IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 8, 2018, accepted December 29, 2018, date of publication January 1, 2019, date of current version January 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2890675

InnoHAR: A Deep Neural Network for Complex Human Activity Recognition

CHENG XU[1,2], (Student Member, IEEE), DUO CHAI[1,2], JIE HE[1,2], XIAOTONG ZHANG[1,2], (Senior Member, IEEE), AND SHIHONG DUAN[1,2]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing 100083, China

Corresponding authors: Jie He (hejie@ustb.edu.cn) and Xiaotong Zhang (zxt@ies.ustb.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFC0901303, in part by the National Natural Science Foundation of China (NSFC) under Projects 61671056, 61302065, 61304257, and 61402033, in part by the Beijing Natural Science Foundation under Project 4152036, and in part by the Tianjin Special Program for Science and Technology under Grant 16ZXCXSF00150.

ABSTRACT Human activity recognition (HAR) based on sensor networks is an important research direction in the fields of pervasive computing and body area network. Existing researches often use statistical machine learning methods to manually extract and construct features of different motions. However, in the face of extremely fast-growing waveform data with no obvious laws, the traditional feature engineering methods are becoming more and more incapable. With the development of deep learning technology, we do not need to manually extract features and can improve the performance in complex human activity recognition problems. By migrating deep neural network experience in image recognition, we propose a deep learning model (InnoHAR) based on the combination of inception neural network and recurrent neural network. The model inputs the waveform data of multi-channel sensors end-to-end. Multi-dimensional features are extracted by inception-like modules by using various kernel-based convolution layers. Combined with GRU, modeling for time series features is realized, making full use of data characteristics to complete classification tasks. Through experimental verification on three most widely used public HAR datasets, our proposed method shows consistent superior performance and has good generalization performance, when compared with the state-of-the-art.

INDEX TERMS Complex human activity, inception neural network, wearable sensor, computational efficiency.

I. INTRODUCTION

Body area network (BAN) [1], [2] is an extension of traditional wireless sensor network, aiming to provide ideal wireless setting for pervasive hearlth care. Human activity recognition (HAR) is a major goal of BAN, which tries to realize the discrimination of complex human actions and behaviors by observing the human body parts and the surrounding environment [1], [8]. During the last decade, more and more technologies and methods have been applied to sensor-based HAR [4], [5], [9]. It has been widely used in medical care [3]–[5], athletic competition [6], smart home [7] and many other applications.

Deep learning makes computer vision (CV) efficient to solve the problem of human activity recognition [7], [13], [15]. However, there are still many deficiencies in the CV-based human activity recognition scheme, such as: 1) the interference of complex and variable backgrounds on activity recognition; 2) the difficulty of positioning, tracking and recognition caused by multiple active subjects simultaneously appearing in the picture; 3) the demands for strict environment conditions for light, brightness and contrast. Besides, targets are easily occluded and it significantly limits the applications of CV in actual scenes.

Wearable sensors are also widely used in human activity recognition and motion capture applications, due to their ease of deployment, high precision, low power consumption, etc. [1], [4]. For instance, bio-sensors are generally used to monitor vital signs such as electrocardiography (ECG), electromyography (EMG), blood pressure, heart rate and temperature [5]. Illnesses such as seizures, hypertension, dysthymias, and asthma can be diagnosed and treated by physiological monitoring. Inclinometers and goniometers are other types of sensors that are used to measure upper/lower limbs kinematics [6]. Even though there are potential gains of a

remote monitoring system using wearable sensors, there are still challenges in terms of technological advancements to design wearable sensors that are easy to use and comfortable for the wearer [7], [11]. An effective spatial-temporal recognition methods to process multi-scaled and noise-mixed signals is another challenge that needs to be addressed.

Statistical learning methods have been widely used to solve activity recognition problems [12], [20]. Chavarriaga *et al.* [21] used a Naïve Bayes(NB) and a K-Nearest Neighbor(KNN) classifier to recognize seven motions, such as walking, running and jumping. However, they relied on hand-crafted features and could not find discriminative features to accurately distinguish different activities. The feature extraction methods such as symbolic representation [22], statistics of raw data [23] and transform coding [24] are widely applied in human activity recognition, but they are heuristic and require expert knowledge to design features [14].

In recent years, with the popularity of deep learning technology, it has also been introduced into the applications of human activity recognition. Distinguished from statistical machine learning methods, deep learning makes it more convenient to extract and classify complex data in the face of a large number of different sensor sources. For example, Convolutional Neural Network (CNN) [9] can automatically extract features, but also fully learn complex high-dimensional nonlinear ones [10], [14]. A few researchers have already done some studies on the applications of deep learning for human activity recognition using wearable sensors. Most of these existing research have only used deep learning as a black box, and the data has been scratched. For example, Ronao and Cho [22] and Yang *et al.* [25] perform feature extraction of the sliding window using only shallow convolutional neural networks. Ordóñez and Roggen [19] use LSTM on this basis, adding timing considerations for human gesture recognition. Yang *et al.* [25] used deep convolutional neural networks to automatically learn features from the original inputs. Through the deep structure, the learning features are considered as higher-level abstract representations of low-level raw time-series signals.

Besides, other challenges also exist in HAR problems, such as large variability of a given action, similarity between classes, time consumption, and the high proportion of Null class [21]. Above mentioned deep convolutional neural networks ignored the temporal dependencies on the features, and was not suitable to recognize real-time sensor signals. Applying the time dependence to the features obtained from the original sensors is a key factor for the success of sequential human activity recognition. Ordóñez and Roggen [19] proposed deep convolutional network with utilizing of CNN and LSTM. This paper took advantage of LSTM to solve sequential human activity recognition problem and achieved a good precision. Hammerla *et al.* [2] rigorously explore deep, convolutional, and recurrent approaches across three representative datasets that contain movement data captured with wearable sensors. Chen *et al.* [14] posit that feature embedding from deep neural networks may convey complementary information and propose a distilling strategy to improve its performance, with handcrafted features utilized to assist a deep long short-term memory (LSTM) network. However, these complex network framework suffered from low efficiency and can hardly meet real-time requirements in practice applications. All of these challenges have led researchers to develop representation methods of systematic features and efficient recognition methods to effectively solve these problems.

