Biomedical Engineering Letters (2020) 10:419–430
https://doi.org/10.1007/s13534-020-00160-x

ORIGINAL ARTICLE

Mobile sensor based human activity recognition: distinguishing of challenging activities by applying long short-term memory deep learning modified by residual network concept

Seyed Vahab Shojaedini[1] · Mohamad Javad Beirami[2]

Received: 21 October 2019 / Revised: 7 June 2020 / Accepted: 15 June 2020 / Published online: 29 June 2020

Abstract

Automated recognition of daily human tasks is a novel method for continuous monitoring of the health of elderly people. Nowadays mobile devices (i.e. smartphone and smartwatch) are equipped with a variety of sensors, therefore activity classification algorithms have become as useful, low-cost, and non-invasive diagnostic modality to implement as mobile software. The aim of this article is to introduce a new deep learning structure for recognizing challenging (i.e. similar) human activities based on signals which have been recorded by sensors mounted on mobile devices. In the proposed structure, the residual network concept is engaged as a new substructure inside the main proposed structure. This part is responsible to address the problem of accuracy saturation in convolutional neural networks, thanks to its ability in jump over some layers which leads to reducing vanishing gradients effect. Therefore the accuracy of the classification of several activities is increased by using the proposed structure. Performance of the proposed method is evaluated on real life recorded signals and is compared with existing techniques in two different scenarios. The proposed structure is applied on two well-known human activity datasets that have been prepared in university of Fordham. The first dataset contains the recorded signals which arise from six different activities including walking, jogging, upstairs, downstairs, sitting, and standing. The second dataset also contains walking, jogging, stairs, sitting, standing, eating soup, eating sandwich, and eating chips. In the first scenario, the performance of the proposed structures is compared with deep learning schemes. The obtained results show that the proposed method may improve the recognition rate at least 5% for the first dataset against its own family alternatives in distinguishing challenging activities (i.e. downstairs and upstairs). For the second data set similar improvements is obtained for some challenging activities (i.e. eating sandwich and eating chips). These superiorities even reach to at least 28% when the capability of the proposed method in recognizing downstairs and upstairs is compared to its non-family methods for the first dataset. Increasing the recognition rate of the proposed method for challenging activities (i.e. downstairs and upstairs, eating sandwich and eating chips) in parallel with its acceptable performance for other non-challenging activities shows its effectiveness in mobile sensor-based health monitoring systems.

Keywords Human activity recognition · Mobile sensor · Deep learning · Convolutional neural networks · Long short-term memory · Residual networks

1 Introduction

Monitoring of the activities of alone old people is one of the important issues in modern electronic healthcare [1]. Recognition and distinguishing the above activities may be utilized for several applications including assistive living, rehabilitation, and surveillance. Although video-based monitoring is the simplest way for human activity recognition but this technique was less appropriately addressed, because of its privacy-invasive nature. In parallel with recent advances in Micro Electro Mechanical Systems (MEMS), low-cost small

✉ Seyed Vahab Shojaedini
shojadini@irost.ir

Mohamad Javad Beirami
mj.Beirami@qiau.ac.ir

[1] Iranian Research Organization for Science and Technology, Tehran, Iran

[2] Faculty of Electrical, Biomedical and Mechatronics Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

size sensors have emerged. Such sensors are widely used in smartphones, smartwatches, and healthcare devices. Some of these sensors (for example accelerometer and gyroscope) enables smartphones and smartwatch to be utilized as equipment for monitoring human activities [2]. The similarity of the signals which are captured from different activities of a person (for example upstairs and downstairs or eating activities) caused the discrimination of several human activities remains a challenging classification problem. In addition, a high amount of recorded data increases the computational cost of recognition algorithms.

In several studies, various linear and nonlinear classification schemes have been proposed to address the above limitations. The main objective of these schemes is to obtain acceptable accuracy, especially when the similarity of the activities increases. Some primary approaches try to distinguish activities based on their simple features such as mean [3–5] or variance [6].Although using such methods shows acceptable results for several simple activities (e.g. standing, sitting, and running) but their outcomes for some complicated activities (e.g. stretching and riding elevator) are not satisfactory.

Some researches make use of Fourier transform domain for extracting suitable features that may distinguish between similar human activities. These features may reflect the frequency-based properties of several activities, therefore they may differentiate between activities based on their various frequency contents. Unfortunately, this family of methods has no sufficient accuracy in recognizing activities in parallel with their high computational cost.

In some studies [7], the support vector machine (SVM) is used as a classifier that makes use of Gaussian kernels. Using this type of kernels results in more flexibility in decision boundaries, therefore ultimately increase the accuracy of the recognition of several activities. However, in the above method, the strategy of selecting SVM parameters is very challenging, because the resultant accuracy is highly dependent on these parameters.

Some sophisticated methods try to classify different activities by constructing the Hidden Markov Model (HMM) [8]. Although this technique has shown better results than many of its older alternatives, however, its performance is highly dependent on the quantity and quality of the extracted features from the recorded signals.

In parallel with recent advances in manufacturing processors with high computational power, deep neural networks have attracted a lot of attention as an effective paradigm to overcome challenges of human activity recognition. In this method, a deep neural network extracts non-handcrafted features from its raw input data [9]. Furthermore, deep neural networks are based on the learning of multiple levels of representations of the data. Such a multi-level representation scheme in parallel with their deep architecture (several processing layers) enables them to get more accurate results [9]. Deep convolutional neural networks (CNNs) are initiated from deep learning theory which is based on large scale data and different types of layers. A portion of this structure is responsible for extracting discriminative features of input data, while others are responsible for the classification of the data based on extracted features. Based on the above abilities, deep convolutional networks have been widely used for separating human activities in recent years containing several standard, partial, or full weight sharing versions (e.g. partial weight sharing in the first convolutional layer and full weight sharing in the second convolutional layer) [10]. Unfortunately, temporal-dependency which is the main characteristic of human activity signals has not been addressed in classic CNN which hampers its performance in human activity recognition. Therefore in complementary researches, the temporal dependency of data has been incorporated in the solution. In some researches, recurrent neural networks (RNNs) were used to consider the time dependence of human activities in constructing deep neural networks [11]. In more complicated solutions several combinations of CNN and long-short term memory (LSTM) were introduced in order to extract temporal and local features simultaneously [12]. Although these methods enabled researchers to improve the results of human activity recognition systems, however the accuracy saturation still remains as an important limiting factor in this application.

