Journal of Computers Vol. 32 No. 6, 2021, pp. 24-37
doi:10.53106/199115992021123206003

Human Activity Recognition with Multimodal Sensing of Wearable Sensors

Chun-Mei Ma[1,2,3], Hui Zhao[4*], Ying Li[1], Pan-Pan Wu[1],
Tao Zhang[1], Bo-Jue Wang[1]

[1] School of Computer and Information Engineering, Tianjin Normal University, Tianjin 300387, China
mcmxhd@163.com,{liying,wupanpan}@tjnu.edu.cn,{taozhanggo,dukewbimcm}@163.com

[2] Postdoctoral Innovation Practice Base, Zhuhai Huafa Properties Co., Zhuhai 519030, Guangdong, China

[3] College of Computer Science, Nankai University, Tianjin 300350, China

[4] School of Cyberspace Security, Dongguan University of Technology, Dongguan 523106, Guangdong, China
zhaohui841010@sina.com

Received 5 January 2021; Revised 16 July 2021; Accepted 15 August 2021

Abstract. Human activity sensed by wearable sensors has multi-granularity data characteristics. Although deep learning-based approaches have greatly improved the accuracy of recognition, most of them mainly focus on designing new models to obtain deeper features, ignoring the different effects of different deep features on the accuracy of recognition. We think that discriminative features learning would improve the recognition performance. In this paper, we propose an end-to-end model ABLSTM that consists of Attention model and BLSTM model to recognize human activities. Specifically, the BLSTM model is used to extract deep features of various activities. After that, the Attention model is used to obtain the discriminative features representation by reducing the irrelevant features and enhancing the positive correlation features to each activity. Therefore, compared with traditional deep learning-based approaches, such as CNN and RNN based etc., the features learned by ABLSTM are more discriminative, which can be in response to the changes of activities. By testing our model on two public benchmark datasets: UCI and Opportunity. The results show that our model can well recognize human activities with F1 scores as high as 99.0% and 92.7% respectively on the two datasets, which pushes the state-of-the-art in human activities recognition of mobile sensing.

Keywords: human activity recognition, multimodal sensory data, discriminative features representation, wearable sensors

1 Introduction

Human activity recognition (HAR) refers to analyze and understand various human movements by computer automatic detection technology based on the perceptual data. This technology has a wide range of application scenarios, such as intelligent monitoring, human-computer interaction, robots, etc. In recent years, with the popularity of smartphones and wearable devices which are built-in multiple sensors, contact human activity recognition is on the rise, which can be directly related to our daily lives, such as medical health monitoring or fitness monitoring, etc. Therefore, human activity recognition for wearable sensors has become a research hotspot in recent years.

The perceptual data collected by various wearable sensors is not only multimodal, but also time series, which reflects the movements of carriers. Hence, human activity recognition based on wearable sensors is generally considered to be a classification problem of heterogeneous time series data [1-2]. For this

[*] Corresponding Author

problem, in the last few years, some researchers propose to utilize data fusion approaches [3-5]. However, the relations of different modal data, such as turning (sensed by the gyroscope sensor) is accompanied by deceleration (sensed by the acceleration sensor) are ignored by these approaches or they never consider the time labels of the data. With the development of deep learning, some deep learning-based approaches are proposed for human activity recognition [6-7]. Although these approaches can obtain better features representation than sensor fusion approaches, most of them focus on designing new models, rather than analyzing the representation features of various activities. Usually, different activities are with different discriminative features. If all the features obtained by deep learning models are treated equally for activity classification, the similar features will degrade the classification performance.

In this paper, to selectively focus on the discriminative features of each activity and ignore irrelevant ones, we propose an end-to-end deep learning model ABLSTM. ABLSTM is a fusion model that extends of BLSTM [8] by attention model [9-10]. Specifically, in our ABLSTM model, the raw sensed data is segmented according to the size of the predefined sampling windows and fed to the BLSTM layer to get the deep features of activities. After that, attention model is used to reduce the irrelevant features and enhance the positive correlation features to obtain discriminative features. This step is very important for obtaining high precision of activity recognition. Finally, the output layer that consists of a fully connected layer and a classifier is used to predict the activity category of the segmented perceptual data. To verify the effectiveness of ABLSTM model, we conduct experiment on two public datasets: Public domain UCI dataset and Opportunity dataset. The experimental results show that our model achieves 99.0% and 92.7% F1 scores on the two datasets, which are higher than the advance approach Res-BLSTM 2.5% and 5.46%, respectively. This demonstrates the excellent recognition capability of our model and its potential for real-life applications.

The original contributions that we have made in this paper as follows:

(1) We provide an end-to-end model ABLSTM that is a deep learning framework composed of attention model and BLSTM. ABLSTM has strong ability to model the multimodal time series data and selectively extract discriminative features which can significantly improve the accuracy of activity recognition.

(2) We introduce the attention mechanism into BLSTM model to reduce the irrelevant features and enhance the positive correlation features, getting more discriminative features, which can optimize the traditional deep learning model BLSTM.

(3) We evaluate the proposed network ABLSTM with the real data multimodal sensory data, which are collected by sensors embedded into smartphones.

The remainder of this paper is structured as follows: Section 2 presents a brief overview of related works of HAR. Then the problem statement and the system architecture are described in Section 3. Section 4 gives a detailed implementation process of our model (ABLSTM). The performance of our model is verified and the experimental results are presented in Section 5. Finally, we conclude this paper in Section 6.

2 Related Work

Motivated by the challenges of features representation of multimodal data for HAR, many researchers have devoted to solve this problem and proposed numerous works, which are mainly divided into two categories: sensor fusion-based and deep learning-based.