Above works use CNN and LSTM to extract the waveform data of human activity. Compared with the classical feature engineering/machine learning method, the existing public datasets have greatly improved the power of deep learning. In view of above mentioned problems, this paper proposes a multi-level neural network structure model based on the combination of Inception Neural Network and GRU. The main contributions are as follows:

- With the use of different scales based convolution kernels, such as 1x1, 1x3, and 1x5, feature extraction and splicing of waveform data are performed to realize multi-scaled human body feature extraction for different durations.
- Pooling layers are applied to filter the interference noise brought by the unconscious jitter of human body, in order to decrease the misjudgment;
- With the concatenation of different scales based convolution layers, pooling layers, and multiple nonlinear activation, above mentioned advantages are superimposed and enlarged, so that the high-dimensional features of a single human activity are more easily extracted, which greatly reduces the interference and misidentification caused by no obvious segmentation.
- By comparison experiments on three most widely used public datasets, it is proved that the blindly superposition of network layers and the accumulation of parameters, in addition to multiplying the training and verification time, reduces the efficiency of learning and prediction. It cannot improve the accuracy of recognition, but lead to a decline in prediction results, perhaps because the network is too deep and insufficient data volume makes the structure difficult to be fully trained.

The rest of this paper is organized as follows: In Section II, we provide a primer on the relevant background in deep learning for human activity recognition. A detailed description is presented in Section III, illustrating the structure of our proposed InnoHAR model. Experiment setup are demonstrated in Section IV, and the results and analysis are discussed in Section IV, which show the superiorities of our proposed model. Then, conclusions come in Section VI.

II. RELATED WORK

Continuous human action recognition is challenging issue in machine learning with some difficulties to determine the parameters and sizing required because highly depended

upon some issues like feature selection in continuously training data streaming, the typical of classifier methods and we have less prior knowledge to determine the final size of training data, and the size of machine learning architectures. In human action data features, we have to deal with variance problem as explained in the survey paper [13], [15], [25]. In a typical classifier approach, it makes uneasy requirements and conditions for any machine learning methods [24], such as neural networks [25], dynamic Bayesian networks [7], extreme learning machine [27], Deep Learning [20] and many others that may not give good generalization accuracy and processing speed for all human action recognition cases. Here, we summarizes state-of-the-arts for e-health monitoring and other proposals for human activity recognition.

A. CONVOLUTIONAL NEURAL NETWORK

Each neuron of Convolutional Neural Network (CNN) [16] is connected to the local acceptor domain of its previous layer. It functions like a filter and is then activated by a nonlinear function, which is formulated as follows:

$$a_{i,j} = f(\sum_{m=1}^{H}\sum_{n=1}^{K} w_{m,n} \cdot x_{i+m,j+n} + b) \tag{1}$$

where, $a_{i,j}$ is the corresponding activation, f is a nonlinear function, $w_{m,n}$ is the $H \times K$ weight matrix of the convolution kernel, b is the offset value, and $x_{i+m,j+n}$ indicates the activation of the upper neurons connected to the neuron (i, j). CNN with several convolutional layers can learn hierarchical representations of data, and deeper convolutional layers characterize data in a more abstract way.

The input of neural network is generally the original signal, however, applying features extracted from the original signal to the neural network tends to improve performance. Extracting more useful features from the original signal requires sufficient expert knowledge, which inevitably limits a systematic exploration of the feature space [17]. Convolutional neural networks have been proposed to address this problem. Generally, CNN can be considered to comprise two parts. The first part is the hierarchical feature extractor, which contains convolutional layers and max-pooling layers. The input of each layer is the output of its previous layer. As a result, the original signal is mapped into feature vectors. The second part is a fully-connected layer, and the feature vectors are classified by the fully-connected layer.

The most widely used deep learning approach in the ubiquitous computing field in general and in human activity recognition using wearables in particular employ CNNs. CNNs typically contain multiple hidden layers that implement convolutional filters that extract abstract representations of input data. Combined with pooling and/or subsampling layers, and fully connected special layers, CNNs are able to learn hierarchical data representations and classifiers that lead to extremely effective analysis systems. A multitude of applications are based on CNNs, including but not limited to [25]–[28]. Recently, sophisticated model optimization techniques have been introduced that actually allow for the implementation of deep CNNs in resource constrained scenarios, most prominently for real-time sensor data analysis on smartphones and even smart watches [29].

B. LONG SHORT-TERM MEMORY

Long short-term memory (LSTM) inputs are sent to different gates, including input gates, output gates, and forgetting gates, representing the long-term, short-term, and near-term memory and control of the information. Each LSTM unit activation is calculated by the following formula:

$$a_t = \sigma(w_{i,h} \cdot x_t + w_{h,h} \cdot a_{t-1} + b) \tag{2}$$

where a_t and a_{t-1} represent respectively the activation at time t and $t-1$, σ is a nonlinear function, $w_{i,h}$ is a connection matrix between the input layer and the hidden layer, and $w_{h,h}$ is a connection matrix to which the hidden layer node is connected, and b is the offset value.