In this paper, a new method is introduced to improve distinguishing different human activities. The proposed algorithm is based on minimizing the accuracy saturation phenomenon along with improving the optimization ability of LSTM-CNN. In our proposed algorithm the temporal deep learning scheme is modified by using the concept of residual network. The resultant architecture utilizes shortcuts to jump over some layers thanks to existing residual networks in its body. Therefore the problem of vanishing gradients is addressed by reusing activations from a previous layer until the layer next to the current one has learned its weights. Consequently, the problem of accuracy saturation is greatly reduced. The paper is organized as follows. In Sect. 2, the proposed approach is demonstrated including dataset and pre-processing, learning deep neural network and the classification of activities. In Sect. 3, the performance of the proposed method is evaluated by comparing its results with the results of deep learning methods. In Sect. 4, the obtained results from the proposed scheme are compared with the results of non-deep techniques. Finally, conclusion is presented in the last section of the paper.

2 Methods

In this section, the details of the proposed method are described. Firstly the LSTM-CNN deep structure is introduced and then the residual network is applied to improve the performance of deep structure against vanishing gradients problem and increases the optimization performance of the network by identity mapping. Finally, the classification is performed in order to distinguish six human activities.

2.1 Convolutional neural network

Convolutional neural network (CNN) is a kind of deep neural network which has high potential in extracting high-level features. The feature extraction is perfumed in so-called convolutional layers of CNN [12] thanks to its linear and nonlinear kernels and regardless of the feature positions which makes them scale-invariant. Suppose input activity signal as:

$$x_i^0 = \left[ax_1, \ldots, ax_N\right] \tag{1}$$

In which x_i^0 may be represented by a matrix of size $3 \times N$ which N refers to the number of incorporated accelerometers. In the same manner, the input of k-th convolutional layer includes $Z_{i,j}$ feature map as demonstrated in Fig. 1. Therefore the component for (i, j) location in k-th layer and l-th feature map may be computed as:

$$z_{i,j}^{l,k} = \sigma\left(\sum_{k'=1}^{k'}\sum_{x=1}^{X}\sum_{y=1}^{Y} w_{x,y,k'}^{l-1,k'} z_{i,j+y-1}^{l-1,k'} + b^{l-1,k}\right) \tag{2}$$

In which σ and k' demonstrate activation function and number of feature maps, respectively in $(l-1)$th layer with kernel size of X and Y. Furthermore, w represents weight matrix and b shows the bias.

Finally, all feature maps are transferred into distinguished classes by using a fully connected layer. For this goal, a dense layer is used with some nodes which are equal to the number of activity classes using bellow *so-called* softmax function.

$$\text{Softmax}\left(Z_{i,j}\right) = \frac{e^{Z_{i,j}}}{\sum_{n=1}^{N} e^{Z_n}}, \quad n = 1, \ldots, N \tag{3}$$

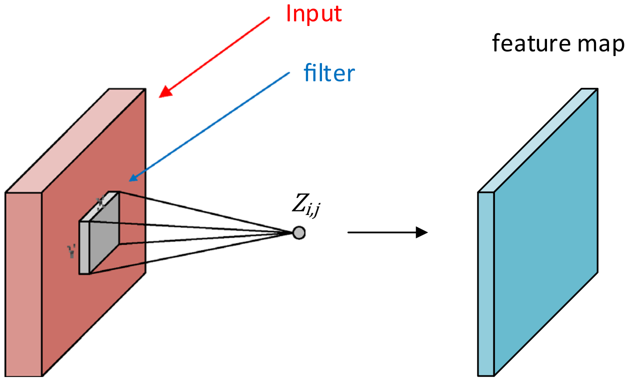

Fig. 1 Illustration of convolution operation with X and Y kernel size

2.2 Long short-term memory

Feedforward networks consider all inputs and outputs as independent elements which are not a valid assumption for time sequence phenomena such as human activity signals. To overcome this limitation, recurrent neural networks (RNNs) are used which have a great potential to model the temporal dependencies thanks to their recurrent unit which serves as a memory. The main challenges in classic RNN are vanishing gradient [13] and the limited number of memory [14] which hamper its long term temporal dependency. This weakness may seriously hamper the effectiveness of RNNs in modelling of long time series of human activity signals.

Long short-term memory (LSTM) is a type of recurrent networks to solve the vanishing gradient problem mentioned above [15]. In this network, the memory cell has been used to save the information instead of the recurrent unit. Memory cells are constructed and updated by using three main gates including write (i.e. controlling input information), read (i.e. controlling output information), and reset (i.e. forgetting useless information) [15] as demonstrated in Fig. 2.

The functionality of the LSTM which was shown in Fig. 2 may be described in details by Eqs. (4–8).

$$\mathrm{i_t} = \sigma_\mathrm{i}\left(\mathrm{w_{zi}z_t} + \mathrm{w_{hi}h_{t-1}} + \mathrm{w_{ci}c_{t-1}} + \mathrm{b_i}\right) \tag{4}$$

$$\mathrm{f_t} = \sigma_\mathrm{f}\left(\mathrm{w_{zf}z_t} + \mathrm{w_{hf}h_{t-1}} + \mathrm{w_{cf}c_{t-1}} + \mathrm{b_f}\right) \tag{5}$$

$$\mathrm{c_t} = \mathrm{f_tc_{t-1}} + \mathrm{i_t}\sigma_\mathrm{c}\left(\mathrm{w_{zc}z_t} + \mathrm{w_{hc}h_{t-1}} + \mathrm{b_c}\right) \tag{6}$$

