http://jhmi.sums.ac.ir
J Health Man & Info 2020, 7(4), 228–239

JHMI
Journal of Health Management and Informatics

Original Article

Residual Network of Residual Network: A New Deep Learning Modality to Improve Human Activity Recognition by Using Smart Sensors Exposed to Unwanted Shocks

Mohammad Javad Beirami[1]*, Seyed Vahab Shojaedini[2]

[1]Faculty of Electrical, Biomedical and Mechatronics Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[2]Associate Professor of Biomedical Engineering, Iranian Research Organization for Science and Technology, Tehran, Iran

Abstract

Introduction: Recently, smartphones have been vastly utilized in monitoring the daily activities of people to check their health. The main challenge in this procedure is to distinguish similar activities based on signals recorded by using sensors mounted on smartphones and smartwatches. Although deep learning approaches have better addressed the above challenge than alternative methods, their performance may be severely degraded, especially when the mounted sensors receive disturbed signals due to smartphones and smartwatches not being in a fixed position.

Methods: In this article, a new deep learning structure is introduced to recognize challenging human activities by using smartphones and smartwatches, even when the recorded signals are noisy due to the sensors being unstable. In the proposed structure, the residual network of residual network (i.e. ROR) is engaged as a new concept inside the deep learning architecture, which provides greater stability against either disturbed or noisy signals.

Results: The performance of the proposed method is evaluated on recorded signals from smartphones and smartwatches and compared with the state of art techniques containing deep learning and classic (non-deep) schemes. The obtained results show that the proposed method may improve the recognition parameters at least 1.79 percent against deep alternatives in distinguishing challenging activities (i.e. downstairs and upstairs). These superiorities reach at least 32.86 percent for classic methods, which are applied on the same data.

Conclusion: The effectiveness of the architecture in recognizing either challenging or non-challenging activities in the presence of unwanted cell phone shocks demonstrates its potential to be used as a mobile application for human activity recognition.

Keywords: Human Activity Recognition, Smartphone, Deep Learning, Gradient Flow, Residual Networks of Residual Networks.

Article History:
Received: 17 February 2020
Accepted: 03 November 2020

Please cite this paper as:
Beirami MJ, Shojaedini SV. Residual Network of Residual Network: A New Deep Learning Modality to Improve Human Activity Recognition by Using Smart Sensors Exposed to Unwanted Shocks. J Health Man & Info. 2020; 7(4): 228-239.

***Correspondence to:**
Mohammad Javad Beirami,
Faculty of Electrical, Biomedical and Mechatronics Engineering,
Qazvin Branch, Islamic Azad University, Qazvin, Iran
Email: Mj.Beirami@qiau.ac.ir

Background

Recently IoT (Internet of Things) has become an effective tool to help humans in their daily life based on its enormous potential of recognition and communication (1). Among various options of this technology, human activity recognition systems have attracted much attention based on this fact that such systems make it possible to promote the quality of some applications, including assistive living and rehabilitation.

Computer vision-based techniques are a group of solutions designed for human activity recognition. Unfortunately, the performance of this category of techniques for activity recognition is highly dependent on the controllability of the application environment. Therefore, such techniques may considerably fail in an environment with the clutter and variable lighting (2, 3). Based on such limitations, wearable sensors have been substituted, which offer a practical and low-cost methodology to analyze human activities.

Nowadays, smartphones are widely available to everyone, so using the signal recorded by the sensors mounted on this type of phones for human activity recognition may be utilized as an easy and cheap modality in the IoT-healthcare domain. The most challenging problem in this system is the similarity between the recorded signals which arise from different activities of a person. This phenomenon makes the recognition of human activity a challenging problem classification paradigm, which severely degrades the recognition rate.

From one point of view, the human activity classification approaches may be divided into two major types, including classical and Neural Network-

based methods. In the domain of classic methods, the Support Vector Machines (SVM) are one of the most widely used techniques for distinguishing human activities (4). Furthermore, several features such as DCT (4), energy (5-7), entropy (6), correlation (5-7) and Fourier features (8) have been utilized in SVM classifier in order to help it to overcome the challenges of human activity recognition system. Unfortunately, the performance of this method is seriously degraded when the data set has more noise (for example when the smartphone is not completely fixed in the individual's pocket).

The random forest algorithm is another examined method in order to address the limitations of human activity classification (9). This method has been frequently used in combination with handcrafted features (10, 11). This strategy, however, has some drawbacks, including poor interpretability, high probability of overfitting, and the need to choose the number of trees.

Some sophisticated methods try to overcome the limitations of human activity recognition problem by modeling it as a state-machine paradigm. Hidden Markov Model (HMM) based algorithms may be considered as representative for this category of solutions (12). Although this technique has led to better performance than many of its alternatives, it is highly sensitive to the quantity and quality of the extracted features from the activity records.

In the last few years, deep learning approaches have attracted much attention to the application of activity recognition, thank their ability to classify poor separable data. These networks are based on the learning of multiple levels of representations of the data. Such a learning strategy, in parallel with their specific structure, which includes several processing layers, enables them to considerably promote the classification performance (13).

One of the earliest works in this field was performed by Jiang et al. which introduced a CNN-based method for human activity recognition by using the gyroscope, total acceleration, and linear acceleration signals (13). In some researches, Deep Belief Networks (DBNs) have been utilized for activity recognition when the sensors were repeatedly exposed to unwanted movements (14). The results obtained from these investigations showed that each limitation in the placement of the sensors in a fixed position might cause a significant decrease in the performance of an activity recognition system. Some sophisticated researches tried to incorporate temporal dependency of human activity signals in training deep networks. This aim led to applying Recurrent Neural Networks (RNNs) for this purpose (15, 16). Some other solutions have utilized Long-short Term Memory (LSTM) concept in parallel with classical CNN to extract the temporal and local features simultaneously (9). Despite all the mentioned research and improvements that they have made, vanishing or exploding of the gradient is still an important weakness for neural network-based approaches. Due to the fact that the learning process in all of the above networks is performed by gradient propagation, the gradient vanishing or exploding may considerably hamper the application of deep structures in human activity recognition.