The defacto standard workflow for activity recognition in ubiquitous and wearable computing [30] treats individual frames of sensor data as statistically independent, that is, isolated portions of data are converted into feature vectors that are then presented to a classifier without further temporal context. However, ignoring the temporal context beyond frame boundaries during modeling may limit the recognition performance for more challenging tasks. Instead, approaches that specifically incorporate temporal dependencies of sensor data streams seem more appropriate for human activity recognition. In response to this, recurrent deep learning methods have now gained popularity in the field. Most prominently models based on so-called LSTM units [31] have been used very successfully. In [19], deep recurrent neural networks have been used for activity recognition on the Opportunity benchmark dataset. The LSTM model was combined with a number of preceding CNN layers in a deep network that learned rich, abstract sensor representations and very effectively could cope with the non-trivial recognition task. Through large scale experimentation in [32] appropriate training procedures have been analyzed for a number of deep learning approaches to HAR including deep LSTM networks. In all of previous works, single LSTM models have been used and standard training procedures have been employed for parameter estimation. The majority of existing methods [19], [30]–[32] are based on (variants of) sliding-window procedures for frame extraction. The focus of this paper is on capturing diversity of the data during training and to incorporate spatial-temporal information into proposed classifiers.

C. GOOGLENET AND INCEPTION MODULE

GoogLeNet based on the Inception module [9] is a new and innovative network structure proposed by Google in the second half of 2014, whose structure is not limited to the traditional sequential model [9]. As shown in Fig. 1, with the result of previous layer as an input, GoogLeNet's Inception module enters concatenation of, from left to right, a 1×1 convolution,

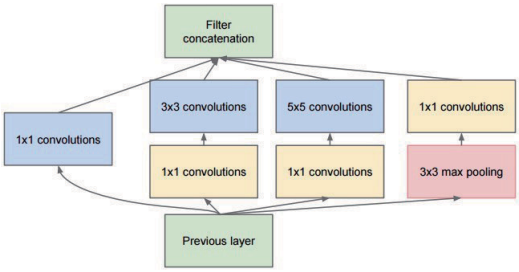

FIGURE 1. Inception module based GoogLeNet [9].

a 1 × 1 convolution tandem 3 × 3 convolution, a 1 × 1 convolution tandem a 5 × 5 convolution, and a maximum pooling layer of 3 × 3 tandem a 1 × 1 convolution. Then, the results calculated by above sub-module are spliced and input to the next layer.

Prior to this, the most direct way to improve the performance of deep neural networks was to enlarge depth by increasing the number of layers, and enlarge the width by increasing the number of nodes in each layer. This is the easiest and safest way to train high quality models [9], especially for a given large-scale tagged data set. However, this simple solution has two major drawbacks. One is that larger network sizes often mean more parameters, which makes the expanded network more prone to overfit, especially when the labeled samples in the training set are limited. The other one is the exponential increase of computing resource requirements caused by blindly increase of the network size.

Inception module brings a variety of proven network construction techniques. For example, a 1 × 1 convolution network implements dimensionality reduction of data and linear combination of multi-channel data. Extensive application of the pooling layer is useful to achieve dimensionality reduction, key feature extraction, and filtering.

III. THE PROPOSED ARCHITECTURE

The network structure of proposed InnoHAR deep neural network model is as shown in Fig. 2. The yellow block is indicated as the pooling layer, and the gray one is the GRU layer. Input data firstly passed through four Inception-like modules, according to the 1 dimensional time-series data features extracted using 1 × 1, 1 × 3, 1 × 5 convolution kernels, and pooling layer. The specific Inception-like network structure will be introduced in Section 3-A. At the same time, after passing by two Inception-like modules, we also connect it with max-pooling layer to help the network better eliminate misjudgment caused by noise disturbance. Finally, the output is passed through two GRU layers, so that the model can better extract the sequential temporal dependencies, as shown in Section 3-B.

A. SPATIAL FEATURES EXTRACTION

In the aspect of feature extraction from sensor waveform data, we use GoogLeNet's Inception module for reference, to implement Inception on three most widely used datasets, as shown in Fig. 3. In each Inception-like module, we also use a 1 × 1 convolution kernel to directly activate the combination of multi-channel information and pass it to the next layer. Two convolution kernels of 1 × 3 and 1 × 5 are cascaded respectively by a 1 × 1 convolution kernel, and the feature information of different scales is extracted for the whole model. The output splicing with only 1x1 convolution, also produces a ResNet residual connection effect. At the same time, there is a 1 × 3 pooling layer followed by a 1 × 1 convolution kernel to provide feature enhancement and filtering. These Inception-like modules use ELU as a nonlinear activation function. These substructures are then stitched together and again passed through the nonlinear activation of ELU and output to the next layer.

B. TEMPORAL FEATURES EXTRACTION

In the extraction of temporal features, we refer to the work of Ordóñez and Roggen [19], and also select two layers of LSTM layers for the extraction of temporal features, which is convenient for comparative analysis of later experiments. Experience has shown that vanilla RNN has a problem of gradient disappearance [18], while in many existing experiments, GRU and LSTM show better performance in dealing with long sequences based problems, while GRU performs better in terms of time efficiency. Based on the relatively complex network structure displayed in Section III-A, we use GRU as the concrete implementation of the loop layer. The entire network structure delivers satisfactory results in both predictive performance and time efficiency.

C. PREPROCESSING

In order to minimize the preprocessing work in the early stage, the end-to-end human activity recognition model is realized while maintaining the same data as the predecessors. Taking Opportunity dataset [21] as an example, as shown in Fig. 4 we used all the 113 channels of sensor data from the human body, and used the same fixed-length 24-line sliding window as that of used by Ordóñez and Roggen [19], sliding 12 lines at a time. A total of 9,984 pieces of data were obtained for testing. We fill the missing values of the sensor by linear fitting, and each sensor channel is normalized to the [0,1] interval. This is a totally 18 classes classification problem, containing NULL classes. Similar procedures are also applied to the other two datasets.

