P_u^2$
      $P_w = \alpha \times P_w^1 + (1-\alpha) \times P_w^2$
    end if
    return The predicted activity from $max(P_u, P_w)$
  else
    return The predicted activity from CNN-7
  end if
end for

In our ensemble learning framework, CNN-7 is used to identify seven activities and CNN-2 is designed to distinguish two confusing human activities which generate highly similar signal patterns: going upstairs and walking. If the output of CNN-7 is neither going upstairs nor walking, this output would serve as the final decision. But if the output of CNN-7 is going upstairs or walking, we will combine the prediction of CNN-2 to improve the recognition accuracy of these two confusing activities.

C. TRADITIONAL CLASSIFICATION ALGORITHM

1) XGBOOST

XGBoost [32] is one of the boosting algorithms, which can promote weak learners to evolve into strong learners. It adds a regularization term to the loss function to control the complexity of the tree compared to the traditional model. XGBoost that comprises multiple classifications and regression trees selects the best classification point according to certain strategies. Disposing the sparse data and adding parallel processing as an optimization method makes XGBoost efficient and robust.

2) RANDOM FOREST

Random Forest [33] adopts the idea of the ensemble learning and integrates decisions from multiple trees. Running efficiency on large databases and ability to handle thousands of input variables without variable deletion are the two principal superiority of random forest, compared with other traditional methods. A random forest consisting of N decision trees will produce N classification results for an input sample, because each of the N decision trees in the random forest is a separate classifier. Using the simple Bagging idea, random forest specifies the category with the most votes from all the classification voting results as the final output.

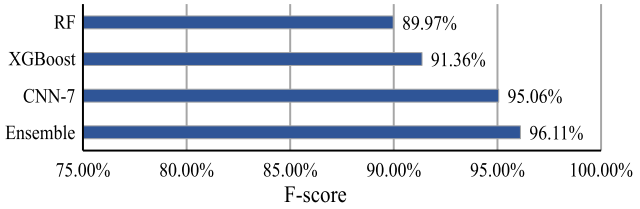


FIGURE 6. Classification results on various classifiers.

V. EXPERIMENTAL RESULTS AND EVALUATION

In this section, we will evaluate the recognition performance of the proposed method through extensive real-world experiments. We first introduce the allocation of training and test dataset before we compare and analyze the recognition accuracy of different classifiers. The impacts of smartphone placements and activity types on recognition accuracy are then evaluated. In the last part, we verify the performance of the novel approach we proposed to improve CNN with the ensemble model.

A. TRAINING AND TESTING PROCEDURE

To evaluate the recognition performance, we divide all collected data samples into training dataset and testing dataset. To make a widely applicable model, we use an individual-based 10-fold evaluation approach where all data samples from 10 random participants out of the 100 participants are selected as the test data and the rest training data. This method takes into account the applicability of the recognition framework for testing data from individuals totally different from the training data and thus examine whether the generalized model can be applied to real world scenarios.

B. CLASSIFICATION ACCURACY

We compared our method with three frequently used methods. Extensive experiments have been conducted and the results show that the classification accuracy of the ensemble model outperforms other models up to 96.11%, which proves the feasibility and effectiveness of the proposed approach in this paper. To analyze the results in more details, we compare the performance of each classifier in different smartphone placement settings of different subjects. To reduce the impact of data imbalance, we calculate the F-score of each activity and finally use average F-score as the criterion. The performance of each classifier is shown in Fig 6.

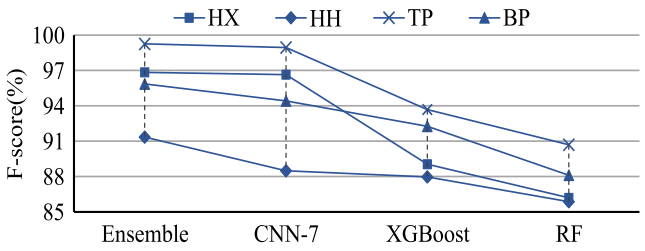


FIGURE 7. F-score of different classifiers for various placement settings. Legend represents the placement settings: "TX"- texting mode, "HH"- handheld mode, "TP"- trouser pocket mode, "BP"- backpack mode.