In this paper, a new method is introduced to improve the gradient propagation in temporal deep neural networks. The proposed scheme tries to overcome the gradient vanishing/exploding problem by using the concept of residual network of residual network (ROR). This idea may optimize shortcut connections by adding them in level by level structure in deep network architecture. Such a strategy utilizes shortcuts to jump over some layers; therefore, the problems related to gradient flow may be addressed by reusing activations from a previous layer until the layer next to the current one has learned its weights. Using the proposed method makes it possible to build deeper structures that exhibit greater robustness against noise and disturbance due to the extraction of more abstract features. While in the conventional case, the degradation problem may seriously hamper the performance of deeper neural networks. Therefore, it is expected that the proposed method may lead to better results, especially for those signals containing unwanted smartphone shocks thanks to its potential in constructing deeper architectures and extracting more robust features.

The paper is organized as follows. In section 2, the proposed approach is analyzed, which consists of exploring the datasets and description of the proposed architecture. In section 3, the performance of the proposed architecture is evaluated by comparing its results with hoset of other deep learning structures. In section 4, the obtained results from the proposed method are compared with the findings of the category of techniques that do not operate on deep learning basis. The conclusion is presented in the last section of the paper.

Materials and Methods

This section reviews the details of the proposed architecture to improve the performance of deep neural networks in human activity recognition. To that end, firstly, the CNN structure was reviewed,

and then the residual network concept was applied to increase the accuracy of HAR. Then, the proposed structure was demonstrated based on residual networks of residual network idea.

Convolutional Neural Network

A significant portion of the capability of CNNs is due to the presence of a part in their structure, which is known as convolutional layer. Basically, the task of this layer is to extract the local features from input (9). Specifically, low-level features can be obtained in lower layers, and high-level features may be extracted as the layers are deepened. Let the input signal be considered as:

$$x = \begin{bmatrix} ax_1 1 & \cdots & ax_1 m \\ \vdots & \ddots & \vdots \\ ax_n 1 & \cdots & ax_n m \end{bmatrix} \quad (1)$$

Where x may represent the window of size $m \times n$, which m refers to window width and n shows window length. The CNN maps the input x to feature space of z, as described in equation (2). In the same manner, $z_{i,j}$ is feature map in location (i , j).

$$z_{i,j}^{l,k} = [z_{i,1}^{l,k}, z_{i,2}^{l,k}, \dots, z_{i,n}^{l,k}] \quad (2)$$

Therefore, each feature map in *k-th* layer and *l-th* feature map may be computed as:

$$z_{i,j}^{l,k} = \sigma(\sum_{k'=1}^{k'} \sum_{x=1}^{X} \sum_{y=1}^{Y} w_{x,y,k'}^{l-1,k'} z_{i,j+y-1}^{l-1,k'} + b^{l-1,k}) \quad (3)$$

In which σ and k' are activation function and number of feature maps, respectively in *(l-1)-th* layer with kernel size of X and Y. Also, we present b as bias term (Figure 1).

Eventually, these layers end up in the final part of the network, which is called fully connected layer, which maps the extracted features into distinguished classes. For this purpose, the Dense layer, the nodes of which represent activity classes has been utilized.

Residual Networks of Residual Network

It has been shown that deeper networks usually lead to better generalization and performance compared to classic networks, thanks to their ability in the modelling of nonlinearities of data (17). On the other hand, most recent works in the HAR domain show that such structures may extract more abstract features, which leads to more accurate classification (18). Despite these advantages, deep structures also have the annoying drawback that they frequently face the problem of gradient vanishing/exploding. It has been shown that such a problem may seriously hamper the convergence of the network and consequently limits its performance in classification (19-21). Therefore, several structures have been proposed to solve or reduce this limitation. An effective approach to address this problem is batch normalization (BN) (22) in which internal covariance reduction along with speed increment relatively is achieved by a well-defined function as bellow (16, 23-25):

$$BN(z) = \gamma \frac{z - E(z)}{\sqrt{(Var(z) + \varepsilon)}} + \beta \quad (4)$$

In the above equation, γ and β are learning parameters. However, batch normalization itself is not enough to perform proper optimization. It was shown that the accuracy saturation phenomenon caused a rapid degradation in the performance of deep networks (20, 26, 27). To address this problem, we proposed residual architecture. The idea behind this approach is deploying shortcut connections along with BN to perform an identity mapping. These skipping connections may lead to easier gradient flow in deep architectures and largely overcome the vanishing/exploding problem (28). As a result, this new structure allows deep network training without suffering from the mentioned poor optimization and degradation (20). The mentioned direct link between the layers in the residual network may be

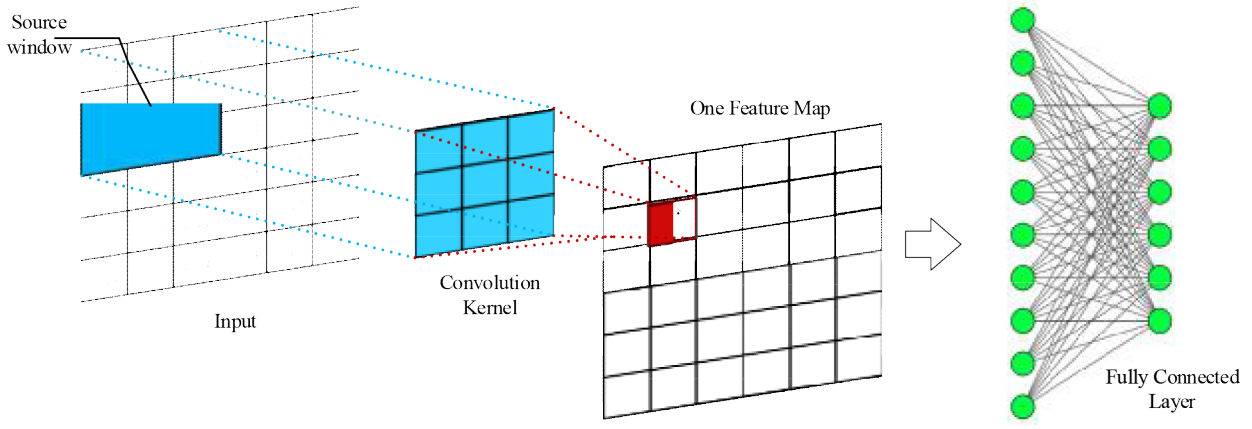

Figure 1: Feature extraction with X and Y kernel size. Fully connected layer with softmax function, maps the extracted features into six activity classes