Sensor fusion-based approach. Sensor fusion combines perceptual data derived from multi-sensors to obtain an invariant feature expression for HAR. Obviously, the use of multi sensors is superior to a single one as it has less uncertainty. Up to now, many of sensor fusion-based methods for recognizing activities of daily living have been proposed [11-14]. In [11], the authors proposed a codebook based perceptual data fusion method. Specifically, the perceptual data will be divided into clusters, and the cluster center is selected as the codeword. For different activities, they own different codewords. Then a support vector machine (SVM) is constructed to perform activity recognition. In [12], the authors proposed a fuzzy logic method to classify human activities. Specifically, the raw data from multi-sensor are together fed into the fuzzy logic model. Then through fuzzification, rule inference and defuzzification process to classify human activities. The disadvantage of these methods is that the temporal correlation feature is ignored, which will increase the misclassification rate. Thus, in [13], the authors used Hierarchical Hidden Markov models to recognize the beginning, estimate on-going activity, detect the end of activities, forming statistical features together with sensory data. Then different classifiers, such as SVM, decision tree etc. are applied for the activity recognition. In [14], the authors tackled the problem of recognizing smoking activity using a single 9-axis inertial sensor measurement unit (IMU) as accelerometer, gyroscope and compass and using the method of quaternions to fuse them. Although these methods derive comprehensive characteristics for activities, the correlation characteristics between the multimodal sensors are ignored, which is easy to lead to inadequate expression of various activities.

Deep learning-based approach. As deep learning network is more flexible in data fitting, it can learn the features of the data by itself. Thus, deep learning-based approaches for human activity recognition have been proposed one after another, especially with the superior performance of deep learning in image and speech recognition. Up to now, the deep learning-based approaches for the activity recognition are mainly divided into convolution neural network (CNN)-based approach [15-17] and recurrent neural network (RNN, LSTM etc.) based approach [18-20]. CNN based approach used the convolution kernel to capture the heterogeneous and continuous characteristics of the sensory data, and finally obtains the feature representation of the data through the deep network structure. Then, a decision tree or support vector machine or multi-classifier softmax is used to realize activity recognition. For example, in [15], the authors presented a feature learning method that deployed convolutional neural networks (CNN) to automate feature learning from the raw inputs in a systematic way. In [16] the authors proposed a two-stage end-to-end CNN method. In [17], the authors proposed a data fusion-based CNN approach that it fused the sensory data from diverse channels and fed the fused data into a CNN models to extract the features and perform classification. Although these methods can capture the correlation characteristics between multimodal sensors and the temporal correlation feature, they are limited by the size of the convolution kernel. Recurrent neural network-based approach used multimodal sequence data as input. Then an iterative model was used to obtain the feature representation of the sensory data and finally a multi-classifier softmax was used to perform classification. These approaches can describe the time characteristics of the multimodal data. For example, in [18] both of a unidirectional LSTM and a bidirectional LSTM were created and compared for skiing activity recognition. In [19-20], the authors present unidirectional, bidirectional and cascaded architectures based on LSTM deep recurrent neural networks for human activity recognition. Besides, in order to make full use of the advantages of CNN and recurrent neural networks, some of mix deep learning models are proposed [21-23]. A comprehensive survey of the deep learning-based human activity recognition can be accessed in [24]. Although these algorithms improve the accuracy of HAR, most of them focus on designing new models. In fact, not all features are equally important for the activity recognition, it is required to selectively pay attention to the discriminative features, and try to ignore the irrelevant ones. To the best of our knowledge, this is the first BLSTM architecture with explicit attention model as its fundamental capability for the multimodal time series data to recognize human activities.

3 Problem Statement and System Design

3.1 Problem Statement

For contact perception of human activity, multimodal time series sensory data is used to recognize these activities. However, even for different activities, some pieces of sensed data may present the same characteristics. Take the Y-gyroscope perceptual data of six different activities in UCI dataset [15] as an example, as shown in Fig. 1. We can observe that the positive related features of each activity are only shown at some point. In addition, some features in an activity may be also shown in others, such as walking downstairs and sitting. Therefore, it is very significant to learning the discriminative features for each activity.

3.2 System Design

In order to learn the discriminative features of each human activity from the multimodal time series perceptual data, a combination of BLSTM model and Attention model ABLSTM is proposed, which introduces the attention mechanism into the BLSTM network for the first time. It can be used not only for processing of the multimodal time series data, but also for discriminative features learning which response to the changes of human activities. The architecture of the ABLSTM model is shown in Fig. 2.

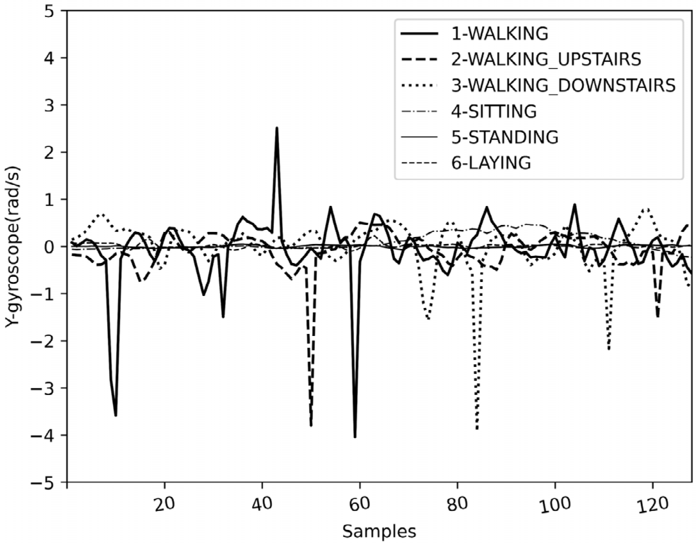

Fig. 1. The raw sensory data of various activities in UCI dataset [24]