TABLE 3. Ensemble model (%).

	TX	HS	TP	BP	M
Test1	96.14	98.38	**99.18**	94.07	97.03
Test2	86.34	94.28	**99.46**	94.79	94.27
Test3	98.00	96.04	**98.98**	95.96	97.21
Test4	98.16	83.07	**100**	99.71	95.26
Test5	**99.23**	100	**99.23**	98.10	97.64
Test6	**99.70**	77.95	99.59	99.09	94.70
Test7	**96.58**	94.54	96.09	91.00	94.63
Test8	97.96	95.36	97.50	**98.48**	97.41
Test9	97.02	89.86	**99.28**	95.61	95.26
Test10	95.58	97.47	**98.95**	90.21	95.39

1) ACTIVITY RECOGNITION UNDER VARIOUS PLACEMENTS

The motion patterns of the mobile phone are very different for different smartphone placements [34], [35]. As a result, the sensor signals can be very different accordingly. For example, smartphones in the backpack are usually looser and deeper than those in trouser pockets. This usually leads to a higher vibration magnitude while walking or running. Moreover, different parts of the body show different patterns. For example, a smartphone placed in the trouser pocket records how the thigh moves while a handheld smartphone records how the arm swings. To investigate the effects of varying sensor placements on activity recognition, we calculate the F-score of each classifier under four different placement settings: (*i*) texting mode; (*ii*) handheld mode; (*iii*) trouser pocket mode; and (*iv*) backpack mode, as shown in in Fig. 7. Table 3-6 provides the detailed classification results of each individual. "M" means no position information and a mix of four positions data together.

It can be observed that the recognition performances of different smartphone placements can vary significantly. For instance, the placement of trouser pocket can be easily recognized and therefore achieves the best recognition accuracy and robustness for each classifier. In comparison, handheld placement is hard to deal with and has the lowest recognition accuracy. The underlying reason lies in the fact that the movements of thighs are quite restricted and thus easy to be identified while the movements of hands can be quite complex and difficult to be identified. Experiment results have shown that deep learning based model significantly outperforms the traditional algorithms in any smartphone placement.

TABLE 4. Convolutional neural network-7 (%).

	TX	HS	TP	BP	M
Test1	96.14	**98.85**	98.53	92.83	96.69
Test2	88.06	97.40	**98.98**	97.00	94.41
Test3	98.99	95.25	**99.73**	90.60	97.47
Test4	98.29	73.27	**99.89**	97.90	92.59
Test5	99.85	100	99.85	**99.79**	98.10
Test6	98.80	78.35	**99.59**	99.31	94.64
Test7	95.87	89.67	**99.75**	93.41	94.80
Test8	89.90	89.93	95.78	**98.58**	95.44
Test9	97.02	91.03	**98.65**	92.67	91.87
Test10	96.10	91.56	**98.76**	82.71	94.89

TABLE 5. eXtreme gradient boosting (%).

	TX	HS	TP	BP	M
Test1	87.58	87.53	**96.10**	95.90	91.70
Test2	56.85	86.93	78.36	**94.48**	84.62
Test3	90.63	94.01	**97.98**	87.34	92.70
Test4	92.48	86.74	**97.78**	91.02	92.34
Test5	89.08	86.87	**93.65**	88.18	90.63
Test6	93.32	92.44	**96.91**	96.83	95.34
Test7	90.67	82.89	**98.27**	92.01	91.22
Test8	**95.55**	87.83	89.74	80.82	91.72
Test9	85.88	78.45	81.10	**91.26**	84.53
Test10	95.68	95.16	**97.39**	89.88	95.71

TABLE 6. Random forests (%).