demonstrated as:

$$y = F(x, w) + x \quad (5)$$

In the above equation, x and y are input and output of the presented layer, respectively. Also, $F(x, w)$ is the function belonging to the block between shortcut connections which may consist of one or more layers. The ResNet architecture is displayed in Figure 2.

The ResNet of ResNet (i.e. ROR) is an idea to improve the performance of ResNet, which is based on optimizing the shortcut connections by adding them in level by level structure (27). Figure 3 shows the proposed architecture in which, at first, the basic residual block was built in the form of BN+CNN+BN structure. Then, a convolution layer was added; afterwards, the same version of this structure was copied. In the next step, the second level of shortcut connection in parallel with convolution mapping was applied, which is called residual group. In the same manner, more residual groups were made, but the important point is that different sizes of convolution filters were utilized (i.e. sizes of 16, 32, and 64 for the first, second, and third residual groups, respectively). Finally, the last shortcut connection as a convolution mapping connected the input to the output of the network, which is called first-level connection. The mentioned multi-level discipline firstly improved the optimization by transforming the learning of $F(x, w)$ to $F(x, w) + x$ and secondly, the gradient propagating between blocks was performed easier.

Results

To evaluate the performance of the proposed method, we applied it on two distinct datasets. The first dataset contained over 1 million 3-axis accelerometer data which were recorded by using some popular brands of smartphones such as HTC Hero, Motorola Backflip, and Nexus (10) from several activities of 36 volunteers. This dataset is known as Wireless Sensor Data Mining (WISDM) and was prepared by the University of Fordham. WISDM included six different types of activities which were walking, jogging, sitting, standing, upstairs, and downstairs (10). The second data set was prepared by using LGG Watch running Android Wear smartwatches which were all equipped with 3-axis accelerometer and fastened on the 51 volunteers' wrists. This dataset was composed of the recorded signals from eight activities, five of which re common with the first dataset and the others were different. Walking, jogging, stairs, standing, sitting, eating soup, eating sandwich, and eating chips were activities whose information was contained in the second dataset (29). The detailed specifications of the test data are shown in Table 1.

One important limitation in the data recording procedure was the conditions of placement of the phone and the watch in the pockets and wrists of people

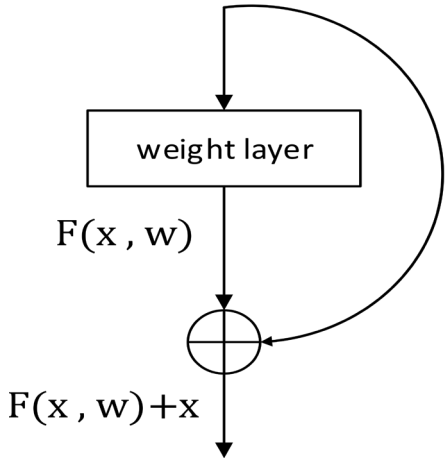

Figure 2: The residual network which including parameter-free connections (identity shortcuts) to connect the input of layer to the output.

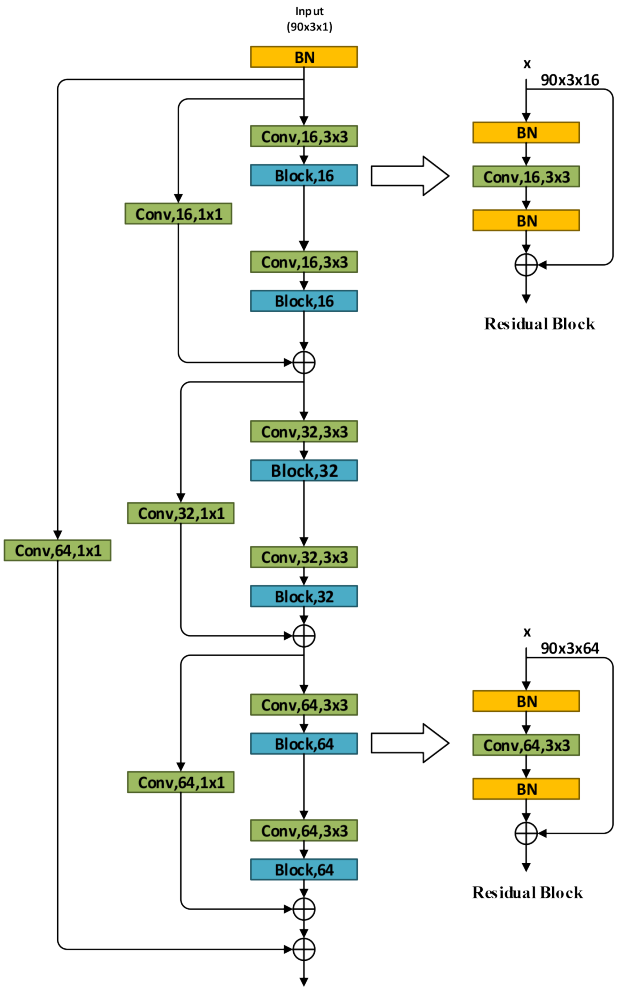

Figure 3: Proposed ROR structure. This architecture includes three shortcut levels; the first level is a basic ResNet level, which contains an identity shortcut connection. The second level is over each first level and finally third level that includes a shortcut to all levels.