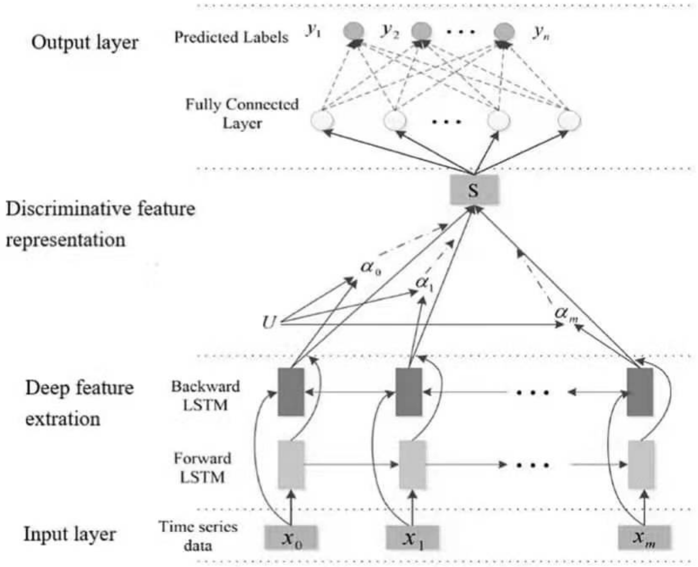

Fig. 2. The architecture of ABLSTM model. The whole architecture is divided into four parts. Deep feature extraction layer is composed by BLSTM. Discriminative feature representation layer consists of attention mechanism which can learn the important degree of each deep feature. The output layer is made up of a fully connected layer and a classifier to classify human activities

It is composed of four layers from bottom-up as: the input layer, the deep feature extraction layer, the discriminative feature representation layer and the output layer. Firstly, the multimodal time series data acquired by multi-sensors is as input to the ABLSTM model. Then the deep feature extraction layer is achieved by BLSTM, which connects the forward LSTM and the backward LSTM. The outputs of the hidden units of BLSTM are the deep features. After that, the discriminative feature representation layer that consists of the attention mechanism is used to reduce the irrelevant features and enhance the positive correlation features of each activity, obtaining more salient features for the activity recognition. Finally, the discriminative features are imported into the output layer which consists of a fully connected layer and a classifier. The fully connected layer is used to integrate the discriminative features of every multimodal time series data, which then fed into a classifier to predict the final recognition results.

4 ABLSTM Model

4.1 Perceptual Data Presentation for Input Layer

Usually, the data collected by different types of sensors have different granularity and quality. For example, acceleration data reflects the change of the speed of an object, while the gyroscope reflects the change of the motion direction of the object. Therefore, the sensory data of human activity collected from different types of sensors are multimodal. Besides, activities have a continuous character that is time related and the sensory data change with the activities. Thus, the sensory data generated by are multimodal and time labelled [25]. An effective sensory data representation is the first step in implementing our ABLSTM. In this paper, the sensory data is represented as: $\{X_i, y^{(i)}\}, \quad i = 1, 2, ..., N$, where $X_i = (x_1^i, x_2^i, ..., x_L^i)$ denotes the time sampling sequences labelled with $y^{(i)}$, L denotes the sampling window and N is the number of sampling windows. For the *t*-th sample sequence x_t^i, it contains *K* features that are the number of sensor categories. Thus, X_i not only reflects the time characteristic of the sensory data but also reflects its multimodal property. It will serve as the input data to our ABLSTM.

4.2 Deep Feature Extraction

For the raw input data of each activity, it contains more detailed information. In order to improve the HAR accuracy, it is significant to extract their deep features, which contain more semantic information. In this paper, we design BLSTM model based deep feature extraction layer, which consists of a forward LSTM and a back LSTM [26]. Thus, BLSTM can not only extract the deep features of the input data by the hidden units, but also well capture the time characteristic of the input data. The structure of BLSTM is illustrated in Fig. 3. For the input data x_t^i, it will be fed to the forward LSTM and back LSTM simultaneously. After that, x_t^i together with h_{t-1}^i that is the output of the previous moment will be fed into the forget gate in the LSTM to control the information inherited by the previous moment, which can be calculated as:

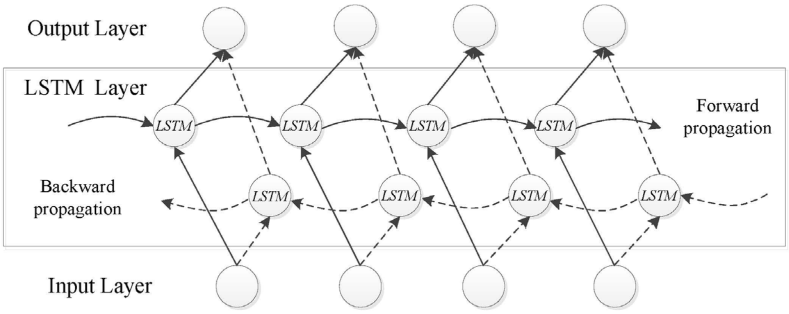

Fig. 3. The architecture of BLSTM model

$$f_t^i = sigmoid(W_f h_{t-1}^i + U_f x_t^i + b_f), \tag{1}$$

where W_f is the wight matrix of the forget gate, U_f is the weight matrix between the input layer of the forgot gate and the hidden layer, b_f is the bias of the forget gate. The information inherited by the x_t^i can be obtained by the input gate in the LSTM, which is given as:

$$i_t^i = sigmoid(W_i h_{t-1}^i + U_i x_t^i + b_i), \tag{2}$$

where W_i is the wight matrix of the input gate, U_i is the weight matrix between the input layer of the input gate and the hidden layer, b_i is the bias of the input gate. By formula (2) it determines how much information of x_t^i is added to the memory stream. Thus, the current state of the input cell in the LSTM is:

$$C_t^{'i} = \tanh(W_c h_{t-1}^i + U_c x_t^i + b_c), \tag{3}$$

where W_c denotes the wight matrix of the current input cell, U_c is the weight matrix between the input layer of the input cell and the hidden layer, b_c is the bias of the input cell. Thus, after filtering out the information that is unimportant by the forget gate and adding the new information obtained by the input gate, the updated input cell can be given as:

$$C_t^i = \tanh(f_t C_{t-1}^i + i_t C_t^{'i}), \tag{4}$$

where C_{t-1}^i is the last state of the input cell. Finally, how much of current input cell can be used for the next layer network update is determined by the output gate which can be given as:

$$o_t^i = sigmoid(W_o h_{t-1}^i + U_o x_t^i + b_o), \tag{5}$$

where W_o denotes the wight matrix of the output gate, U_o is the weight matrix between the input layer of the output gate and the hidden layer, b_o is the bias of the output gate. Thus, the final output h_t^i of the LSTM is:

$$h_t^i = o_t^i C_t^i. \tag{6}$$

Suppose $\vec{h}_t$ is the output of the forward LSTM and $\overleftarrow{h}_t$ is the output of the back LSTM. Thus, the output of BLSTM which is also the deep feature of x_t^i can be expressed as:

$$H_t^i = \overrightarrow{h_t^i} + \overleftarrow{h_t^i}. \tag{7}$$

Hence, the deep features of the multimodal time series sensory data X_i are $H^i = H_1^i + H_2^i + \cdots + H_L^i$.