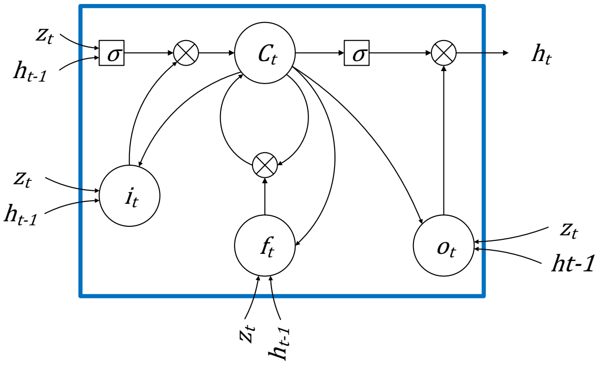

Fig. 2 LSTM cell including write, read and forget gate to save information

$$o_t = \sigma_o\left(w_{zo}z_t + w_{ho}h_{t-1} + w_{co}c_t + b_o\right) \tag{7}$$

$$h_t = o_t\sigma_h\left(c_t\right) \tag{8}$$

In the above equations i, f, o, and c represents input gate, forget gate, output gate, and cell activation functions respectively. The combination of CNN and LSTM may be used to model local and temporal dependencies for long time series signals [12].

2.3 Batch normalization

Generally, the change of input distribution may cause several problems in the learning process of deep neural networks [16]. On the other hand, the presence of any amount of variance in the input of each layer shows itself in the form of more intense changes in the next layer, due to nonlinear and deep structure of CNN [17]. In order to reduce the impact of this unfavourable phenomenon, the normalization is widely used between successive layers [9, 18, 19]. In this research Batch Normalization (BN) function is applied to reduce the internal covariance shift between layers in parallel with increasing learning speed [20] as illustrated in Eq. (9) and Fig. 3.

$$BN(z) = \gamma \frac{z - E\{z\}}{\sqrt{(Var\{z\} + \varepsilon)}} + \beta \tag{9}$$

In above equation γ and β are learning parameters.

2.4 Residual network

Although depth increment leads to construct more fitted model between input and output in CNN [20], but such a deep structure faces some limitations in its training procedure. One of the important problems is the vanishing/exploding gradient [19, 21] which occurs with the stacking of more layers [22]. Degradation with the network depth increasing, cause the accuracy gets saturates and then reduces fast. This problem points out that maybe the network has problems with approximating identity mapping due to stacking nonlinear layers [22].

In this research, residual network [22] is incorporated in the combination of CNN and LSTM structures (i.e. ConvLSTM) to overcome the above-mentioned problem. This solution utilizes parameter-free connections (identity shortcuts) for connecting the input of layer to output as shown in Fig. 4.

This shortcut connections help the function to play its role more optimal. Therefore, such direct input–output connections enable the deep network to overcome accuracy saturation and overfitting problems by skipping some layers. As shown in Fig. 4, the shortcut connections can help the solver function to map the identity function easier.

The final structure of our proposed network has been shown in Fig. 5 in which the convolutional layers extract local features from a 3-axis Mobile sensor-based accelerometer as feature maps. LSTM layers used to model temporal dependency existing in the feature maps. As shown in this figure in parallel with feature extraction, BNs are applied between layers to reduce the variance of each layer. Finally, the fully connected layer maps the result of ConvLSTM into six classes of activities as shown in the right end of Fig. 5.

The complete procedure of the proposed method may be observed in pseudocode of Fig. 6.

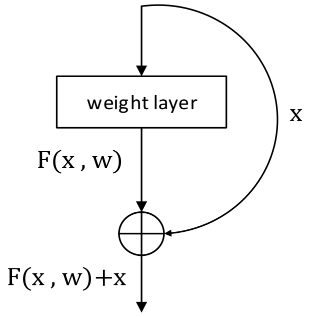

Fig. 4 Residual network which including parameter free connections (identity shortcuts) to connect the input of layer to output

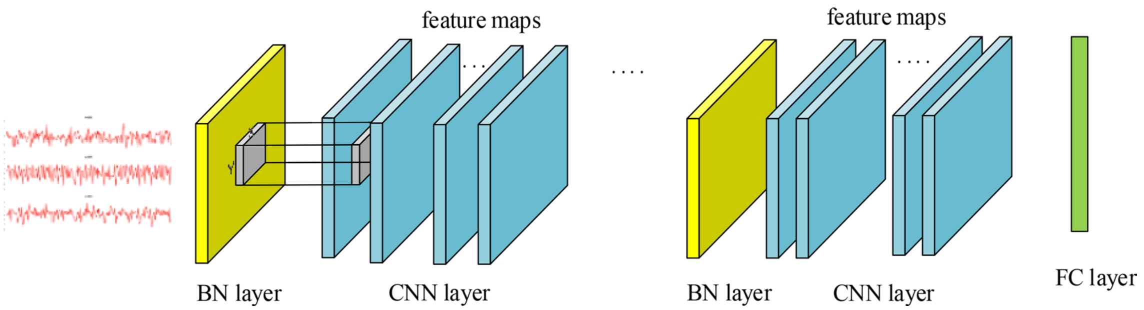

Fig. 3 Locating BN in CNN structure. Before each CNN layer, a BN layer was placed to reduce internal covariance shift

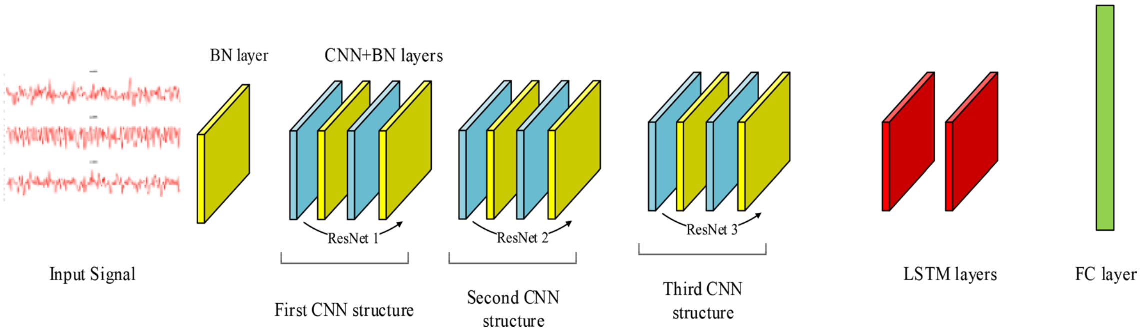

Fig. 5 The final structure of the proposed method: three CNN structures which each include two convolution layers and two BN layers. Residual shortcuts connect the first CNN output to the output of the last BN layer in each CNN section. Two LSTM layers were implanted to model temporal dependencies. Finally, the fully connected layer with softmax function, map the features into desired activity classes