Table 1: Description of details of WISDM dataset

Dataset	Activities	Class distribution	Sample Rate	Measurement Time	Number of Users	Number of data
First dataset	Walking	Walking: 38.6%	20 Hz	10 min	36	1,098,207 sample
	Jogging	Jogging: 31.2%				
	Upstairs	Upstairs:11.2%				
	Downstairs	Downstairs: 9.1%				
	Sitting	Sitting: 5.5%				
	Standing	Standing: 4.4%				
Second dataset	Walking	walking: 12.5%	20 Hz	3 min	51	1,676,282 sample
	Jogging	Jogging: 12.5%				
	Stairs	Stairs:12.5%				
	Standing	Sitting: 12.5%				
	Sitting	Standing: 12.5%				
	EatingSoup	EatingSoup: 12.5%				
	Eating Sandwich	Eating Sandwich: 12.5%				
	Eating Chips	Eating Chips: 12.5%				

under testing. The smartphones and smartwatches were not fixed in the pockets and wrists deliberately; therefore, the captured data included considerable components due to unwanted motions and excessive shocks. This fact may be considered as high recording noise and axis displacement between the users for the same activities. For example, Figure 4 shows axis displacement for "sitting" and "standing" between three different users, which is not expected for such a stationary activity. Such a diversity between several records from the same activity may hamper the ability of the algorithms, especially in recognizing more exciting activities. Furthermore, the first dataset faced another limitation, which in turn hampered its recognition results. Unfortunately, the volumes of data belonging to several activities were not equal in this dataset because of two reasons. Firstly, some users did not perform some of the activities due to their physical restrictions. Secondly, some activities (i.e. sitting and standing) were limited to only a few minutes because it was expected that the data would remain almost constant over a long time. Figure

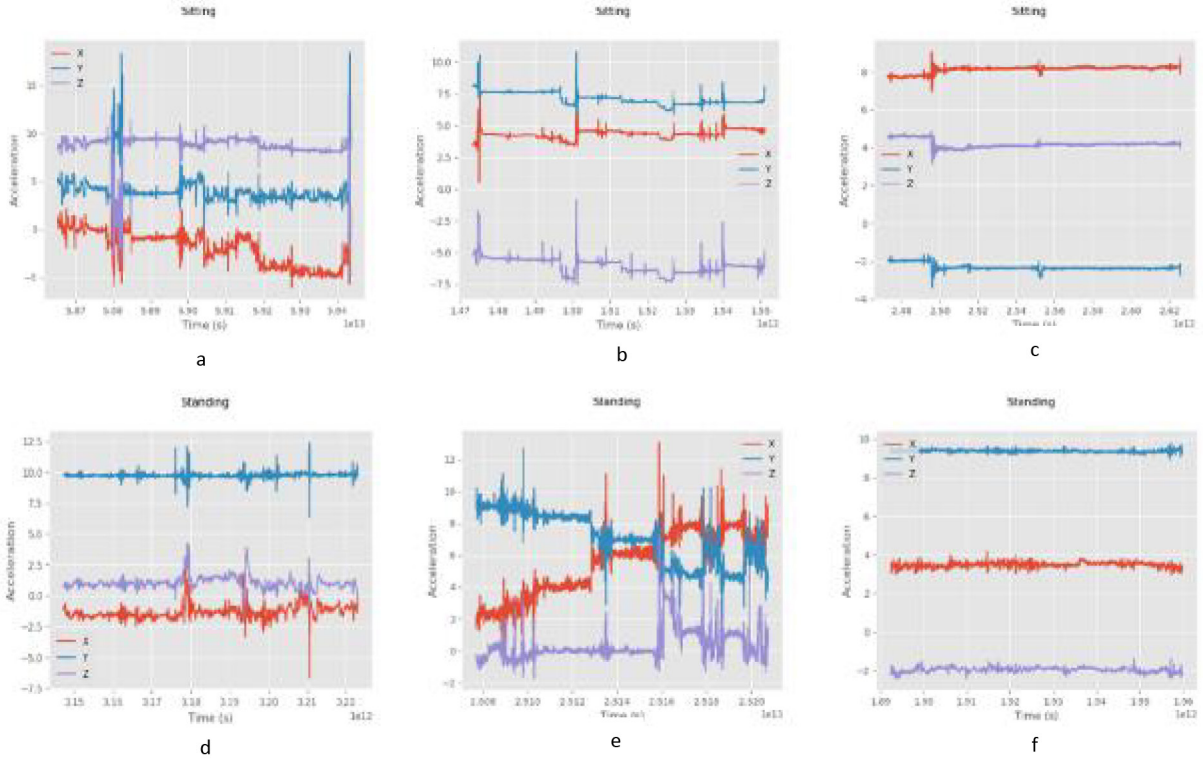

Figure 4: Six recorded signals belonging to six deferent users. a, b and c are "sitting" and d, e and f are "standing" signals. As the figures shows, axis orders are deferent for thiese stationary activities between users.

4 shows some signals which were captured from activities in WISDM, which clearly demonstrates lack of significant differences between signals recorded from different activities (e.g. standing and sitting). This fact illustrates why recognizing these activities may be considered as a challenge in the human activity recognition paradigm.

The proposed structure and its alternatives were implemented in the TensorFlow framework by using Keras API. The training procedure was performed by Tesla T4 GPU with 2560 CUDA cores. In this study, in order to assess the effectiveness of the proposed method, we implemented two alternative methods, including Convolutional Neural Network (i.e. CNN) (9, 16) and Residual based modified Convolutional Neural network (20) *which is called as ResNet for brevity in rest of this article.* The most important common point of the above alternative methods is that both of them are based CNN concept; therefore, they may be considered as members of the same family with the proposed method.