	TX	HS	TP	BP	M
Test1	82.67	82.44	**95.28**	92.71	87.93
Test2	62.88	84.34	79.45	**88.30**	83.39
Test3	90.51	93.86	**95.72**	90.83	93.23
Test4	91.24	88.93	**94.50**	92.28	91.86
Test5	89.75	90.71	**93.32**	87.08	90.54
Test6	92.28	83.40	82.68	**97.09**	89.14
Test7	90.65	80.06	**95.74**	85.20	88.23
Test8	87.72	84.27	**92.11**	91.13	89.07
Test9	81.09	78.29	80.85	**83.90**	81.27
Test10	85.21	**91.54**	85.37	84.87	86.76

2) ACTIVITY RECOGNITION UNDER VARIOUS ACTIVITIES

There are numerous and complex human activities, which makes the boundaries of different activities blurry. Even for the same activity, different individuals have very different motion modes. Fig. 8 evaluates the performance of different classifiers for each activity.

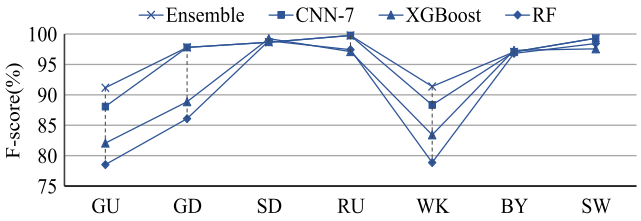


FIGURE 8. F-score of different classifiers for various activities.

TABLE 7. Ensemble model (%).

	GU	DU	SD	RU	WK	BY	SW
Test1	99.28	95.61	98.20	99.08	92.94	96.38	100
Test2	82.89	99.43	99.77	100	90.24	99.29	100
Test3	96.42	98.56	97.24	100	97.58	95.12	95.43
Test4	86.34	98.96	99.93	100	85.23	99.86	99.67
Test5	95.74	98.31	100	100	96.27	99.88	100
Test6	80.18	98.99	100	100	85.42	99.40	99.69
Test7	93.05	96.59	96.88	100	92.13	93.01	99.36
Test8	92.06	98.13	99.59	99.71	94.11	99.13	99.68
Test9	90.29	98.32	97.76	99.49	88.99	98.29	99.71
Test10	93.24	97.01	97.11	99.41	90.98	90.54	99.68

TABLE 8. Convolutional neural network-7 (%).

	GU	DU	SD	RU	WK	BY	SW
Test1	97.31	95.61	98.20	99.08	91.71	96.38	100
Test2	80.13	99.43	99.77	100	87.80	99.29	100
Test3	95.69	98.56	97.24	100	96.93	95.12	95.43
Test4	81.39	98.96	99.93	100	76.95	99.86	99.67
Test5	95.04	98.31	100	100	95.65	99.88	100
Test6	79.77	98.99	100	100	85.35	99.40	99.69
Test7	88.47	96.59	96.88	100	89.12	93.01	99.36
Test8	86.82	98.13	99.59	99.71	89.89	99.13	99.68
Test9	89.28	98.32	97.76	99.49	87.79	98.29	99.71
Test10	83.56	97.01	97.11	99.41	80.96	90.54	99.68

From Fig. 8, we can infer that all classifiers can achieve an F-score above 95% in standing, running, bicycling and swinging. Note that even if there are a very limited number of examples of bicycling and swinging, we can still identify these activities quite well. Nevertheless, the performance of distinguishing between going upstairs and walking is quite unsatisfactory, especially for the traditional classifier. Deep learning algorithm achieves the best prediction accuracy among all models, slightly above 97% excluding the recognition of going upstairs and walking. To tackle the confusion of these two activities, we further propose a CNN with ensemble model. Table 7-10 shows a further relation between recognition accuracy and individuals in detail. We can find that recognition accuracy may be drastically different between different individuals even under the same activity and algorithm. Therefore, the increase of data from different individuals could improve human activity recognition accuracy.