Fig. 6 Description of the pseudocode of the proposed method

```
Require: Human activity raw signals
  1- Data preparation:
  2- Make train, validation and test datasets using the data of step 1
  3- While (training data is available) %fit the model on training data using step 4 to 25
        4- Apply the first BN layer
        5- Apply the previous step output as input to the first convolutional layer
        6- The second BN layer has applied to the output of step 5
        7- Use the output of step 6 to the input of the second convolutional layer
        8- Apply 3rd BN layer to the output of second convolutional layer
        9- Add residual connection to connect the 5th step output to the 3rd BN output
        10- Add the 3rd convolutional layer to the output of the last layer
        11- Apply the 4th BN layer to output of the last layer
        12- Use the output of step 11 to input of the 4th convolutional layer
        13- Apply the 5th BN layer to output of the last layer
        14- Add residual connection to connect the 10th step output to the 5th BN output
        15- Using output of the last layer as input of the 5th convolutional layer
        16- Apply the 6th BN layer to output of the last layer
        17- Use the last layer output to the input of the 6th convolutional layer
        18- Apply the 7th BN layer to the output of the last layer
        19- Add residual connection to connect the 15th step output to the 7th BN output
        20- Using flat layer to reshape the output of last layer to a vector over time steps
        21- Apply first LSTM layer to the output of last step
        22- Apply second LSTM layer
        23- Use flat layer to reshape the 2-D output of the last step to a vector
        24- Use dense layer to map output of the last layer to 6 types of activities
        25- Set the next batch as input
  26- End while
  27- While (validation data is available)
        28- Get batch of validation data
        29- Use fitted model to tune batch of validation data
        30- Apply next batch
  31- End while
  32- Use the test data to evaluate the trained network
```

3 Results

The proposed method was applied on two datasets from wireless sensor data mining (WISDM) lab [23, WISDM2] which includes over one million and 15 million raw time series data respectively. These signals have been captured from the smartphone's 3-axis accelerometer of 36 volunteers for first the dataset and data from the accelerometer sensor from smartphone and watch as 51 subjects performed 18 activities for the second one. The smartphone was fixed in the right pocket of each volunteer's pants to achieve the maximum robustness in recorded signals.

The first dataset which we used included six different activities consist of walking, jogging, upstairs, downstairs, standing, and sitting which has been captured over 10 min per each. From the second dataset, we choose eight activities from the smartwatch's accelerometer consist of walking, jogging, stairs, standing, sitting, eating soup, eating sandwich, and eating chips. For the first dataset, the volumes of data belonging to several activities are not equal because some users did not perform some of the activities due to their physical restrictions. Furthermore, some activities (i.e. sitting and standing) were limited to only a few minutes because it has been expected that the data would remain almost constant over time. More information about this dataset may be found in Table 1.

In Figs. 7 and 8 for instance, some sample signals from both datasets have been shown. These examples belong to upstairs and downstairs activities for the first dataset and two

Table 1 The details of two WISDM datasets

Dataset	Activities	Class distribution	Sample rate	Measurement time	Number of users	Number of data
First dataset	Walking, jogging, upstairs, downstairs, standing and sitting	Walking: 38.6% Jogging: 31.2% Upstairs:11.2% Downstairs: 9.1% Sitting: 5.5% Standing: 4.4%	20 Hz	10 min	36	1,098,207 sample
Second dataset	Walking, jogging, stairs, standing, sitting, eating soup, eating sandwich and eating chips	Walking: 12.5% Jogging: 12.5% Stairs:12.5% Sitting: 12.5% Standing: 12.5% Eating soup: 12.5% Eating sandwich: 12.5% Eating chips: 12.5%	20 Hz	3 min	51	1,676,282 sample