4.3 Discriminative Feature Representation

After the deep features H^i of the perceptual data X_i are obtained, the attention mechanism is used to adjust the attention probability of the deep features. It enables ABLSTM model to selectively focus on discriminative features and reduce the irrelevant features. Firstly, the deep features H_t^i are used to obtain its implicit representation u_t through a nonlinear transformation, which can be expressed as:

$$u_t = \tanh(W_\omega H_t^i + b_\omega). \tag{8}$$

Based on a context vector u_ω, the similarity representation between u_t and u_ω is calculated to obtain the importance of the deep features. Thus, the attention probability of the deep features α_t can be given by normalizing the importance of the deep features. Here u_ω is a random initialization matrix that can focus on the important information over u_t. The attention probability of the deep features can be expressed as:

$$\alpha_t = \frac{\exp(u_t u_\omega)}{\sum \exp(u_t u_\omega)}. \tag{9}$$

Finally, the discriminative features s can be computed via the weighted sum of H_t based on α_t. s can be given as:

$$s = \sum \alpha_t H_t^i. \tag{10}$$

4.3 Activity Recognition Output

The discriminative features vector *s* generated from the attention mechanism is used for HAR. After a linear transformation of the full connection layer on *s*, a *softmax* classifier is applied to predict the probability of activities, which can be expressed as:

$$P(y=a^k)=soft\max(w_h s+b_h), \tag{11}$$

where a^k is the *k*-th category activity, w_h represents the weight matrix of the classifier, which can map *s* to a new vector with length *h*, *h* is the number of categories of activities. In order to train our ABLSTM, a binary cross entropy loss function is designed, as show in formula (12). After that, the Back Propagation Through Time (BPTT) algorithm [27] is applied to train and update the parameters in ABLSTM to minimize the cross entropy of the real and predicted activities.

$$loss=-\sum\sum y_i^k \ln a_i^k+(1-y_i^k)\ln(1-a_i^k), \tag{12}$$

where *i* is the index of the activity data and y^k is its real activity label.

5 Evaluation

To verify the effectiveness of the proposed ABLSTM, we conduct extensive experiments. Specifically, we evaluate our model on two public datasets: UCI dataset [15] and the Opportunity dataset [28]. Firstly, we evaluate the effects of the parameters of our ABLSTM model. Then, we use the best parameters to evaluate the performance of the ABLSTM model. Finally, we compare our model with some state-of-the-art works on the two datasets.

5.1 Benchmark Datasets

The final result of HAR is closely related to the dataset. We consider two typical datasets for human activity recognition. In the following, we introduce each dataset in detail.

The Public domain UCI dataset. The dataset includes six types of human activities, which are standing, sitting, lying, walking, going upstairs and going downstairs, respectively. They are recorded by accelerometer and gyroscope sensors built in smartphones. Since the accelerometer and the gyroscope are both three-axial sensors, it is a six modal for each activity. In this dataset, the sampling frequency of the sensors is 50Hz and the sampling window is set to 128 frames.

The Opportunity dataset. The dataset consists of annotated recordings from on-body sensors, which are taken by four participants who are instructed to carry out common kitchen activities. The perceptual data is recorded at a frequency of 30Hz and annotated by 18 mid-level gesture annotations (e.g. Open Door/Close Door). The dataset contains about six hours recordings in total. During the process of recordings, each participant performs a session of activities of daily living (ADL) five times and one drill session. During each ADL session, activities are performed by the participants with a loose description, such as checking ingredients and utensils in the kitchen, preparing and drinking a coffee, preparing and eating a sandwich, cleaning up etc. During the drill sessions, the participants perform 20 repetitions of a predefined sorted by 17 activities. The Null class refers to either irrelevant activities or non-activities. Therefore, there are totally 18 classes in the Opportunity dataset.

Since the data in this dataset is recorded continuously, the sampling window and its label should be determined first. In this paper, a sliding window scheme is used to the sample perceptual data and its label is determined by the last frame, that is to say the label of the last frame in each sampling window is set as its label. Besides, in order to add some redundancy information, the sliding window is overlapped by 50%. The detail process of sliding and labelling is shown in Fig. 4. Repeating the operation above can generate a dataset suitable for the training and testing. To determine the optimal sampling window, the dataset with different sampling windows are tested with ABLSTM model, the result is shown in Fig. 5. From this figure, we can observe that when the sampling window is set to 3000ms, F1 score of ABLSTM model is the highest. Thus, the sampling window of the Opportunity dataset is set to 3000ms with a step size of 1500ms.