TABLE 9. eXtreme gradient boosting (%).

	GU	DU	SD	RU	WK	BY	SW
Test1	90.84	89.61	98.83	95.35	83.03	92.63	89.13
Test2	65.77	91.79	100	89.68	66.76	98.61	96.91
Test3	78.83	95.14	96.90	99.45	89.54	89.29	99.73
Test4	87.28	86.81	100	96.64	84.92	99.17	98.69
Test5	83.99	68.47	99.84	99.61	82.38	99.76	99.08
Test6	81.77	94.90	99.85	99.50	91.97	99.70	99.69
Test7	69.42	95.33	99.21	99.36	78.70	97.91	100
Test8	79.87	92.32	99.84	96.80	85.02	97.30	99.37
Test9	67.39	81.76	97.89	97.72	67.90	98.14	98.84
Test10	87.69	95.68	99.77	98.01	92.20	99.00	100

TABLE 10. Random forests (%).

	GU	DU	SD	RU	WK	BY	SW
Test1	81.59	88.59	98.83	97.03	70.62	86.82	92.81
Test2	63.19	85.87	99.92	94.35	61.36	99.72	99.66
Test3	82.58	94.29	95.56	99.52	91.23	85.98	98.61
Test4	86.30	87.85	99.86	96.55	83.26	97.44	98.69
Test5	82.90	73.23	99.69	99.41	81.40	99.52	99.70
Test6	65.68	91.39	99.54	99.43	71.88	99.10	99.39
Test7	68.33	90.31	98.05	98.45	72.68	94.43	100
Test8	75.68	84.24	99.59	96.20	83.90	95.21	99.37
Test9	62.63	75.69	98.10	98.00	58.97	98.29	99.71
Test10	71.61	87.14	99.77	94.13	68.27	99.00	99.68

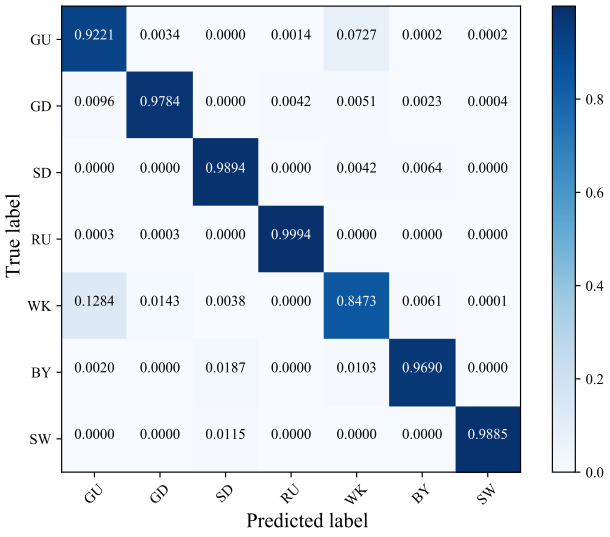


FIGURE 9. The normalized confusion matrix of CNN-7.

3) PERFORMANCE OF ENSEMBLE MODEL

The task is to distinguish between two confusing activities: going upstairs and walking which bring the major prediction errors known from confusion matrices. The CNN-7 in Fig. 9 indicates that going upstairs is usually misclassified to walking, which leads to a 7.27% decrease in prediction accuracy. Similarly, walking is misclassified as going upstairs, leading to a further 12.84% decrease in prediction accuracy.