D. MODEL IMPLEMENTATION

We use Keras 2 to build our network structure. Keras is a high-level neural network API written in Python with optional Tensorflow or Theano as the backend. We chose Tensorflow as the backend in the experiment and run it on the GPU. The hardware environment is introduced as follows in Table 1.

IV. EXPERIMENT SETUP

In this paper, we conduct experiments on three benchmark datasets representative of the problems typical for HAR

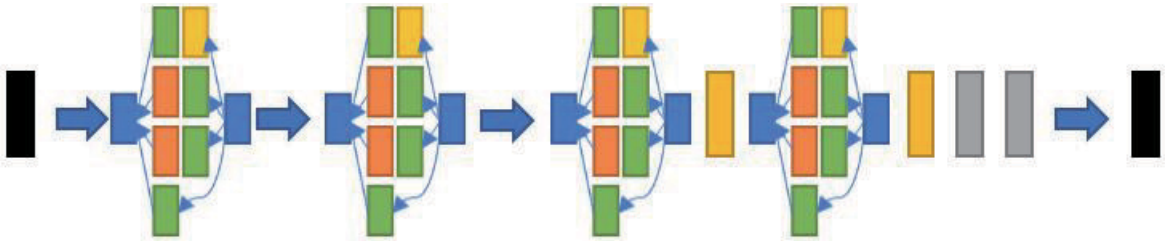

FIGURE 2. Inception Neural Networks for Human Activity Recognition Model Architecture. The yellow block is indicated as the pooling layer, and the gray one is the GRU layer. The other parts are summarized as convolution layers in Inception-like module, which is described in Fig. 3.

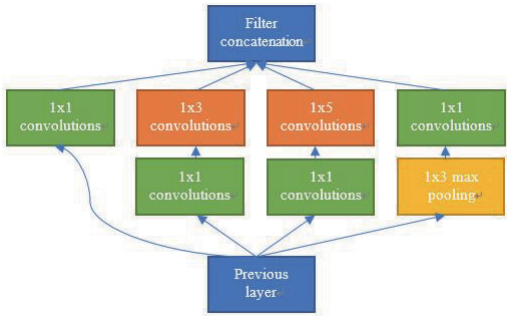

FIGURE 3. Inception-like module used in our proposed InnoHAR network.

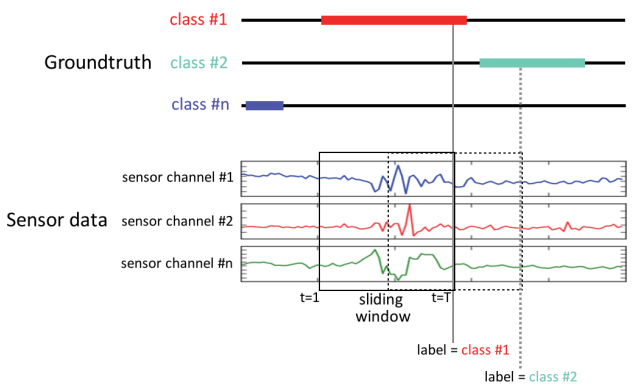

FIGURE 4. Preprocess sensor waveform data using a sliding window.

TABLE 1. Hardware parameters.

CPU Type	Intel Xeon E3-1505M v5
CPU Main Frequency	2.8GHz
Kernels/Threads Number	4/8
RAM Capacity	16GB(8GB×2)
Graphic Card Chip	NVIDIA Quadro M5000M
Graphic Memory Capacity	8GB
Clock Frequency	1050MHz
Cuda kernels	1536

(described below) to train and test our model. They are composed of a set of complex human natural activities collected in an environment where rich sensors are installed [21], [33].

A. OPPORTUNITY DATASET

Opportunity activity recognition dataset [33] is of complex naturalistic activities with a particularly large number of atomic activities (more than 27,000) collected in a sensor rich environment. Overall, it comprises recordings of 12 subjects using 15 networked sensor systems, with 72 sensors of 10 modalities, integrated in the environment, in objects, and on the body. These characteristics make it well suited to benchmark various activity recognition approaches.

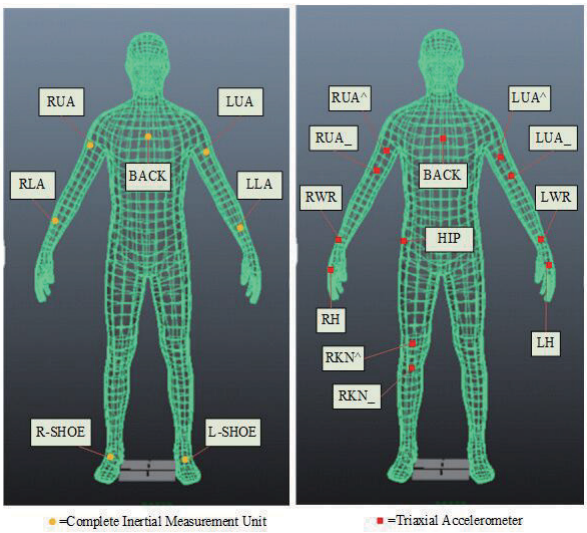

FIGURE 5. Position of on-body sensors used in the OPPORTUNITY dataset (left: IMU sensors; right: 3-axis accelerometers) [21].

We only consider the on-body sensors, including inertial measurement units and 3-axis accelerometers. The wearing position of the sensors are shown in Fig. 5. Each sensor channel is treated as an individual channel, a total of 113 channels. The sampling frequency of these sensors is 30Hz. OPPORTUNITY dataset contains several gestures and postures, and we mainly realize the recognition of gestures either including or ignoring the Null class. This is an 18-class classification problem, the gestures in the dataset are summarized in Table 2.