Table 2 shows the specifications of the best-fitted structures for CNN, ResNet, and proposed algorithms, which led to the best results for distinguishing activities.

Four standard factors were measured to evaluate the performance of each method. The measured parameters consisted of True Positive (TP) which shows the number of correctly identified activities, True Negative (TN) which shows the wrong activities which were rejected correctly, and False Positive (FP) which is the number of false detections and False Negative (FN) which shows the number of missed activities. Finally, Recall and Precision were estimated by using the following formulas for all examined methods to compare their effectiveness in recognizing activities.

Recall means the probability that an activity is identified if it exists. This parameter is defined as:

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

Precision illustrates the correct percentage of activity recognition; in other words, it reports that from a constant numbers of recorded signals, how many have been actually recognized. This parameter is calculated as:

$$Precision = \frac{TP}{FP+TP} \quad (7)$$

The results of the proposed method and its two deep base alternatives are shown in Tables 3 to 8. Firstly, the performance of CNN was evaluated in distinguishing several activities of the two datasets.

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

	CNN	**Value**		**ResNet**	**Value**		**ROR**	**Value**
	Network parameter			**Network parameter**			**Network parameter**	
	Input size	**90×3**		**Input size**	**90×3**		**Input size**	**90×3**
Layer	Parameter	Value	Layer	Parameter	Value	Layer	Parameter	Value
1-2	Feature map Filter size Activation function	16 3×3 ReLU	1-2	Feature map Filter size Activation function	16 3×3 ReLU	1-4	Feature map Filter size Activation function	16 3×3 ReLU
3-4	Feature map Filter size Activation function	32 3×3 ReLU	3-4	Feature map Filter size Activation function	32 3×3 ReLU	5-8	Feature map Filter size Activation function	32 3×3 ReLU
			5-6	Feature map Filter size Activation function	64 3×3 ReLU	9-12	Feature map Filter size Activation function	64 3×3 ReLU
						13-16	Feature map Filter size Activation function	128 3×3 ReLU

Table 3: Results of the first dataset classification by using CNN

	CNN	**Predicted class**						**Recall**
	Activity	**Walking**	**Jogging**	**Sitting**	**Standing**	**Upstairs**	**Downstairs**	
Class	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%
Precision		97.82%	97.77%	97.43%	100%	83.1%	88.58%	93.74% 94.11%

Table 4: Results of the second dataset classification by using CNN

CNN		Predicted class								Recall
	Activity	Walking	Jogging	Stairs	Sitting	Standing	Soup	Sandwich	Chips	
Class	Walking	841	3	85	0	0	0	1	0	90.48%
	Jogging	6	905	3	0	0	0	0	0	99.01%
	Stairs	87	8	792	6	11	4	7	7	85.90%
	Sitting	4	1	4	762	38	37	35	65	80.55%
	Standing	2	0	3	73	817	18	11	39	84.84%
	Soup	0	0	2	18	32	734	77	67	78.92%
	Sandwich	2	0	3	48	43	111	546	179	58.58%
	Chips	1	0	12	67	37	129	205	455	50.22%
Precision		89.18%	98.69%	87.61%	78.23%	83.54%	71.05%	61.90%	56.03%	78.56% 78.28%

Table 5: Results of the first dataset classification by using ResNet

ResNet		Predicted class						Recall
	Activity	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	
Class	Walking	1847	0	0	0	26	13	97.93%
	Jogging	9	1499	0	3	13	0	98.55%
	Sitting	0	0	266	0	0	0	100%
	Standing	0	0	2	206	7	0	95.81%
	Upstairs	7	17	3	0	503	16	92.12%
	Downstairs	9	4	0	0	29	404	90.58%
Precision		98.66%	98.62%	98.15%	98.56%	87.02%	93.30%	95.83% 95.72%

Table 6: Results of the second dataset classification by using ResNet

ResNet		Predicted class								Recall
	Activity	Walking	Jogging	Stairs	Sitting	Standing	Soup	Sandwich	Chips	
Class	Walking	890	3	41	0	0	0	1	0	95.19%
	Jogging	6	907	0	1	0	0	0	0	99.23%
	Stairs	60	11	814	5	11	3	9	9	88.28%
	Sitting	1	4	4	809	14	31	28	55	85.52%
	Standing	0	0	2	67	812	16	25	41	84.32%
	Soup	0	0	5	15	17	693	102	98	74.51%
	Sandwich	3	0	1	43	29	96	580	180	62.23%
	Chips	3	1	9	54	20	85	205	529	58.39%
Precision		92.42%	97.95%	92.92%	81.39%	89.92%	75.00%	61.05%	58.00%	80.96% 81.08%

Table 8: Results of the first dataset classification by using ROR

ROR		Predicted class						Recall
	Activity	Walking	Jogging	Sitting	Standing	Upstairs	Downstairs	
Class	Walking	1852	1	0	0	25	8	98.19%
	Jogging	0	1512	0	0	7	2	99.40%
	Sitting	0	0	266	0	0	0	100%
	Standing	0	0	1	212	0	2	98.14%
	Upstairs	6	12	2	0	510	16	94.32%
	Downstairs	2	3	0	1	26	414	92.37%
Precision		99.57%	98.95%	98.89%	99.53%	89.79%	93.66%	97.07% 96.73%

As demonstrated in the Tables 3 and 4, this approach has obtained fully acceptable results on those activities in which their recorded signals were not so similar to each other. (i.e. all activities except upstairs and downstairs for the first dataset and three eating activities for the second dataset).

However, the performance of this network was dropped when it was applied to recognize challenging activities (i.e. downstairs, upstairs, eating soup, eating sandwich, and eating chips which caused similar signals) in such a way the recall for these two activities for the first dataset were obtained 85.53 and 86.99 percent for upstairs and downstairs, respectively, and for the second one 78.92, 58.58, and 50.22 percent for eating soup, sandwich and chips, respectively. These values were at least 10 percent lower than those which have been obtained for other activities. Also, precision for upstairs and downstairs was 83.1 and 88.58 percent, respectively for the first dataset, and for the second one it was 71.05, 61.90, and 56.03 percent for three eating actions. Also, these results showed the precision of detecting downstairs and upstairs was at least 10 percent lower than what was obtained for other activities.