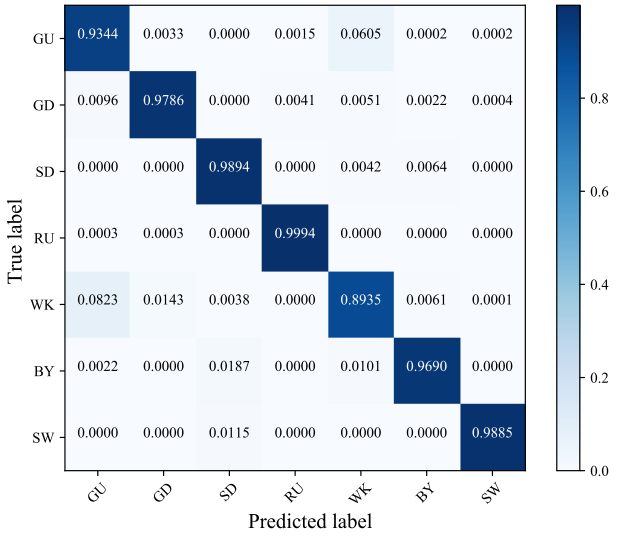


FIGURE 10. The normalized confusion matrix of ensemble of CNN.

To reduce the confusion between these two activities, we propose an ensemble model based on CNN which can achieve up to 96.11% accuracy. The confusion matrix of the model is shown in Fig. 10. We can see that the prediction accuracy has been improved significantly, even if going upstairs and walking are still the two most confusing activities.

VI. CONCLUSION

This paper has proposed a CNN-based human activity recognition model using the nine-axis motion signals of accelerometer, gyroscope and magnetometer in common smartphones. We have compared and analyzed the performance of different algorithms with seven daily activities and four different placements of smartphones. In order to further improve the recognition accuracy, this paper has developed an ensemble model based on CNN which extracts the local dependence and scale invariant characteristics of the sensor time series and reached an accuracy up to 96.11%. In the future, to verify the robustness and practicality of the model, we will conduct further experiments with larger datasets to recognize more human activities under more placements of smartphones.

REFERENCES

[1] M. Elhoushi, J. Georgy, A. Noureldin, and M. J. Korenberg, "Motion mode recognition for indoor pedestrian navigation using portable devices," *IEEE Trans. Instrum. Meas.*, vol. 65, no. 1, pp. 208–221, Jan. 2016.

[2] W. Kang and Y. Han, "SmartPDR: Smartphone-based pedestrian dead reckoning for indoor localization," *IEEE Sensors J.*, vol. 15, no. 5, pp. 2906–2916, May 2015.

[3] P. Wu, H.-K. Peng, J. Zhu, and Y. Zhang, "SensCare: Semi-automatic activity summarization system for elderly care," in *Mobile Computing, Applications, and Services*. Berlin, Germany: Springer, 2011, pp. 1–19.

[4] G. Sagl, B. Resch, and T. Blaschke, "Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities," *Sensors*, vol. 15, no. 7, pp. 17013–17035, Jul. 2015.

[5] U. A. Akansha, M. Shailendra, and N. Singh, "Analytical review on video-based human activity recognition," in *Proc. 3rd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, New Delhi, Delhi, Mar. 2016, pp. 3839–3844.

[6] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognition—A review," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 865–878, Nov. 2012.

[7] M. Elhoushi, J. Georgy, A. Noureldin, and M. J. Korenberg, "A survey on approaches of motion mode recognition using sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 7, pp. 1662–1686, Jul. 2017.

[8] G. De Leonardis, S. Rosati, G. Balestra, V. Agostini, E. Panero, L. Gastaldi, and M. Knaflitz, "Human activity recognition by wearable sensors: Comparison of different classifiers for real-time applications," in *Proc. IEEE Int. Symp. Med. Meas. Appl. (MeMeA)*, Rome, Italy, Jun. 2018, pp. 1–6.

[9] M. Ermes, J. PÄrkkÄ, J. MÄntyjÄrvi, and I. Korhonen, "Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 1, pp. 20–26, Jan. 2008.

[10] Z. Xiao, H. Wen, A. Markham, and N. Trigoni, "Robust pedestrian dead reckoning (R-PDR) for arbitrary mobile device placement," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Busan, South Korea, Oct. 2014, pp. 187–196.