B. PAMAP2 DATASET

It consists of recordings from 9 participants (8 males and 1 female) instructed to carry out 18 lifestyle activities,

TABLE 2. Class labels for mode of Opportunity dataset.

Open Dishwasher	Close Dishwasher
Open Fridge	Close Fridge
Open Drawer 1	Close Drawer 1
Open Drawer 2	Close Drawer 2
Open Drawer 3	Close Drawer 3
Open Door 1	Close Door 1
Open Door 2	Close Door 2
Drink from Cup	Clean Table
Toggle Switch	Null

TABLE 3. Class labels for mode of PAMAP2 dataset.

lie	sit	stand
walk	run	cycle
Nordic walk	iron	vacuum clean
rope jump	ascend stairs	descend stairs
watch TV	computer work	drive car
fold laundry	clean house	play soccer

including household activities (lie, sit, stand, walk, run, cycle, Nordic walk, iron, vacuum clean, rope jump, ascend and descend stairs) and a variety of leisure activities (watch TV, computer work, drive car, fold laundry, clean house, play soccer) [37], as summarized in Table 3. Accelerometer, gyroscope, magnetometer, temperature and heart rate data are recorded from inertial measurement units located on the hand, chest and ankle over 10 hours (in total). The resulting dataset has 52 dimensions. We used runs 1 and 2 for subject 5 in our validation set and runs 1 and 2 for subject 6 in our test set. The remaining data is used for training. In our analysis, we downsampled the accelerometer data to 33.3Hz in order to have a temporal resolution comparable to the Opportunity dataset. For frame-by-frame analysis, we replicate previous work with non-overlapping sliding windows of 5.12 seconds duration with one second stepping between adjacent windows (78% overlap) [37]. The training-set contains approx. 473k samples (14k frames).

C. SMARTPHONE DATASET

Smartphone database is built from the recordings of 30 subjects performing activities of daily living (ADL) while carrying a waist-mounted smartphone with embedded inertial sensors [38].

The experiments have been carried out with a group of 30 volunteers within an age bracket of 19-48 years. Each person performed six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) wearing a smartphone (Samsung Galaxy S II) on the waist. Using its embedded accelerometer and gyroscope, we captured 3-axial linear acceleration and 3-axial angular velocity at a constant rate of 50Hz. The experiments have been video-recorded to label the data manually. The obtained dataset has been randomly partitioned into two sets, where 70% of the volunteers was selected for generating the training data and 30% the test data.

The sensor signals (accelerometer and gyroscope) were pre-processed by applying noise filters and then sampled in fixed-width sliding windows of 2.56 sec and 50% overlap (128 readings/window). The sensor acceleration signal, which has gravitational and body motion components, was separated using a Butterworth low-pass filter into body acceleration and gravity. The gravitational force is assumed to have only low frequency components, therefore a filter with 0.3 Hz cutoff frequency was used. From each window, a vector of features was obtained by calculating variables from the time and frequency domain, using proposed and comparative methods.

D. PERFORMANCE MEASURE

Human activity datasets collected in natural scenes are often imbalanced between classes [26]. Some classes may contain a large number of samples while other classes have only a few samples. The gestures of OPPORTUNITY dataset are extremely imbalanced, the Null class accounts for more than 70% of all the data. The classifier predicts the classification accuracy of each class, Null class can achieve very high accuracy. The overall classification accuracy is not an appropriate index for performance evaluation. F-measure (F_1) considers the correct classification of each class as equally important. It takes into account both the precision and the recall of each class to compute the score and can evaluate the model better than the precision. Precision is defined as $P = \frac{TP}{TP+RP}$, and recall corresponds to $R = \frac{TP}{TP+PN}$, where TP and FP are the number of true and false positives, respectively, and FN corresponds to the number of false negatives. Class imbalance is countered by weighting classes according to their sample proportion:

$$F_1 = \sum_i 2 * w_i \frac{precision_i \cdot recall_i}{precision_i + recall_i} \tag{3}$$

where $w_i = n_i/N$ is the proportion of samples of the ith class, with n_i being the number of samples of the ith class and N being the total number of samples.

V. RESULTS AND ANALYSIS

With the consideration of human activity recognition applications, we apply our proposed InnoHAR model on above-mentioned three public dataset for verification. We compared our results with both baseline classifiers and state-of-the-arts deep networks. All proceeded results are

TABLE 4. Best results (F-measure) obtained for each model and dataset, along with some baselines for comparison.

Methods	Oppotunity dataset	PAMAP2 dataset	Smartphones dataset
CNN [25]	0.851	0.88	0.915
DeepConvLSTM [19]	0.915	0.905	0.91
Proposed method (InnoHAR)	**0.946**	**0.935**	**0.945**