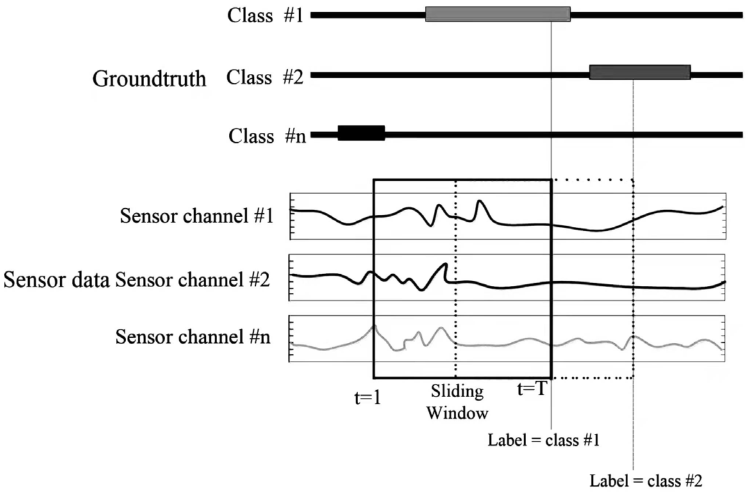

Fig. 4. The sampling window label on the Opportunity dataset. The label of the last frame in each sampling window is set as its label

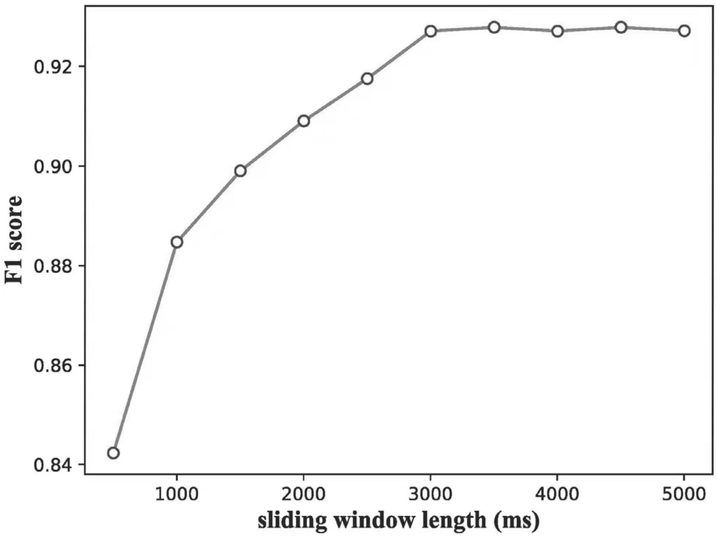

Fig. 5. F1 score performance of ABLSTM model on the Opportunity dataset with different sliding window length

5.2 Evaluation Metric

Since the Opportunity dataset is extremely imbalanced, activity classes with fewer samples are more likely to be misclassified than those with more samples. If we use accuracy as the indicator to evaluate the performance, this could achieve higher accuracy than its actual situation. Therefore, we choose F1 score as the indicator to evaluate the performance on the Opportunity dataset. The public domain UCI dataset is relatively balanced that we can use accuracy and the F1 score to evaluate this model. Thus, both accuracy and F1 score are used in this paper. F1 score combines with precision and recall measure. Precision is defined as TP/(TP+FP), and recall corresponds to TP/(TP+FN), where TP presents the value of correctly classified activities. FP presents the value that other activities misclassified to the current activity. FP presents the value that the current activity misclassified to other activities. Besides, it is a multi-class classification for human activity recognition. Thus, the evaluation metrics of the F1 score is defined as:

$$F1=\sum_{k}2\times\omega_k\frac{\text{Precision}_k\cdot \text{Re}call_k}{\text{Precision}_k+\text{Re}call_k}, \quad \textbf{(13)}$$

where k is the class index. $\omega_k = n_k/N$ is the proportion of samples of class k with n_k being the number of samples of the k-th class and N being the total number of samples.

5.3 Parameter Optimization

For the deep learning models, their parameters determination mainly relays on the experiences of designers and the optimal parameters usually different for different tasks. Thus, to determine the optimal parameters of ABLSTM, we conduct experiments on the two datasets with different parameters.

(1) Hidden layer units (H) of ABLSTM: The number of hidden layer units at each level is an important parameter, which directly affects the ability of deep features extraction. Generally speaking, the more hidden layer nodes, the stronger feature extraction ability, but also leading to higher computational complexity. Fig. 6 depicts F1 score as a function of the hidden layer units on the Public domain UCI dataset and Opportunity dataset. As the number of hidden layer units increases, the F1 score on both datasets increases. After H = 32, F1 score of ABLSTM is stable, to further increase in the number of hidden layer units does not improve the performance greatly as well. Thus, the hidden layer units of ABLSTM model is set to H = 32.

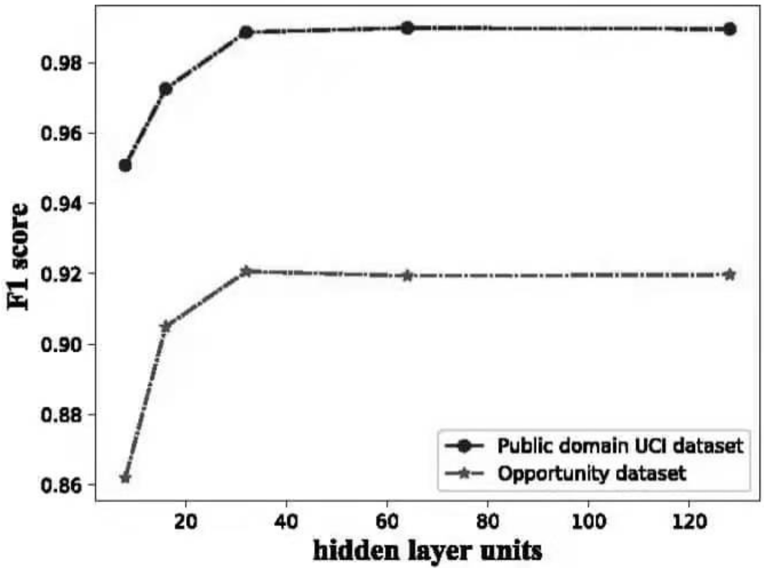

Fig. 6. The Performance of ABLSTM model with the different number of hidden layer units

(2) Learning rate (η) of ABLSTM: The learning rate determines the update magnitude of ABLSTM's parameters. If the amplitude is too large, the parameters may fetch values on both sides of the optimal values, leading to hard convergence. On the contrary, if the amplitude is too small, convergence is guaranteed, but the optimization speed will be greatly reduced. Fig. 7 depicts F1 score as a function of the learning rate on the Public domain UCI dataset and Opportunity dataset. It can be observed the increase in the value of η can improve the F1 score of the model. When $\eta = 0.0025$, the F1 score gets a higher value and more stable state. Consequently, $\eta = 0.0025$ is selected as the learning rate for ABLSTM model.