[11] Y.-S. Lee and S.-B. Cho, "Activity recognition using hierarchical hidden Markov models on a smartphone with 3D accelerometer," in *Proc. 6th Int. Conf. Hybrid Artif. Intell. Syst.*, Berlin, Germany, 2011, pp. 460–467.

[12] F. Foerster, M. Smeja, and J. Fahrenberg, "Detection of posture and motion by accelerometry: A validation study in ambulatory monitoring," *Comput. Hum. Behav.*, vol. 15, no. 5, pp. 571–583, Sep. 1999.

[13] V. Ayumi, "Pose-based human action recognition with extreme gradient boosting," in *Proc. IEEE Student Conf. Res. Develop. (SCOReD)*, Kuala Lumpur, Malaysia, Dec. 2016, pp. 1–5.

[14] M. T. Uddin, M. M. Billah, and M. F. Hossain, "Random forests based recognition of human activities and postural transitions on smartphone," in *Proc. 5th Int. Conf. Informat., Electron. Vis. (ICIEV)*, Dhaka, Bangladesh, May 2016, pp. 250–255.

[15] Z. Chen, Q. Zhu, S. Y. Chai, and L. Zhang, "Robust human activity recognition using smartphone sensors via CT-PCA and online SVM," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3070–3080, Dec. 2017.

[16] K. G. M. Chathuramali and R. Rodrigo, "Faster human activity recognition with SVM," in *Proc. Int. Conf. Adv. ICT Emerg. Regions (ICTer)*, Colombo, Sri Lanka, Dec. 2012, pp. 197–203.

[17] T. Huynh and B. Schiele, "Analyzing features for activity recognition," in *Proc. Joint Conf. Smart Objects Ambient Intell., Innov. Context-Aware Services, Usages Technol.*, New York, NY, USA, 2005, pp. 159–163.

[18] X.-S. Wei, C.-W. Xie, and J. Wu, "Mask-CNN: Localizing parts and selecting descriptors for fine-grained image recognition," 2016, *arXiv:1605.06878*. [Online]. Available: https://arxiv.org/abs/1605.06878

[19] P. Le-Hong and A.-C. Le, "A comparative study of neural network models for sentence classification," in *Proc. 5th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, Ho Chi Minh City, Vietnam, Nov. 2018, pp. 360–365.

[20] M. D. E. Beily, M. D. Badjowawo, D. O. Bekak, and S. Dana, "A sensor based on recognition activities using smartphone," in *Proc. Int. Seminar Intell. Technol. Appl. (ISITIA)*, Lombok, Indonesia, Jul. 2016, pp. 393–398.

[21] A. Bhavan and S. Aggarwal, "Stacked generalization with wrapper-based feature selection for human activity recognition," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Bangalore, India, Nov. 2018, pp. 1064–1068.

[22] I. Cleland, B. Kikhia, C. Nugent, A. Boytsov, J. Hallberg, K. Synnes, S. McClean, and D. Finlay, "Optimal placement of accelerometers for the detection of everyday activities," *Sensors*, vol. 13, no. 7, pp. 9183–9200, 2013.

[23] H. Gjoreski, M. Lustrek, and M. Gams, "Accelerometer placement for posture recognition and fall detection," in *Proc. 7th Int. Conf. Intell. Environ.*, Nottingham, U.K., Jul. 2011, pp. 47–54.

[24] Y.-S. Lee and S.-B. Cho, "Activity recognition using hierarchical hidden Markov models on a smartphone with 3D accelerometer," in *Hybrid Artificial Intelligent Systems*, vol. 6678. Amsterdam, The Netherlands: IOS Press, 2011, pp. 460–467.

[25] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[26] L. Sun, D. Zhang, B. Li, B. Guo, and S. Li, "Activity recognition on an accelerometer embedded mobile phone with varying positions and orientations," in *Proc. Int. Conf. Ubiquitous Intell. Comput.*, 2010, pp. 548–562.