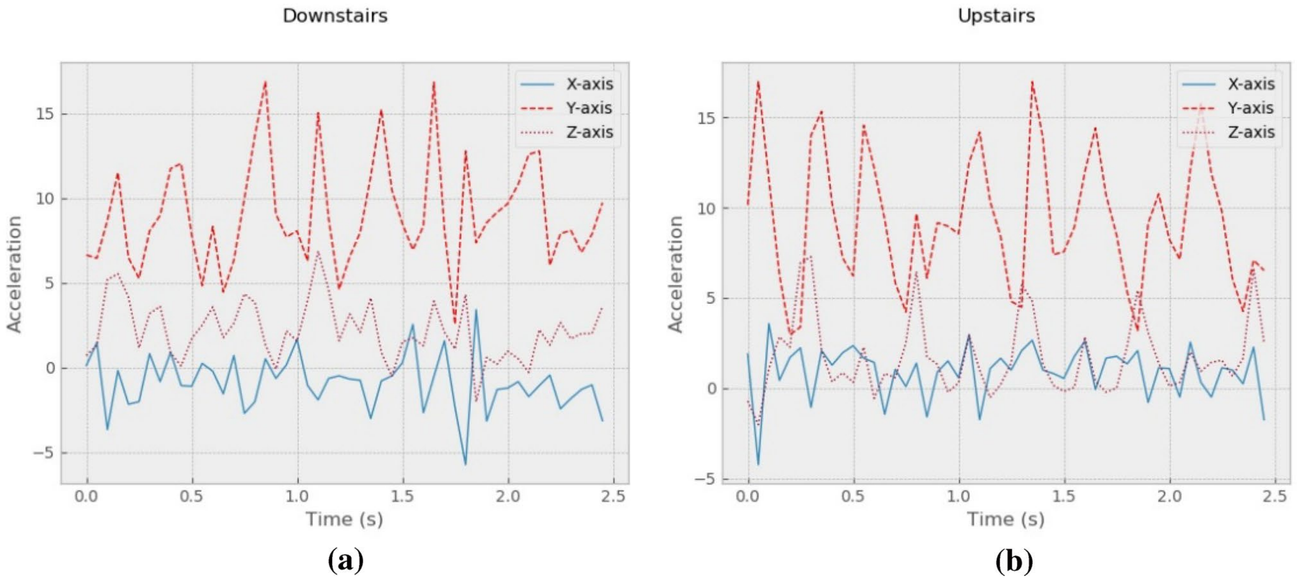

Fig. 7 Two recorded signals belong to **a** downstairs and **b** upstairs activities

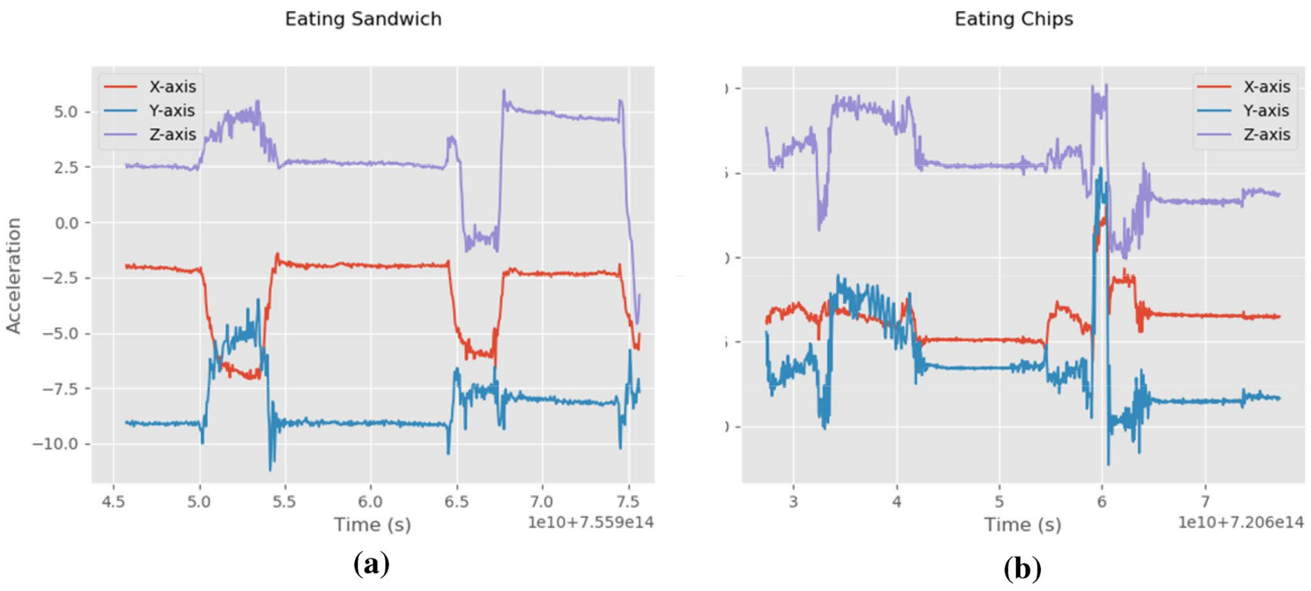

Fig. 8 Two recorded signals belong to **a** eating sandwich and **b** eating chips activities

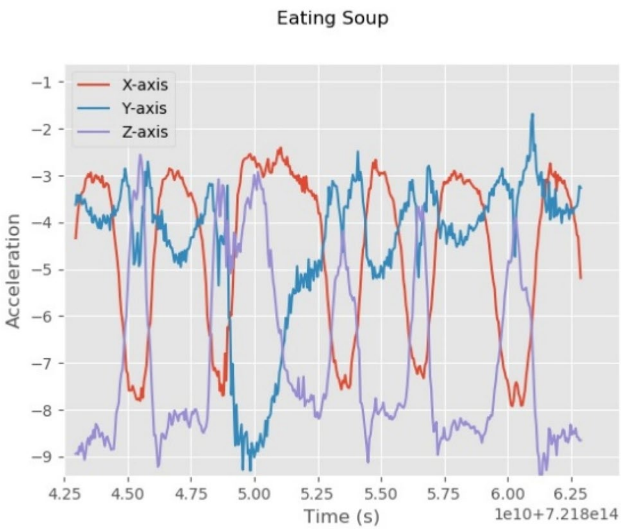

Fig. 9 Recorded signal belongs to eating soup

different eating activities for the second dataset respectively. These figures clearly show that there is no significant difference between two recorded signals belonging to downstairs and upstairs activities for the first dataset and eating sandwich and eating chips for the second one. Therefore it illustrates why recognizing these activities may be considered as a challenge in the domain of human activity recognition.

However for the Eating Soup activity, the signal is more distinguishable from the other two eating activities as showed in Fig. 9.

The proposed method was implemented on the Tensorflow framework, using the tensor processing unit (TPU) hardware developed by Google collaboratory. Furthermore, three deep learning-based alternative methods were also implemented to compare with the proposed scheme including: (a) basic CNN algorithm, (b) combination of CNN and LSTM which is called ConvLSTM for brevity in the rest of the article, and (c) ConvLSTM which has been modified by residual network (ResNet) concept which is called ConvLSTM + ResNet for brevity in the rest of the article. The accuracy of each method was obtained on the same data set and reported to evaluate the performance of several examined algorithms.

The first step of human activity recognition is data segmentation, and the most traditional approach is to use a sliding window. In this paper, the window size of 90 with a 50% overlap was used.