5.4 Performance Analysis of ABLSTM

Table 1 shows the confusion matrix on the Public domain UCI dataset for our ABLSTM model. WK, WU, WD, ST, SA and LA represent WALKING, WALKING UPSTAIRS, WALKING DOWNSTAIRS, SITTING, STANDING and LAYING, respectively. The confusion matrix contains information about the actual and the predicted activity categories. The horizontal axis of the confusion matrix is the category of the predicted activities, and the vertical axis is the category of the actual activities. The diagonal line indicates the number of samples that are correctly predicted. From this table, we can observer that if there is a big difference between activities sensed by accelerometer and gyroscope sensors, such as WK, WU, WD AND LA, they can be well identified and there is no confusion between them at all. Else, the

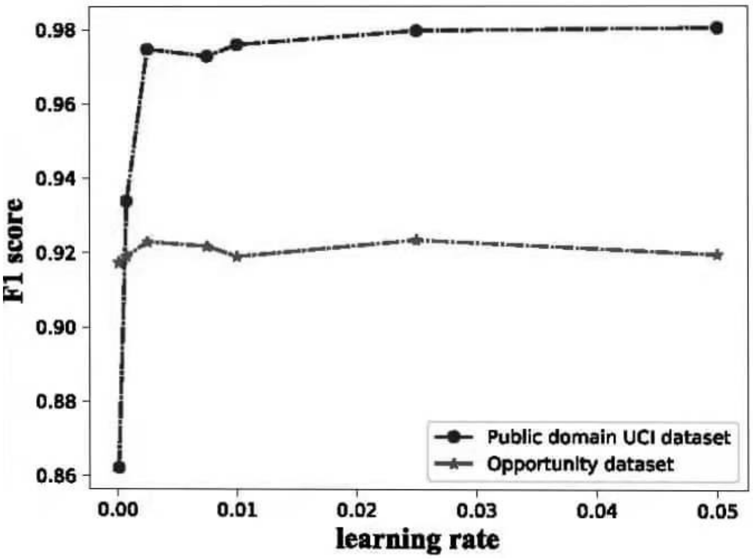

Fig. 7. The Performance of the ABLSTM model with different values of learning rate

Table 1. Confusion matrix yielded by ABLSTM model on the public domain UCI dataset

	WK	WU	WD	ST	SA	LA	Recall
WK	**496**	0	0	0	0	0	99.6%
WU	0	**471**	0	0	0	0	100%
WD	0	0	**491**	0	0	0	99.3%
ST	0	0	0	**454**	37	0	100%
SA	0	0	0	15	**517**	0	95.4%
LA	0	0	0	0	0	**537**	100%
Precision	100%	100%	100%	95.3%	100%	100%	**99.0%**

activities that have similar movement characteristics, such as ST and SA that the gyroscope sensors are insensitive to the two activities, are easily misclassified. Finally, the precision of the six types of activities are in the range of 95.3% to 100%, and the recall of them are in the range of 95.4% to 100%. F1 score of ABLSTM model on the Public domain UCI dataset can be up to 99.0%.

The confusion matrix on the Opportunity dataset for ABLSTM model is illustrated in Table 2. Different from the Public domain UCI dataset, the Opportunity dataset is extremely imbalanced as the Null class is more than 75% of the recorded data. Consequently, most classification errors are related to this class, as shown in this table. Almost every activity can be misclassified by NULL class to some degree. Besides, two group activities (Open Door 1-Close Door 1 and open drawer 1-Close drawer 1) tend to be misclassified by each other since they have similar motion characteristics. Thus, they involve the activation of the same type of sensors, but only with a different sequentiality. Except for these two groups of activities, the precision and recall of the other activities are all above 89.3% and 91.1%, respectively. F1 score of ABLSTM model on the Opportunity dataset is 92.7%.

5.5 Comparison to the State of the Art

To verify the effectiveness of ABLSTM model, we choose some advanced works of activity recognition in [28-32] for the performance comparison. In [28], the authors introduced a versatile human activity dataset Opportunity recorded in a sensor-rich environment and reported the performance achieved by standard classification techniques such as k-NN, NCC, LDA, and QDA. In [29], the authors proposed a systematic feature learning method for HAR problem. This method adopted a deep convolutional neural networks (CNN) to automate feature learning from the raw inputs in a systematic way on the Opportunity dataset. In [30], the authors proposed using deep network architecture comprised of Deep Convolutional and LSTM Recurrent neural network for human behavior recognition on the Opportunity dataset. In [31], the authors evaluated four different recognition algorithms based on deep convolutional neural network and recurrent neural network to identify the most effective method for human activity recognition on the

Public domain UCI dataset. In [32], the authors used the deep residual LSTM network to recognize human activities on the Opportunity dataset and the Public domain UCI dataset. The residual network was used to improve the learning ability and recognition rate of BLSTM. At the same time, LSTM-based and BLSTM-based networks are also optimized for the two datasets by deeper network layer and parameter optimization. The comparison results among these works on the Public domain UCI dataset and Opportunity dataset are shown in Table 3 and Table 4, respectively.