Table 5 shows the results obtained from the modification of CNN by using shortcut connections (i.e. ResNet). These results demonstrate although such modification may marginally improve the recall and precision in distinguishing similar activities (i.e. 6.59 and 3.59 percent improvement for recall in recognizing upstairs and downstairs and 3.92 and 4.72 percent improvement in the precision of recognizing these activities), these improvements were not enough to make them completely acceptable. Also, Table 6 shows 3.65 and 8.17 improvement in eating sandwich and chips; however, it drops 4.41 accuracy for eating soup. Furthermore, these tables show that other non-challenging activities for both datasets have had no meaningful difference from the results which had been obtained from basic CNN (e.g. these differences were about 1 percent for both recall and precision).

Finally, Tables 7 and 8 demonstrate that the proposed method had significantly increased the obtained accuracies for challenging activities compared to previously examined schemes. The results showed that the Residual Network of Residual Network (i.e. ROR) concept might improve the recall against the basic CNN network by extents of 8.79 and 5.38 percent for upstairs and downstairs activities for the first dataset and 4.72 and 7.62 percent for eating sandwich and chips for the second dataset, respectively. Also, our proposed scheme compensated for the accuracy drops for eating soup in ResNet (i.e 1 percent drop which is not meaningful). Furthermore, it promoted precision by extents of 6.69 and 5.08 percent against the above algorithms for the first dataset. For the second dataset, the promotions were equal to 6.24 percent for eating soup and 3.58 percent for eating chips. In recognizing the activity of eating sandwich, no meaningful change was observed (less than 1 percent). In a similar manner, recall improvements were obtained as 2.2 and 1.79 percent for the same activities in the first dataset compared to the ResNet scheme. For the second dataset, improvement were 1.07 percent for eating sandwich and 3.45 percent for eating soup. The above improvements have been 2.77 and 0.36 percent for precision in upstairs and downstairs activities for the first dataset and 2.29 percent for eating soup in the second one. For the two other eating actions, the change was around 0.5 percent.

As some of activity datasets are imbalanced, as described in Table 1, overall accuracy is not an appropriate criterion for evaluating the methods. Thus, along with Precision and Recall calculated in Tables 3 to 8, F1 score was also calculated as a measure that combines Recall and Precision in the form of the following equation:

$$F_1 = 2 \times w \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

In which w is defined for each classe as the ratio of the members of that class versus total data. Table 9 made use of F-score to show how the proposed method promoted the results.

Cross Entropy Analysis

It is essential to understand that cross-entropy loss function, which we have used in this contribution, measures the performance of the classification model in the probabilistic paradigm. To obtain a decisive decision, the output of this framework was transformed into a range of [0,1] by utilizing the softmax function. To evaluate the loss for each method as a measure of confidence for its classification, the following equation was applied:

$$L = -y.\log(y') \quad (9)$$

In which, L is loss, y is actual and y' is predicted value.

For instance, in the prediction between three classes whose probabilities were equal to (0.5 0.25 0.25), the first class was chosen (i.e., maximum probability), but it is clear that although the network performed correct classification,it had a poor confidence levele.g. based on low entropy of the result) (30). Based on this logic, the Loss parameter

Table 9: Results of the second data classification by using ROR

ROR		Predicted class								Recall
	Activity	Walking	Jogging	Stairs	Sitting	Standing	Soup	Sandwich	Chips	
Class	Walking	894	0	41	0	0	0	0	0	95.61%
	Jogging	6	906	2	0	0	0	0	0	99.12%
	Stairs	60	8	830	4	4	2	7	7	90.02%
	Sitting	3	4	2	810	13	30	40	43	85.62%
	Standing	1	0	1	70	825	10	19	37	85.67%
	Soup	0	0	1	14	14	725	98	78	77.96%
	Sandwich	0	0	2	44	23	83	590	190	63.30%
	Chips	4	0	8	58	20	88	204	524	57.84%
Precision		92.35%	98.69%	93.57%	81.00%	91.77%	77.29%	61.59%	59.61%	81.89% 81.98%

was computed for all activities, and all methods and the results are shown in Tables 10 and 11. As Table 10 shows, ResNet architecture had lower loss against CNN in four of six activities, but slightly higher loss for the other two activities (0.03 and 0.08 for walking and downstairs, respectively). Our proposed method showed lower loss for five activities between all the examined methods, but a bit higher loss only for walking activity. These results indicated that our proposed method was more approprite for the classification of activities, especially for jogging, upstairs, and downstairs.

Table 10: F1 score for WISDM datasets

Method	Dataset	F_1 Score
CNN	1	0.9537
ResNet	1	0.9681
ROR	1	0.9769
CNN	2	0.7842
ResNet	2	0.8102
ROR	2	0.8193

For the second dataset, the results showed considerable loss reduction for most of activities (e.g. 0.266 for walking, 0.0913 for jogging, 0.3154 for stairs, 0.932 for eating sandwich, and 1.3894 for eating chips). However, the loss for standing and eating soup increased about 0.2252 and 0.6647, respectively (Table 12).

Discussion

In the previous section, we compared the proposed scheme to several existing deep based methods on the WISDM dataset. These comparisons demonstrated higher performance of the proposed algorithm against alternative architectures. The common aspect of all those algorithms was that all of them belong to deep neural networks family. Consequently, 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, the following activity recognition methods were applied on the first dataset:

(i). Basic features + RF: Basic features included hand-crafted features, which is the most traditional approach for activity recognition. The random forest is also known to be a good method for activity recognition task, and a combination of these two methods, which is called Basic features + RF, is one of our comparison alternatives (10, 11).