In this contribution, three different structures were proposed which for each of them the best configuration (number of layers and hyperparameters) were determined and showed in Table 2. The weights initialized randomly for each training procedure using stochastic gradient decent (SGD) optimizer [24] with the momentum of 0.9 and the initial learning rate of 0.01 and decay rate of 50% per 10 epochs. Table 2 shows the main parameters of examined structures.

Firstly, the performance of CNN was evaluated in distinguishing several activities of the two mentioned datasets. As demonstrated in Tables 3 and 4, this approach has obtained utterly acceptable results on those on-challenging activities in which their recorded signals were not so similar to each other.

However, the performance of this network was dropped when it was applied to recognize challenging activities (i.e. downstairs and upstairs for the first dataset and eating

Table 2 Description of main parameters of proposed method and its deep based alternatives

	Conv	ConvLSTM	ConvLSTM + ResNet
Number of parameters	57,366	39,158	164,730
Number of layers	6	8	17
Input	90 × 3 × 1	90 × 3×1 × t	90 × 3 × 1 × t
Convolution 1	[2-D conv (3 × 3), 16 2-D conv (3 × 3), 16]	[2-D conv (3 × 3), 16 2-D conv (3 × 3), 16]	[2-D conv (3 × 3), 16 + BN 2-D conv (3 × 3), 16 + BN]
Convolution 2	[2-D conv (3 × 3), 32 2-D conv (3 × 3), 32]	[2-D conv (3 × 3), 16 2-D conv (3 × 3), 16]	[2-D conv (3 × 3), 32 + BN 2-D conv (3 × 3), 32 + BN]
Convolution 3			[2-D conv (3 × 3), 64 + BN 2-D conv (3 × 3), 64 + BN]
LSTM 1		Cells = 32	Cells = 32
LSTM 2		Cells = 32	Cells = 32
Output	Softmax	Softmax	Softmax

Table 3 Results for the first dataset of WISDM classification by using CNN

	Predicted classes						Accuracy (%)
	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	
Activity classes							
Walking	1839	3	0	0	25	18	97.56
Jogging	14	1490	0	0	15	2	97.96
Sitting	0	0	266	0	0	0	100
Standing	5	0	3	203	4	0	94.42
Upstairs	15	31	3	0	467	30	85.53
Downstairs	7	0	1	0	51	388	86.99
Overall							93.74

Table 4 Results for the second dataset of WISDM classification by using CNN

	Predicted classes								Accuracy (%)
	Walking	Jogging	Sitting	Standing	Stairs	Eating soup	Eating sandwich	Eating chips	
Activity classes									
Walking	841	3	0	0	85	0	1	0	90.48
Jogging	6	905	0	0	3	0	0	0	99.01
Sitting	4	1	762	38	4	47	35	65	80.55
Standing	2	0	73	817	4	37	35	65	84.84
Stairs	87	8	6	38	792	4	7	7	85.90
Eating soup	0	0	18	32	2	734	77	67	78.92
Eating sandwich	2	0	48	43	3	111	546	179	58.58
Eating chips	1	0	67	37	12	129	205	455	50.22
Overall									78.56

sandwich and eating chips for the second dataset which caused similar signals). These decrements occurred in such way that for the first dataset the accuracies were obtained 85.53 and 86.99% for upstairs and downstairs respectively and for the second dataset, the accuracies were 58.58 and 50.22 for eating sandwich and eating chips respectively.

Tables 5 and 6 show the results obtained from applying the ConvLSTM scheme. These results demonstrate although the modification of CNN by long short-term memory may marginally improve the accuracies for the mentioned similar activities. For the first dataset, 4.03 and 2.02% improvement was obtained for upstairs and downstairs and in a similar manner for the second dataset, the improvements were 2.36 and 8.06% for eating sandwich and chips respectively). Note that these improvements were not enough to make the results acceptable for these challenging activities.

Furthermore, Tables 3 and 4 show that the accuracies of the other activities have had no meaningful difference from the results which had been obtained from basic CNN (e.g. these differences were about 1%).

Finally, Tables 7 and 8 demonstrate that the proposed method has significantly increased the obtained accuracies for two challenging activities compared to CNN. The obtained results showed that the residual network concept may improve the recognition accuracies against the basic CNN network for the first dataset by extents of 9.16 and 6.96% for upstairs and downstairs activities respectively. These improvements were obtained as 5.13 and 4.94 for the same activities compared to the ConvLSTM scheme. However, these results illustrated a tiny accuracy decrement in walking compared to basic CNN. For the second dataset, accuracies which were obtained for eating sandwich and chips were improved about 7.41 and 9.71% compared to the CNN and 5.05 and 1.65% compared to the ConvLSTM.

Table 5 Results for the first dataset of WISDM classification by using ConvLSTM

	Predicted classes						Accuracy (%)
	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	
Activity classes							
Walking	1839	6	0	0	21	19	97.56
Jogging	10	1470	0	0	35	6	96.64
Sitting	0	0	266	0	0	0	100
Standing	5	0	4	205	1	0	95.34
Upstairs	6	22	4	0	489	25	89.56
Downstairs	10	5	0	0	34	397	89.01

Table 6 Results for the second dataset of WISDM classification by using ConvLSTM

	Predicted classes								Accuracy (%)
	Walking	Jogging	Sitting	Standing	Stairs	Eating soup	Eating sandwich	Eating chips	
Activity classes									
Walking	852	1	0	0	81	1	0	0	91.12
Jogging	4	900	0	0	10	0	0	0	98.47
Sitting	5	0	775	17	5	34	31	79	81.92
Standing	0	0	58	813	2	17	25	48	84.42
Stairs	69	13	7	2	817	6	5	3	88.61
Eating soup	0	0	12	30	1	712	51	124	76.55
Eating sandwich	2	0	35	25	1	80	569	221	60.94
Eating chips	2	0	63	21	12	91	189	528	58.28
Overall									80.04

Table 7 Results for the first dataset of WISDM classification by using ConvLSTM + ResNet