Table 2. Confusion matrix yielded by ABLSTM model on the Opportunity dataset

	NU-LL	open door1	open door2	close door1	close door2	open fridge	close fridge	open dish washer	close dish washer	open drawer 1	close drawer 1	open drawer 2	close drawer 2	open drawer 3	close drawer 3	clean table	drink cup	toggle switch	Recall (%)
NULL	**6539**	1	1	4	0	1	11	1	0	3	6	1	0	0	2	1	6	5	99.0
open door1	6	**37**	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	85.1
open door2	4	0	**88**	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	96.2
close door1	8	3	0	**35**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	79.5
close door2	3	1	0	0	**59**	0	0	0	0	0	0	0	0	0	0	0	0	0	94.4
open fridge	22	0	0	0	0	**196**	0	0	1	0	0	0	0	0	0	0	0	1	93.3
close fridge	11	0	0	0	0	2	**180**	0	0	0	0	0	0	0	0	0	0	1	93.0
open dish washer	2	0	0	0	0	0	0	**93**	0	0	0	0	0	0	0	0	0	1	97.4
close dish washer	3	0	0	0	0	0	1	0	**69**	0	0	0	0	0	0	0	0	1	95.2
open drawer1	1	0	0	0	0	1	0	0	0	**25**	0	1	0	0	0	0	0	1	84.7
close drawer1	6	0	0	0	0	0	1	0	0	1	**29**	1	0	0	0	0	0	0	79.5
open drawer2	2	0	0	0	0	0	0	0	0	0	0	**40**	0	0	0	0	0	0	94.1
close drawer2	2	0	0	0	0	0	0	0	0	0	0	0	**25**	0	0	0	0	1	94.3
open drawer3	2	0	0	0	0	0	0	0	0	0	0	0	0	**50**	0	0	0	0	98.0
close drawer3	4	0	0	0	0	0	0	0	0	0	0	0	0	0	**62**	0	0	0	95.4
clean table	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**78**	0	0	98.1
drink cup	3	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	**295**	0	98.0
toggle switch	2	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	**113**	93.8
Precision (%)	99.3	82.2	94.6	72.9	93.7	89.1	92.8	96.9	93.2	86.2	76.3	95.2	89.3	96.2	93.9	97.5	98.3	96.6	**92.7**

Table 3. Comparison results on the public domain UCI dataset

Method	Accuracy	F1 score
LSTM [31]	87.38%	-
BLSTM [31]	84.54%	-
CNN [31]	85.4%	-
Dropout [31]	90.98%	-
LSTM [32]	90.77%	90.77%
BLSTM [32]	91.09%	91.11%
Res-BLSTM [32]	93.57%	93.54%
ABLSTM	**99.4%**	**99.0%**

Table 4. Comparison results on the Opportunity dataset

Method	F1 score
LDA [28]	69%
QDA [28]	53%
NCC [28]	51%
1NN [28]	87%
3NN [28]	85%
UP [28]	64%
NStar [28]	64%
SStar [28]	86%
BDN [32]	73%
CNN [32]	85.1%
DeepConvLSTM [30]	89.5%
Res-BLSTM [32]	90.2
ABLSTM	**92.7%**

From Table 3, we can observe that the accuracy and the F1 score of ABLSTM model are 99.4% and 99%, respectively, which are higher than other models. This is because our ABLSTM not only can capture the multimodal time series features of the sensory data, but also it can learn more discriminative features representation by the attention mechanism than other models. In addition, it can also observe that the deeper LSTM-based and BLSTM-based networks proposed in [32], the higher classification accuracy. The reason is deeper network layer can capture more discriminative features, but it will increase the computational complexity. Besides, even though they can learn more discriminative features, they treat all the features equally. Thus, they have lower classification accuracy than our ABLSTM. For CNN-based methods, they have to use a fixed convolution kernel to extract deep features, which cannot well capture the dynamic features of the time series data. Besides, although Res-BLSTM introduces the residual network into BLSTM to enhance the deep features expression, discriminative features learning scheme plays a more important role in activity recognition. Thus, ABLSTM model is better than Res-BLSTM as well.

Table 4 presents the comparison results of F1 score on the Opportunity dataset. These models are based on a sliding window-based data, but the feature extraction and classifier are different. From this table, we can see that ABLSTM model outperforms other models, improving the non-recurrent deep learning by 11% on average. Compared to DeepConvLSTM and Deep-Res-BLSTM, F1 score of ABLSTM is higher because it incorporates attention mechanism which highlights the impact of deep key features to obtain more differentiated features. Furthermore, if the learned feature of the activities are not sufficient, deep-learning based activity classification performed worse than some standard classification techniques, such as K-NN and SStar, which further verifies the importance of discriminative feature learning. These results prove that ABLSTM model can offer a significant advantage across very different scenarios.

In summary, ABLSTM model achieves the highest recognition accuracy for time series data, which is sufficient for human activity recognition. Therefore, ABLSTM model is an effective activity recognition method.

6 Conclusion

In this paper, we demonstrate the advantages of the combination of attention mechanism and BLSTM model for activity recognition of mobile sensing. Firstly, the multimodal sensory data is segmented according to the size of the predefined sampling windows and fed to the BLSTM layer to get the deep features of activities. After that, the attention mechanism is used to selectively focus on discriminative deep features of the deep features. Through the experiments on two datasets derived from real world scenarios, the results show that the recognition accuracy of the public domain UCI dataset is improved by 5.83% and the F1 score of the Opportunity dataset is increased by 2.1% compared with the state of the art. Hence, we believe that the proposed method can be used as a powerful tool for human activity recognition problem. Besides, there is still a limitation that ABLSTM just lets the network learn the discriminative features of various activities, it doesn’t fundamentally solve the problem of features overlap. In the future, we will focus on key frame sampling, removing the overlapped data to enhance the discriminability of various activities.

Acknowledgements

This work is supported in part by the Natural Science Foundation of China under Grants 61902282, the Natural Science Foundation of Tianjin under Grants 18JCYBJC85900, 18JCQNJC74100 and 20JCZDJC00400, the Tianjin Municipal Education Commission Project for Scientific Research Plan under Grant 2018KJ155, the Science and Technology Development Fund of Tianjin under Grant No.JW1702 and the Science and Technology Project of Guangdong Province under Grant 2017KQNCX194.

References

[1] J. Wang, Y. Chen, S. Hao, X. Peng, L. Hu, Deep learning for sensor-based activity recognition: A survey, Pattern Recognition Letters, 119(2017) 3-11.

[2] A.-D. Ignatov, V.-V. Strijov, Human activity recognition using quasi periodic time series collected from a single tri-axial accelerometer, Multimedia Tools & Applications 75(12)(2016) 7257–7270.