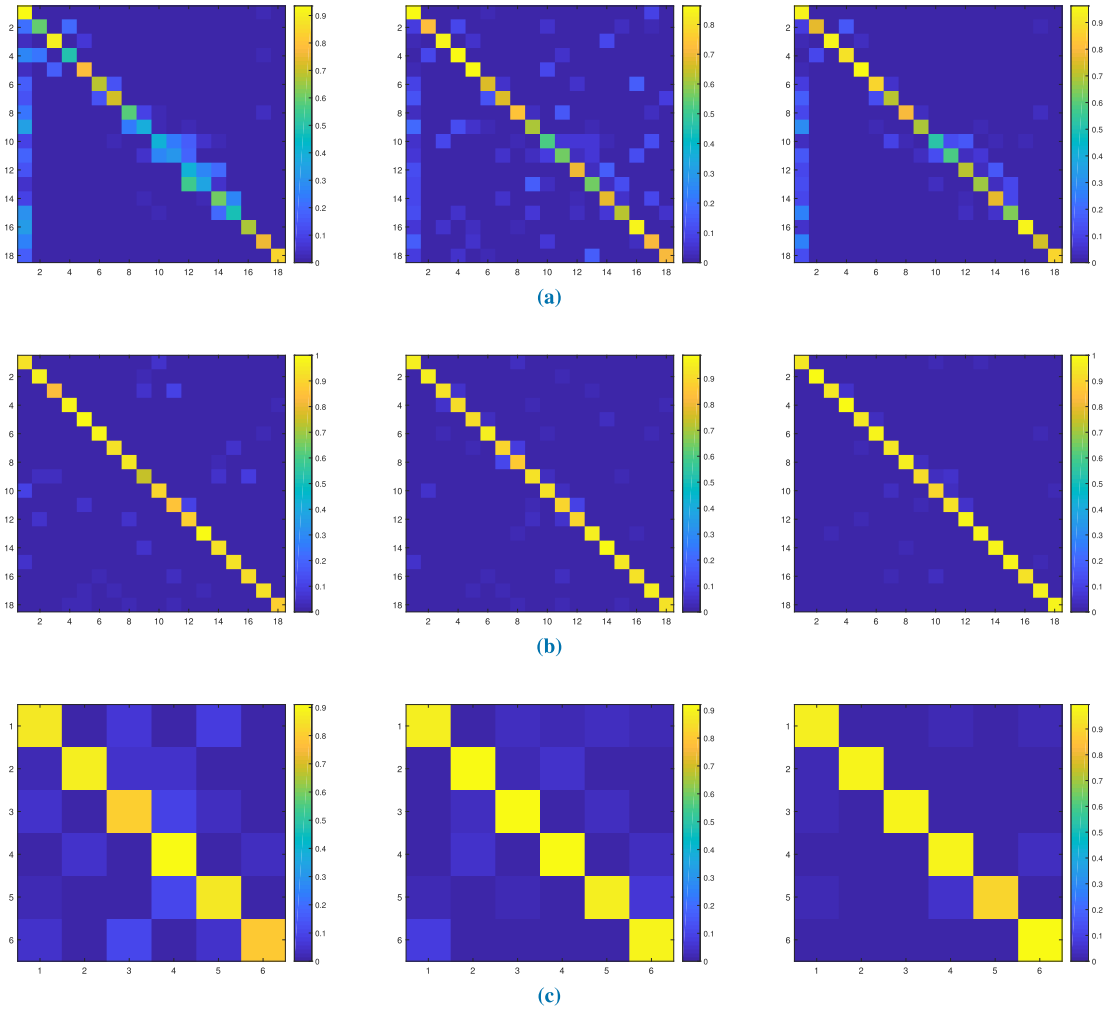

FIGURE 6. Confusion matrix of proposed method and two state-of-the-arts in one test run. The most left one in (a),(b),(c) is the confusion matrix of CNN [25], the middle one is that of DeepConvLSTM [19] and the most right one is that of proposed method.

verified by F1-score means to ensure the fairness and consistency of the following comparison results. The good performance of recurrent approaches, which model movement at the sample level, holds the potential for novel (real-time) applications in HAR, as they alleviate the need for segmentation of the time-series data.

A. COMPARED WITH BASELINES AND STATE-OF-THE-ARTS

In recent years, we have also seen some methods of deep neural network used in human activity recognition [7], [16], [19], [23], [25]. Others have tried to use deep neural network to re-attack Opportunity data sets and Opportunity Challenge, such as Yang *et al.* [25] and Ordóñez and Roggen [19]. DeepConvLSTM is the previous state-of-the-arts model on Opportunity dataset. We test and evaluate the recognition accuracy of Yang *et al.* [25] CNN, DeepConvLSTM and our proposed InnoHAR human activity recognition network model under the same experiment scenario.

Furthermore, we generalize and verify these three models to three most widely used public datasets, namely Opportunity dataset [21], PAMAP2 dataset [37] and smartphone dataset [38]. Table 4 shows the evaluation results of the above various deep neural models.

In terms of overall performance, the experimental results show that our proposed InnoHAR model based on Inception-like module has better performance than both CNN [25] and DeepConvLSTM [19]. For Opportunit dataset, our proposed model has a significant increase of about 9% compared with the CNN model of Yang *et al.* [25], as well as a 3% improvement compared with DeepConvLSTM. For PAMAP2 dataset, we can also observe nearly a 5% gap between the best (proposed model) and worst (CNN [25]), and InnoHAR also has a 3% performance advantage over DeepConvLSTM. For smartphones dataset, InnoHAR is also superior to the other two methods and maintains good generalization performance.

To be more specific, Fig. 6 shows the confusion matrix of proposed method and two state-of-the-arts in one test run, from which we can see that proposed method maintains consistent superior performance on different public datasets and has good generalization performance in the recognition of complex human activities. It could be because of the multi-scaled and spatial-temporal feature extraction characteristics, due to the combination of inception and GRU.

B. COMPARED WITH IMPROVED STATE-OF-THE-ARTS

At the same time, we also tested the "simple and safer" traditional method in GoingDeeper's [9] to improve the performance of proposed model. We modified DeepConvLSTM with adding a layer of CNN with the same kernel size as the original ones; in DeepConvLSTM with bi-LSTM, we replace the original unidirectional LSTM layers with two bi-LSTM with positive and negative bidirection, and these two bi-LSTMs are spliced into the next layer.