	Predicted classes						Accuracy (%)
	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	
Activity classes							
Walking	1837	0	0	0	29	19	97.45
Jogging	12	1483	1	0	23	2	97.5
Sitting	0	0	262	0	0	4	98.5
Standing	0	0	5	206	4	0	95.81
Upstairs	3	4	4	0	517	18	94.69
Downstairs	6	2	4	0	15	419	93.95
Overall							96.32

Table 8 Results for the second dataset of WISDM classification by using ConvLSTM + ResNet

	Predicted classes								Accuracy (%)
	Walking	Jogging	Sitting	Standing	Stairs	Eating soup	Eating sandwich	Eating chips	
Activity classes									
Walking	892	0	0	0	43	0	0	0	95.4
Jogging	6	906	0	0	2	0	0	0	99.12
Sitting	2	6	816	10	4	24	22	62	86.26
Standing	1	0	62	807	5	17	24	47	83.8
Stairs	59	5	2	3	835	3	9	6	90.56
Eating soup	1	0	7	8	3	725	91	95	77.96
Eating sandwich	1	0	42	20	0	91	615	163	65.99
Eating chips	4	0	54	14	8	76	207	543	59.93
Overall									82.38

4 Discussion

In the previous section, the superiority of the proposed algorithm against CNN based schemes has been investigated. The common aspect of all those algorithms was that all of them belong to deep neural networks family, hence all of them extract features from raw data by using their convolutional layers. In this section, the performance of the proposed algorithm is compared with the feature-based classifiers as an alternative family for deep methods. To perform such comparison, five feature-based activity recognition methods were applied on the first dataset of WISDM. The alternative algorithms include (a) a combination of hand-crafted features and Random Forest classifier which is called for brevity as Basic features + RF in this article [23, 25], (b) principal component analysis (PCA) based on empirical cumulative distribution function which is called for brevity as PCA + ECDF in this article [26, 27], (c) logistic regression [23, 28], (d) a decision tree algorithm used for classification which is called J48 algorithm in the article [23, 28] and finally (e) multilayer perceptron [23]. The classification accuracy was calculated for the proposed method and the all above alternatives to compare their effectiveness. Table 9 shows the obtained accuracies for all examined methods. This table describes that for upstairs activity the recognition accuracy of the proposed algorithm was 33.23%, 35.43%, 67.17%, 35.14%, and 28.02% better than multilayer perceptron, J48, logistic regression, PCA + ECDF, and basic features + RF methods, respectively. Also, this table shows that for downstairs activity, the recognition accuracy of the proposed algorithm was 49.65%, 38.47%, 81.69%, 54.35%, and 44.13% better than the above alternatives.

For the other three non-challenging activities (e.g. sitting, standing and jogging) although the proposed algorithm recognized activities overall better than its alternatives, but for most of the test items the performances of examined methods were in acceptable range in accordance with those which were described in deep learning methods (see Tables 3, 5, 7). However, it is important to note that even for these non-challenging activities the superiorities of the proposed scheme against alternative methods have reached up to 16.03% (e.g. proposed method vs Basic features + RF

Table 9 Comparison of Results first dataset of WISDM classification by using proposed method and its feature based alternatives

Activity type	Basic features + RF (%)	PCA + ECDF (%)	Logistic regression (%)	J48 (%)	Multilayer perceptron (%)	CNN (%)	CNNLSTM (%)	CNN + LSTM + ResNet (%)
Walking	83.56	98.54	93.58	89.90	91.68	97.56	97.56	97.45
Jogging	94.72	95.44	97.95	96.52	98.33	97.96	96.64	97.50
Sitting	82.47	100	92.20	95.74	95.05	100	100	98.5
Standing	95.76	100	86.99	93.27	91.93	94.42	95.34	95.81
Upstairs	66.67	59.55	27.52	59.26	61.46	85.53	89.56	94.69
Downstairs	49.82	39.60	12.26	55.48	44.30	86.99	89.01	93.95
Overall	78.83	82.19	68.42	81.69	80.46	93.74	94.68	96.32

in sitting activity). Finally, in case of walking activity, the results of the proposed scheme were 13.89%, 3.87%, 7.55%, and 5.77% higher than those which had been obtained by basic features + RF, logistic regression, J48, and multilayer perceptron methods. However, for this activity, the result of the proposed method has been lower than PCA + ECDF by extents of 1.09%. The above results confirmed the results which had been obtained by using deep learning-based methods which showed that the proposed algorithm caused a great accuracy improvement in recognizing challenging activities (i.e. downstairs and upstairs). On the other hand, for other non-challenging activities although the proposed method showed a slight accuracy increase or decrement compared to existing methods, but the recognition accuracies belonging to this method still remained within the acceptable range.

5 Conclusion

In recent years, deep neural networks have been widely utilized for human activity detection which the most famous among them is convolutional neural network (CNN). Despite the considerable potential of CNNs in recognizing human activities, unfortunately, such networks face with accuracy saturation phenomenon which hampers their performance in real-world applications. In this paper, a new structure was introduced to address this problem based on a combination of long short-term memory (LSTM) and residual network structures. The performance of the proposed structure was evaluated on two real data sets contained the recorded signals belonged to six human activities including walking, jogging, upstairs, downstairs, sitting, and standing for the first dataset. The second dataset contained walking, jogging, stairs, sitting, standing, eating soup, eating sandwich, and eating chips. Two different scenarios were adopted to compare the performance of the proposed method with two main categories of existing techniques. In the first scenario, the performance of the proposed method was compared with those of its own family, all based on deep learning. The obtained results showed that for the first dataset, the proposed scheme distinguished both of downstairs and upstairs (as the most challenging activities) almost 5% better than its closest deep based alternative. For the second data set improvements were 5 and 1.65% for those results which had been obtained for eating sandwich and eating chips respectively. On the other hand, the performances of the proposed method and its deep based alternatives had no meaningful difference among four other (i.e. non-challenging) activities for both datasets. The second scenario was dedicated to comparing the performance of the proposed structure and non-deep techniques. The results obtained in this scenario also indicated the superiority of the proposed method against five well known non-deep techniques in recognition of challenging activities for the first dataset. The obtained results showed that the proposed scheme distinguished downstairs and upstairs (as the most challenging activities) almost 38 and 28% better than its closest feature-based alternative. Similar to the previous scenario, the performance of the proposed method and its alternatives had no meaningful difference and both are in acceptable range when they were examined on the other four non-challenging activities. Based on the above analyses it may be concluded that the proposed structure has considerable potential to be used as a low-cost and non-invasive diagnostic modality to implement as mobile software.