[27] S. Ha and S. Choi, "Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 381–388.

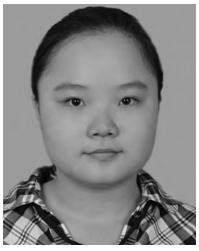
[28] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. 24th Int. Conf. Artif. Intell.*, Buenos Aires, Argentina, 2015, pp. 3995–4001.
[29] M. Panwar, S. R. Dyuthi, K. C. Prakash, D. Biswas, A. Acharyya, K. Maharatna, A. Gautam, and G. R. Naik, "CNN based approach for activity recognition using a wrist-worn accelerometer," in *Proc. 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Seogwipo, South Korea, Jul. 2017, pp. 2438–2441.
[30] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th Int. Conf. Mobile Comput., Appl. Services*, Austin, TX, USA, Nov. 2014, pp. 197–205.
[31] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Kowloon, China, Oct. 2015, pp. 1488–1492.
[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," 2016, *arXiv:1603.02754*. [Online]. Available: https://arxiv.org/abs/1603.02754
[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.
[34] A. Henpraserttae, S. Thiemjarus, and S. Marukatat, "Accurate activity recognition using a mobile phone regardless of device orientation and location," in *Proc. Int. Conf. Body Sensor Netw.*, Dallas, TX, USA, May 2011, pp. 41–46.
[35] Y. Chen and C. Shen, "Performance analysis of smartphone-sensor behavior for human activity recognition," *IEEE Access*, vol. 5, pp. 3095–3110, 2017.

RAN ZHU received the B.Eng. degree from Jilin University, Changchun, China, in 2018. She is currently pursuing the master's degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. Her current research interests include indoor localization, and machine learning techniques for sensor networks and indoor localization.

ZHUOLING XIAO received the Ph.D. degree from the University of Oxford, where he was a Postdoctoral Researcher. He is currently an Associate Professor with the University of Electronic Science and Technology of China. His research interests include localization protocols for networked sensor nodes and machine learning techniques for sensor networks and localization. He has several international patent applications and over 30 papers published in leading journals and conferences, including several best paper awards from leading conferences, including IPSN and EWSN.

YING LI received the B.Eng. degree from Qingdao University, Qingdao, China, in 2018. She is currently pursuing the master's degree with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. Her current research interests include indoor localization, machine learning, and information fusion.

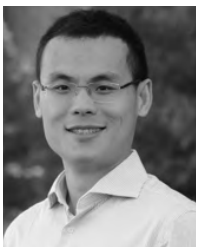
MINGKUN YANG received the B.Eng. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2018, where he is currently pursuing the master's degree with the School of Information and Communication Engineering. His research interests focus on the application of machine learning techniques in sensor networks and indoor localization.



YAWEN TAN is currently pursuing the B.Eng. degree with the Glasgow College, University of Electronic Science and Technology of China, Chengdu, China. Her current research interests include indoor localization and mobile sensor systems.

LIANG ZHOU received the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2014, where he is currently an Associate Professor. His research interests are focused on wireless sensor networks, indoor positioning/tracking, and signal processing for wireless communications.

SHUISHENG LIN received the master's and Ph.D. degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 1989 and 1998, respectively, where he is currently a Professor and the Dean of the Internet of Things Engineering. He has authored or coauthored 14 books, nine invention patents, and over 40 scientific research papers in his research areas. His research interests include wireless and mobile communications, the Internet of Things, ad hoc networks, and embed systems.



HONGKAI WEN received the D.Phil. degree from the University of Oxford, where he was a Postdoctoral Researcher with the Oxford Computer Science and Robotics Institute. He is currently an Assistant Professor with the Department of Computer Science, University of Warwick. His research interests include mobile sensor systems, human-centric sensing, and pervasive data science.