(ii). PCA+ECDF: Principal Component Analysis (PCA) is a method for decorrelation and dimensionality reduction of data (31). PCA is a method of feature learning which transforms data into uncorrelated linear data, which is called principal components. Applying this method based don the Empirical Cumulative Distribution function is called

Table 11: Test Loss of the first dataset for three examined deep learning methods.

	Test Loss			
	Activity	CNN	ResNet	ROR
Class	Walking	0.07	0.1	0.129
	Jogging	0.138	0.063	0.019
	Sitting	0.05	$1.1e^{-0.5}$	$8.8e^{-0.6}$
	Standing	0.20	0.19	0.159
	Upstairs	0.375	0.36	0.23
	Downstairs	0.64	0.72	0.44

Table 12: Test Loss of the first dataset for three examined deep learning methods.

	Test Loss			
	Activity	**CNN**	**ResNet**	**ROR**
Class	Walking	0.4850	0.2996	0.219
	Jogging	0.1419	0.1325	0.0506
	Stairs	0.7323	0.7361	0.4169
	Sitting	0.9892	0.8663	1.014
	Standing	0.6123	0.8809	0.8375
	Soup	0.7433	1.59	1.408
	Sandwich	1.151	1.41	0.219
	Chips	1.44	1.70	0.0506

PCA+ECDF and has been utilized for comparisons of this section (29, 32).

(iii). Logistic Regression: It is a classifier for supervised learning using the Logistic function, which has been used for the classification of human activities before (10, 33).

(iii). J48: It is a classification algorithm that is based on the decision tree concept and has led to a considerable performance in human activity recognition (10, 31).

(iv). Multilayer Perceptron: It is a basic neural network-based scheme that belongs to non-deep classical neural networks and was traditionally used in researches of human activity recognition (10). Finally, the accuracy of classification was calculated both for the proposed method and all the above alternatives, as demonstrated in Table 6. This Table describes that the most superior of the proposed method against its alternatives was obtained in recognizing upstairs and downstairs activities.

This Table demonstrates that the recall parameter which was obtained for recognizing upstairs activity was 32.86%, 35.06%, 66.8%, and 34.77% better than Multilayer Perceptron, J48, Logistic Regression, and PCA+ECDF methods, respectively. In a similar manner, the precision of recognizing the above activities by using the proposed method was 30.92%, 26.08%, 47.66, and 19.75% better than same alternatives. Also, this Table shows that the recognition statistics of downstairs activity were comparatively similar to those reported for upstairs. The proposed algorithm was promoted in recall by 48.07%, 36.89%, 80.11%, and 52.77% better than its alternatives. Furthermore, 38.28%, 38.65%, 63.66%, and 18.85% precision improvement has obtained by applying the proposed scheme compared to its feature-based alternative methods (Table 13).

Corresponding to the other four activities (e.g. sitting, standing, jogging, and walking), our proposed method still showed acceptable performance in precision and recall, which were a bit better than its alternative (see Table 8). However, it is important to note that even for these non-challenging activities the superiority of the proposed scheme against alternative methods reached 42.94 percent in precision (e.g. proposed method vs. PCA+ECDF in standing activity) and 11.15 percent in recall (e.g. proposed method vs. Logistic Regression in standing activity).

The results reported in the last two sections indicate 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 performance of the proposed method showed a slight increase or decrement compared to the existing methods, the outcomes of this method remained within the acceptable range.

Conclusion

In this paper, a new method was introduced to promote the performance of deep neural networks in distinguishing human activities signals when they

Table 13: Comparison of Results of WISDM classification by using proposed method and its features based on alternatives

Activity type	**PCA+ECDF**		**Logistic Regression**		**J48**		**Multilayer Perceptron**		**ROR**	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
Walking	98.54%	92.05%	93.58%	73.94%	89.90%	87.76%	91.68%	88.73%	**98.19%**	99.57%
Jogging	95.44%	98.93%	97.95%	95.78%	96.52%	96.15%	98.33%	96.94%	**99.40%**	98.95%
Sitting	100%	95.95%	92.20%	92.53%	95.74%	96.77%	95.05%	95.37%	**100%**	98.89%
Standing	100%	56.59%	86.99%	91.94%	93.27%	96.30%	91.93%	94.04%	**98.14%**	99.53%
Upstairs	59.55%	70.04%	27.52%	42.13%	59.26%	63.71%	61.46%	58.87%	**94.32%**	89.79%
Downstairs	39.60%	74.81%	12.26%	30%	55.48%	55.01%	44.30%	55.38%	**92.37%**	93.66%

were recorded along with the unwanted mobile shocks. In this case, the sensors mounted on the smartphones and smartwatches recorded the signals of various activities with disturbance due to their deviation of orientation. The strategy of the proposed method for achieving this improvement was to address the gradient vanishing/exploding problem and digging the performance of structure by identity mapping in deep neural networks by using residual of residual (ROR) structure. Such a strategy made it possible to construct deeper networks, which led to obtaining more noise-resistant and discriminative features. The performance of the proposed architecture was evaluated on two datasets which contained walking, jogging, sitting, standing, upstairs, and downstairs activities for the first dataset and walking, jogging, stairs, sitting, standing, eating soup, eating sandwich, and eating chips for the second one. Two different scenarios were considered to compare the results of the proposed scheme and results obtained from the existing methods. In the first scenario, the performance of the proposed method was compared with the family of deep learning based approaches. Such a comparison showed that the proposed architecture could improve the confidence parameter considerably in addition to improving the recall and precision of recognizing challenging activities.

In the second scenario of evaluations, the performance of the proposed structure was compared to the existing non-deep methods to distinguish the same activities. The obtained results showed a dramatic increase in precision and recall of the proposed method in recognizing challenging activities (by extents of 19% compared to its closest alternative). Based on the above results, it may be concluded that the proposed structure has a considerable potential to be considered as a suitable activity recognition mobile software.