Acknowledgements Both co-authors testify that our article entitled "Mobile Sensor Based Human Activity Recognition: Distinguishing of Challenging Activities by Applying Long Short-Term Memory Deep Learning Modified by Residual Network Concept" has not been published in whole or in part elsewhere is not currently being considered for publication in another journal and both authors have been personally and actively involved in substantive work leading to the manuscript, and will hold themselves jointly and individually responsible for its content.

Funding This work was not funded.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Tahavori F, Stack E, Agarwal V, Burnett M, Ashburn A, Hoseini tabatabaei SA, Harwin W. Physical activity recognition of elderly people and people with parkinson's (PwP) during standard mobility tests using wearable sensors. In: 2017 international smart cities conference (ISC2). Wuxi, ChinaSept, 14–17 September 2017; 2012. p. 403–407.
2. Dernbach S, Das B, Krishnan NC, Thomas LB, Cook JD. Simple and complex activity recognition through smart phones. In: 2012 eighth international conference on intelligent environments, Guanajuato, Mexico, 26–28 June 2012; 2012. p. 214–221.
3. Foerster F, Smeja M, Fahrenberg J. Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. Comput Hum Behav. 1999;15(5):571–83.
4. Aminian K, Robert Ph, Buchser EE, Rutschmann B, Hayoz D, Depairon M. Physical activity monitoring based on accelerometry: validation and comparison with video observation. Med Biol Eng Compu. 1999;37(3):304–8.
5. Bao L, Intille S. Activity recognition from user-annotated acceleration data. In: International conference on pervasive computing,

pervasive 2004: pervasive computing, 1–17, Linz and Vienna, Austria, 21–23 April 2004; 2004.

6. Plotz P, Hammerla NY, Olivier P. Feature learning for activity recognition in ubiquitous computing. In: IJCAI'11 Proceedings of the twenty-second international joint conference on artificial intelligence, vol. 2, Barcelona, Spain, 16–22 July 2011; 2011. pp. 1729–1734.
7. He Z, Jin L. Activity recognition from acceleration data based on discrete consine transform and SVM. In: 2009 IEEE international conference on systems, man and cybernetics, San Antonio, TX, USA, 11–14 October 2009; 2009. p. 5041–5044.
8. Kim Y-J, Kang B-N, Kim D: Hidden markov model ensemble for activity recognition using tri-axis accelerometer. In: 2015 IEEE international conference on systems, man, and cybernetics, Kowloon, China, 9–12 October 2015; 2015. p. 3036–3041.
9. Lee S-M, Yoon SM, Cho H. Human activity recognition from accelerometer data using convolutional neural network. In: 2017 IEEE international conference on big data and smart computing (BigComp), Jeju, South Korea, 13–16 February 2017; 2017. p. 131–134.
10. Ha S, Choi S. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. Vancouver, BC, Canada, 24–29 July 2016; 2016. p. 381–388.
11. Singh D, Merdivan E, Psychoula E, Kropf J, Hanke S, Geist M, Holzinger A. Human activity recognition using recurrent neural networks. In: International cross-domain conference for machine learning and knowledge extraction, Reggio, Italy, 29 August 2017; 2017. p. 267–274.
12. Ordóñez FJ, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors. 2016;16:115.
13. Cho K, van Merrienboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: encoder–decoder approaches, association for computational linguistics. In: Proceedings of SSST-8, eighth workshop on syntax, semantics and structure in statistical translation, Doha, Qatar, October 2014; 2014. p. 103–111.
14. Arifoglu D, Bouchachia A. Activity recognition and abnormal behaviour detection with recurrent neural networks. Procedia Comput Sci. 2017;110(86–93):12.
15. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
16. Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. J Stat Plan Inference. 2000;90(2):227–2441.
17. NadeemHashmi S, Gupta H, Mittal D, Kumar K: A lip reading model using CNN with batch normalization. In: 2018 eleventh international conference on contemporary computing (IC3), 1–8, New Delhi, India, August 2018; 2018.
18. LeCun Y, Bottou LB, Orr G, Müller K-R (2002) Efficient backprop, neural networks: tricks of the trade. Lecture notes in computer science, vol. 1524. Springer, Berlin; 2002.
19. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. J Mach Learn Res. 2010;9:249–56.
20. Zhang Y, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). New Orleans, LA, USA, 5–9 March 2017; 2017. p. 4845–4849.
21. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. IEEE Trans Neural Netw. 1994;5(2):157–66.
22. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; 2016. p. 770–778.
23. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. ACM SIGKDD Explor Newslett. 2010;12(2):74–82.
24. Bottou L. Large-scale machine learning with stochastic gradient descent. In: International conference on computational statistics (COMPSTAT); 2010. p. 177–186.
25. Ignatov A. Real-time human activity recognition from accelerometer data using convolutional neural networks. Appl Soft Comput. 2017;62:915–22.
26. Zeng M, Nguyen LT, Yu B, Mengshoel OJ, Zhu J, Wu P, Zhang J. Convolutional neural networks for human activityrecognition using mobile sensors. In: 6th international conference on mobile computing, applications and services, Austin, TX, USA, 6–7 November 2014; 2014. p. 197–205.
27. Plotz T, Hammerla NY, Olivier P. Feature learning for activity recognition in ubiquitous computing, 1729–1735. In: Proceedings of the twenty-second international joint conference on artificial intelligence, Barcelona, Catalonia, Spain, 16–22 July 2011; 2011.
28. Witten H, Frank EI. Data miningpractical machine learning tools and techniques, a volume in the morgan kaufmann series in data management systems; 2011.