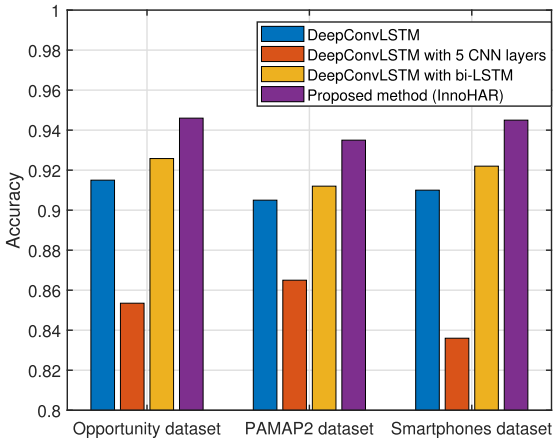

FIGURE 7. Performance comparison with modified state-of-the-arts on typical public datasets.

In Fig. 7, we can see that when an additional CNN layer is added to simply increase the depth of model, the recognition accuracy model is not improved as expected, but has a certain degree of decline. We speculate that because a large CNN layer contains a huge number of parameters, and the CNN layer is connected in a fully connection manner. Each additional layer may cause an exponential rise in model parameters. In case that the amount of data is not large enough, the gradient change of the last layer cannot be fully transmitted to the previous level by Back-Propagation (BP) algorithm, which leads to over-fitting of the whole model, resulting in decreasing the recognition accuracy.

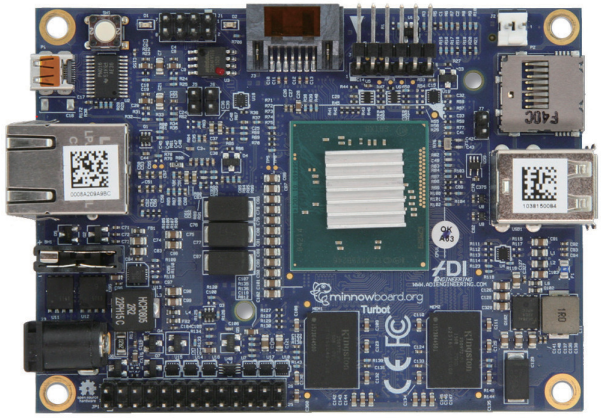

FIGURE 8. MinnowBoard Turbot Dual Core Board [27].

In contrast, when DeepConvLSTM was transformed into a bidirectional LSTM, the accuracy did increase slightly. However, LSTM was used as a network layer for timing modeling. The calculation of the latter node must wait until that of the previous node completes. It is difficult to parallelize the calculation by GPU. Therefore, in actual training process, a large part of time is often occupied in trainging process, but the income is relatively little.

C. EFFICIENCY ANALYSIS

Considering running time and efficiency, we did not simply count the running time and present the result as usual. We believe that human activity recognition should be proceeded in following steps: 1) train a model with collected sensor data; 2) import this model into an embedded system; 3) the embeded system reads real-time data and run pre-trained model, working out the prediction. Therefore, regardless of the recognition accuracy, we must consider the actual operation of the model in real-time systems, which can not exceed the limit of computing resources consumption, affecting the real-time prediction.

The embedded system we used in this experiment was the MinnowBoard Turbot Dual Core Board [27]. It is a small embedded platform released by Intel that is equipped with Intel Atom E3826 processor, with 1.46 GHz clock speed and 2GB DDR3L 1067MT/s DRAM. It is an outstanding embedded platform with both performance and price advantages.

We run Ubuntu 14.04 operating system on it, testing our proposed model with the same framework and package dependencies. We import the model json file and weight file into the system. We start timing after the model is loaded and begins to predict. In the prediction of all previous test

data, our model took 152.02 s and completed the prediction of all 9894 sliding windows. The predicted speed reached 65.09 pieces/s. The original test data is recorded at 30 Hz, and our model can easily predict the real-time activity on this platform.

VI. CONCLUSION

In this paper, we conceptually proposed an InnoHAR model for wearable sensor based human activity recognition applications by concatenating convolution kernels of different scales and splicing with max-pooling layers. Compared with baselines and state-of-the-arts, our proposed method shows consistent superior performance and has good generalization performance on three most widely used public datasets. In the experiment, we also proved that our innovative structure has more potential in realtime applications by practice test on MinnowBoard Turbot Dual Core Board.

For our future work direction, we will first continue to adjust our network structure, including the size of kernels and the connection method. Besides, we may explore further the problem of data imbalance in real-life human activity recognition applications.

REFERENCES

[1] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Inf. Fusion*, vol. 35, pp. 68–80, May 2017.

[2] N. Y. Hammerla, S. Halloran, and T. Ploetz. (2016). "Deep, convolutional, and recurrent models for human activity recognition using wearables." [Online]. Available: https://arxiv.org/abs/1604.08880

[3] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, "Enabling effective programming and flexible management of efficient body sensor network applications," *IEEE Trans. Human-Mach. Syst.*, vol. 43, no. 1, pp. 115–133, Jan. 2013.

[4] C. Xu, J. He, X. Zhang, C. Yao, and P.-H. Tseng, "Geometrical kinematic modeling on human motion using method of multi-sensor fusion," *Inf. Fusion*, vol. 41, pp. 243–254, May 2017.

[5] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *Proc. 23rd Int. Conf. Archit. Comput. Syst. (ARCS)*, Feb. 2010, pp. 1–10.

[6] J. Margarito, R. Helaoui, A. M. Bianchi, F. Sartor, and A. G. Bonomi, "User-independent recognition of sports activities from a single wrist-worn accelerometer: A template-matching-based approach," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 788–796, Apr. 2016.

[7] P. C. Roy, S. Giroux, and B. Bouchard, "A possibilistic approach for activity recognition in smart homes for cognitive assistance to Alzheimer's patients," in *Activity Recognition in Pervasive Intelligent Environments*. Paris, France: Atlantis Press, 2011, pp. 33–58.