Acknowledgment

Both co-authors testify that our article entitled " Residual Network of Residual Network: A New Deep Learning Modality to Improve Human Activity Recognition by Using Smartphone Exposed to Unwanted Shocks" has not been published in whole or in part elsewhere and is not currently being considered for publication in another journal; 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.

Conflict of Interest: None declared.

References

1. Castro D, Coral W, Rodriguez C, Cabra J, Colorado J. Wearable-based human activity recognition using an iot approach. *Journal of Sensor and Actuator Networks*. 2017;6(4):28. doi: 10.3390/jsan6040028.
2. Zhang S, Wei Z, Nie J, Huang L, Wang S, Li Z. A Review on Human Activity Recognition Using Vision-Based Method. *J Healthc Eng*. 2017;2017:3090343. doi: 10.1155/2017/3090343.
3. Chaquet JM, Carmona EJ, Fernández-Caballero A. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*. 2013;117(6):633-59. doi: 10.1016/j.cviu.2013.01.013.
4. He Z, Jin L. Activity recognition from acceleration data based on discrete consine transform and SVM. *2009 IEEE International Conference on Systems, Man and Cybernetics*. 2009:5041-4. doi: 10.1109/ICSMC.2009.5346042.
5. Wang S, Yang J, Chen N, Chen X, Zhang Q. Human activity recognition with user-free accelerometers in the sensor networks. *2005 International Conference on Neural Networks and Brain*. 2005;2:1212-7.
6. Bao L, Intille SS. Activity recognition from user-annotated acceleration data. *International conference on pervasive computing*. 2004:1-17. doi: 10.1007/978-3-540-24646-6_1.
7. Ravi N, Dandekar N, Mysore P, Littman ML. Activity recognition from accelerometer data. *Aaai*. 2005;5(2005):1541-6.
8. Mantyjarvi J, Lindholm M, Vildjiounaite E, Makela S-M, Ailisto H. Identifying users of portable devices from gait pattern with accelerometers. *Proceedings(ICASSP'05) IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*. 2005;2:ii/973-ii/6 Vol. 2.
9. Ordonez FJ, Roggen D. Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors (Basel)*. 2016;16(1). doi: 10.3390/s16010115.
10. Kwapisz JR, Weiss GM, Moore SA. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*. 2011;12(2):74-82. doi: 10.1145/1964897.1964918.
11. Ignatov A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing*. 2018;62:915-22. doi: 10.1016/j.asoc.2017.09.027.
12. Kim TW, Lee SM, Seong SC, Lee S, Jang J, Lee MC. Different intraoperative kinematics with

comparable clinical outcomes of ultracongruent and posterior stabilized mobile-bearing total knee arthroplasty. *Knee Surg Sports Traumatol Arthrosc*. 2016;24(9):3036-43. doi: 10.1007/s00167-014-3489-0.
13. Lee S-M, Yoon SM, Cho H. Human activity recognition from accelerometer data using Convolutional Neural Network. *2017 ieee international conference on big data and smart computing (bigcomp)*. 2017:131-4.
14. Alsheikh MA, Selim A, Niyato D, Doyle L, Lin S, Tan H-P. Deep activity recognition models with triaxial accelerometers. *arXiv preprint arXiv:151104664*. 2015.
15. Singh D, Merdivan E, Psychoula I, Kropf J, Hanke S, Geist M, et al. Human activity recognition using recurrent neural networks. *International cross-domain conference for machine learning and knowledge extraction*. 2017:267-74. doi: 10.1007/978-3-319-66808-6_18.
16. Shojaedini SV, Beirami MJ. Mobile sensor based human activity recognition: distinguishing of challenging activities by applying long short-term memory deep learning modified by residual network concept. *Biomed Eng Lett*. 2020;10(3):419-30. doi: 10.1007/s13534-020-00160-x.
17. Zhang Y, Chan W, Jaitly N. Very deep convolutional networks for end-to-end speech recognition. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017:4845-9.
18. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. *European conference on computer vision*. 2014:818-33. doi: 10.1007/978-3-319-10590-1_53.
19. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010:249-56.
20. Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw*. 1994;5(2):157-66. doi: 10.1109/72.279181.
21. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016:770-8.
22. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *International conference on machine learning*. 2015:448-56.
23. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. *Journal of Machine Learning Research*. 2010;9:249-56.
24. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE international conference on computer vision*. 2015:1026-34.
25. LeCun YA, Bottou L, Orr GB, Müller K-R. Efficient backprop. Neural networks: Tricks of the trade: Springer; 2012. p. 9-48.
26. He K, Sun J. Convolutional neural networks at constrained time cost. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015:5353-60
27. Srivastava RK, Greff K, Schmidhuber J. Highway networks. *arXiv preprint arXiv:150500387*. 2015.
28. Zhang K, Sun M, Han TX, Yuan X, Guo L, Liu T. Residual networks of residual networks: Multilevel residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*. 2017;28(6):1303-14. doi: 10.1109/TCSVT.2017.2654543.
29. Weiss GM, Yoneda K, Hayajneh T. Smartphone and smartwatch-based biometrics using activities of daily living. *IEEE Access*. 2019;7:133190-202. doi: 10.1109/ACCESS.2019.2940729.
30. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16(5):412-24. doi: 10.1093/bioinformatics/16.5.412.
31. Plötz T, Hammerla NY, Olivier PL, editors. Feature learning for activity recognition in ubiquitous computing. Twenty-second international joint conference on artificial intelligence; 2011.
32. Zeng M, Nguyen LT, Yu B, Mengshoel OJ, Zhu J, Wu P, et al. Convolutional neural networks for human activity recognition using mobile sensors. *6th International Conference on Mobile Computing, Applications and Services*. 2014:197-205. doi: 10.4108/icst.mobicase.2014.257786.
33. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. (The Morgan Kaufmann Series in Data Management Systems). 3rd Edition. Amsterdam: Elsevier. 2011.