[3] Z.-W.-H. Xia, J. Wang, Human activity recognition based on transformed accelerometer data from a mobile phone, Journal of Xian University of Posts & Telecommunications 29(13)(2016) 1981-1991.

[4] X.-U. Chuan-Long, G.-U. Qinlong, Y. Minghai, Activity recognition method based on three-dimensional accelerometer, Computer Systems & Applications 22(6)(2013) 132-135.

[5] W. Wu, S. Dasgupta, E.-E. Ramirez, C. Peterson, G.-J. Norman, Classification accuracies of physical activities using smartphone motion sensors, Journal of Medical Internet Research 14(5)(2012) e130.

[6] M. Chunmei, Z. Jinqi, W. Shang, L. Bin, Representation learning from time labelled heterogeneous data for mobile crowdsensing, Mobile Information Systems (2016)(PT.4) 2097243.1-2097243.10.

[7] B. Alejandro, S. Yago, I. Pedro, Evolutionary design of convolutional neural networks for human activity recognition in sensor-rich environments, Sensors 18(4)(2018) 1288.

[8] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlter, H. Ney, A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition, in: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

[9] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, Y. Bengio, End-to-end attention-based large vocabulary speech recognition, in: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

[10] M.-T. Luong, H. Pham, C. Manning, Effective approaches to attention-based neural machine translation, Computer ence, arXiv:1508.04025 (2015).

[11] L. Koping, K. Shirahama, M. Grzegorzek, A general framework for sensor-based human activity recognition, Computers in Biology & Medicine (2018) S0010482517304225.

[12] N. Wichit, Multisensor data fusion model for activity detection, in: Proc. 2014 Twelfth International Conference on ICT and Knowledge Engineering, 2015.

[13] P. Asghari, E. Soleimani, E. Nazerfard, Online human activity recognition employing hierarchical hidden markov models, Journal of ambient intelligence and humanized computing 11(3)(2020) 1141-1152.

[14] A. Parate, M. Chiu, RisQ: Recognizing smoking gestures with inertial sensors on a wristband, in: Proc. International Conference on Mobile Systems, Applications and Services, 2014.

[15] D. Anguita, A. Ghio, L. Oneto, X. Parra, J.-L. Reyesortiz, A public domain dataset for human activity recognition using smartphones. <https://upcommons.upc.edu/handle/2117/20897>, 2013.

[16] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, J. Zhou, TSE-CNN: A Two-Stage End-to-End CNN for Human Activity Recognition, IEEE Journal of Biomedical and Health Informatics 24(1)(2020) 292-299.

[17] X. Hanjian, Y.-E. Luo, H. Zhou, The Effect of Axis-Wise Triaxial Acceleration Data Fusion in CNN-Based Human Activity Recognition, IEICE Transactions on Information and Systems E103.D(4)(2020) 813-824.

[18] M. Pawlyta, M. Hermansa, A. Szczsna, M. Janiak, K. Wojciechowski, Deep Recurrent Neural Networks for Human Activity Recognition During Skiing, Man-Machine Interactions 6(2020) 136-145.

[19] M. Abdulmajid, P. Jaeyoung, Deep Recurrent Neural Networks for Human Activity Recognition, Sensors 17(11)(2017) 2556.

[20] A. Anagnostis, L. Benos, D. Tsaopoulos, A. Tagarakis, D. Bochtis, Human Activity Recognition through Recurrent Neural Networks for Human-Robot Interaction in Agriculture, Applied Sciences 11(2188))(2021) 2188.

[21] S. Tongtong, S. Huszhi, M. Chunmei, J. Lifen, X. Tongtong, HDL: Hierarchical Deep Learning Model based Human Activity Recognition using Smartphone Sensors, in: Proc. 2019 International Joint Conference on Neural Networks (IJCNN), 2019.

[22] C.-F. Martindale, V. Christlein, P. Klumpp, B.M. Eskofier, Wearables-based multi-task gait and activity segmentation using recurrent neural networks, Neurocomputing 432(6)(2020).

[23] V. Hernandez, T. Suzuki, G. Venture, Convolutional and recurrent neural network for human activity recognition: Application on American sign language, PLoS ONE 15(2)(2020)e0228869.

[24] E. Sansano, R. Montoliu, S. Fernández, A study of deep neural networks for human activity recognition, Computational Intelligence 36(6)(2020).

[25] Q. Luyue, W. Sheng, Review on time-series data visualization, Microcomputer & Its Applications 34(12)(2015) 7-10.

[26] K. Greff, R.-K. Srivastava, J. Koutnłk, B.-R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, IEEE Transactions on Neural Networks & Learning Systems 28(10)(2016) 2222–2232.

[27] Z. Ming, T.-N. Le, Y. Bo, O.-J. Mengshoel, J. Zhang, Convolutional neural networks for human activity recognition using mobile sensors, in: Proc. 6th International Conference on Mobile Computing, Applications and Services, 2015.

[28] R. Chavarriaga, H. Sagha, A. Calatroni, S.-T. Digumarti, G. TrSter, J.-D.-R. Milln, D. Roggen, The opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognition Letters 34(15)(2013) 2033–2042.

[29] J.-B. Yang, M.-N. Nguyen, P.-P. San, X.-L. Li, P.-K. Shonali, Deep convolutional neural networks on multichannel time series for human activity recognition, in: Proc. 2015 International Joint Conferences on Artificial Intelligence (IJCAI), 2015.

[30] O. Francisco, R. Daniel, Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition, Sensors 16(1)(2016) 115.

[31] K. Xiaohua, H. Jun, H. Zhaohua, Z. Yuan, Comparison of deep feature learning methods for human activity recognition, Application Research of Computers 35(9)(2018) 2815-2822.

[32] Y. Zhao, R. Yang, G. Chevalier, M. Gong, Deep residual bidir-lstm for human activity recognition using wearable sensors, Mathematical Problems in Engineering 2018(2018).