[8] S. Galzarano, R. Giannantonio, A. Liotta, and G. Fortino, "A task-oriented framework for networked wearable computing," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 2, pp. 621–638, Apr. 2016.

[9] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[10] C. Zhenghua, J. Chaoyang, and X. Lihua, "A novel ensemble ELM for human activity recognition using smartphone sensors," *IEEE Trans. Ind. Informat.*, to be published.

[11] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," *Inf. Fusion*, vol. 22, pp. 50–70, Mar. 2015.

[12] Z. Wang, D. Wu, R. Gravina, G. Fortino, Y. Jiang, and K. Tang, "Kernel fusion based extreme learning machine for cross-location activity recognition," *Inf. Fusion*, vol. 37, pp. 1–9, Sep. 2017.

[13] A. Bux, P. Angelov, and Z. Habib, "Vision based human activity recognition: A review," *Advances in Computational Intelligence Systems*. Cham, Switzerland: Springer, 2017, pp. 341–371.

[14] Z. Chen, L. Zhang, Z. Cao, and J. Guo, "Distilling the knowledge from handcrafted features for human activity recognition," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4334–4342, Oct. 2018.

[15] S.-R. Ke, H. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[17] D. Figo, P. C. Diniz, D. R. Ferreira, and J. M. Cardoso, "Preprocessing techniques for context recognition from accelerometer data," *Pers. Ubiquitous Comput.*, vol. 14, no. 7, pp. 645–662, 2010.

[18] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, Aug. 2003.

[19] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[20] W. Liu, J. Yang, L. Wang, C. Wu, and R. Zhang, "Movement behavior recognition based on statistical mobility sensing," *Adhoc Sensor Wireless Netw.*, vol. 25, nos. 3–4, pp. 323–340, 2015.

[21] R. Chavarriaga *et al.*, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 2033–2042, 2013.

[22] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Syst. Appl.*, vol. 59, pp. 235–244, Oct. 2016.

[23] M. Zeng *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. MobiCASE*, Nov. 2014, pp. 197–205.

[24] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 3, pp. 1192–1209, 3rd Quart., 2013.

[25] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 3995–4001.

[26] C. A. Ronaoo and S.-B. Cho, "Evaluation of deep convolutional neural network architectures for human activity recognition with smartphone sensors," in *Proc. KIISE Korea Comput. Congr.*, 2015, pp. 858–860.

[27] J. B. 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. IJCAI*, 2015, pp. 1–7.

[28] N. M. Rad, A. Bizzego, S. M. Kia, G. Jurman, P. Venuti, and C. Furlanello. (2015). "Convolutional neural network for stereotypical motor movement detection in autism." [Online]. Available: https://arxiv.org/abs/1511.01865

[29] S. Bhattacharya and N. D. Lane, "Sparsification and separation of deep learning layers for constrained resource inference on wearables," in *Proc. SenSys*, 2016, pp. 176–189.

[30] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, p. 33, 2014.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] N. Y. Hammerla, J. M. Fisher, P. Andras, L. Rochester, R. Walker, and T. Plotz, "PD disease state assessment in naturalistic environments using deep learning," in *Proc. AAAI*, 2015, pp. 1–7.

[33] D. Roggen *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *Proc. 7th Int. Conf. Netw. Sens. Syst. (INSS)*, Jun. 2010, pp. 233–240.

[34] C. X. Ling and V. S. Sheng, "Class imbalance problem," in *Encyclopedia of Machine Learning*. New York, NY, USA: Springer, 2011.

[35] *MinnowBoard*. Accessed: Jan. 11, 2019. [Online]. Available: https://www.minnowboard.org/

[36] M. Bachlin *et al.*, "Potentials of enhanced context awareness in wearable assistants for Parkinson's disease patients with the freezing of gait syndrome," in *Proc. Int. Symp. Wearable Comput. (ISWC)*, Sep. 2009, pp. 123–130.

[37] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," in *Proc. 16th Int. Symp. Wearable Comput. (ISWC)*, Jun. 2012, pp. 108–109.

[38] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proc. ESANN*, 2013, pp. 1–6.

CHENG XU received the B.E. and M.S. degrees from the University of Science and Technology Beijing, China, in 2012 and 2015, respectively, where he is currently pursuing the Ph.D. degree with the Data and Cyber-Physical System Lab. His research interests include patter recognition and the Internet of Things. He is a Student Member of the IEEE and CCF.

DUO CHAI received the degree from the University of Science and Technology Beijing. He was a Research Assistant with the Data and Cyber-Physical System Lab, University of Science and Technology Beijing. He was a Machine Learning Engineer with the 360 Search Lab, focusing on natural language processing. His research interests include patter recognition, machine learning, and deep learning, especially in natural language processing.

JIE HE received the B.E. and Ph.D. degrees in computer science from the University of Science and Technology Beijing (USTB), China, in 2005 and 2012, respectively. From 2011 to 2012, he was a Visiting Ph.D. Student with the Center for Wireless Information Network Studies, Worcester Polytechnic Institute. Since 2015, he has been an Associate Professor with the School of Computer and Communication Engineering, USTB. His research interests include wireless indoor positioning, human gesture recognition, and motion capture.

XIAOTONG ZHANG received the M.S. and Ph.D. degrees from the University of Science and Technology Beijing, in 1997 and 2000, respectively. He was a Professor with the Department of Computer Science and Technology, University of Science and Technology Beijing. His research interests include quality of wireless channels and networks, wireless sensor networks, networks management, cross-layer design and resource allocation of broadband and wireless networks, signal processing of communication, and computer architecture.

SHIHONG DUAN received the Ph.D. degree in computer science from the University of Science and Technology Beijing, where she is currently an Associate Professor with the School of Computer and Communication Engineering. Her research interests include wireless indoor positioning, human gesture recognition, and motion capture.