Human Activity Recognition Based on Residual Network

Jingxuan Zhang*, Shaojie Qiao, Zhiyu Lin and Yang Zhou

1 School of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China
2 School of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China
Email: jingxuan10@126.com

Abstract. With the development of information technology and the blooming of artificial intelligence, human activity recognition (HAR) has become a hot research topic. HAR has a certain prospect and practical value for mobile medicine and security monitoring, which has attracted the attention of many researchers. However, most of the existing HAR methods rely on the shallow feature learning mechanism, and the features learned by these methods are not completely consistent with the actual activities. To solve this problem, a human motion recognition method based on residual network (ResNet) is proposed. In addition, we modified the internal structure of ResNet by limiting the number of its network layers within a certain range to prevent over fitting and degradation. Our proposed method has better recognition performance than methods using pure convolutional neural networks (CNN), and it achieves 95.66% recognition accuracy.

1. Introduction
With the development of information technology and the blooming of intelligence, HAR has become a hot topic in the field of artificial intelligence research because HAR plays an important role in daily life. HAR is widely used in digital health and mobile medicine. For example, doctors can use HAR to monitor and analyze patients. This involves wearable sensors, such as accelerometer sensors. At the same time, HAR is widely used in many fields, such as gait recognition, bad behavior detection, military exercises, and identity authentication.

Although the study of HAR has been active for more than 10 years, there are still some problems. More specifically, traditional HAR methods can only learn shallow features, making the learning effect inconsistent with normal activities. While pure CNN models can automatically learn the features, however, the features learned and the classification performance is highly dependent on the depth and hyperparameters of the neural networks. For this reason, we propose the ResNet-based method in this paper. In comparison with vanilla CNN methods, the accuracy of our method is notably higher.

The rest of the paper is structured as follows. In Section II, we present current research on HAR. In Section III, our activity recognition method based on ResNet is clarified. Section IV describes the experimental work, including data sets, baseline method and evaluation metrics, network parameters, and experimental results. The conclusion is drawn in section V.

2. Related Work
Different sensors will reflect different values in different environments. In the study of HAR, there are two main research directions: HAR based on visual sensors and HAR based on wearable acceleration sensors.
2.1. HAR Based on Visual Sensors
In recent decades, many important achievements have been made in HAR research based on computer vision. Holte et al. [1] introduced the multi-viewpoint method for human 3d attitude estimation and activity recognition in recent years. This paper discusses the application fields of human attitude estimation and activity recognition, including human-computer interaction, interactive games, video surveillance and so on. Popoola et al. [2] introduced the recognition of human behavior and activity patterns with environmental anomalies. This paper focuses on the activity recognition of abnormal environments in video surveillance applications. The construction of the visual system of the scene and the semantic reasoning based on the observable characteristics of the moving object.

2.2. HAR Based on Wearable Acceleration Sensors
The acceleration sensor has been widely used for its low cost and small volume compared with the vision sensor. Ding, et al. [3] conducted the recognition process of 10 activities employing the random forest approach with an accuracy of 93.01%, but declining energy consumption by 74.9%. Karantonis et al. [4] proposed a real-time classification of types of human activity system. The system can identify the direction of the body's movement posture and judge whether the person is active or at rest. Overall accuracy was 90.8%.

Researchers not only use traditional methods to study HAR but also use deep learning methods. Pham et al. [5] proposed a RECOGNITION model to identify 7 daily human activities, and the accuracy of the system reached 93% on average. Murad et al. [6] used a deep recursive neural network (DRNNs) recognition model and proposed the recognition performance of Long Short-Term Memory (LSTM) in DRNNs for one-way, two-way and cascade architectures. The unidirectional DRNN model has a higher accuracy of 96.7%, while The cascaded DRNN has a slightly lower accuracy of 92.6%. Although the result is good, the disadvantage is that the time complexity is higher than the traditional algorithm. Chen et al. [7] used LSTM recursive neural network to analyze the activity recognition of sensors from accelerometers and gyroscopes, and proposed a position sensing method to improve the recognition accuracy. But the effect is not very ideal, the accuracy rate is 82.57%. Ronao et al. [8] proposed an adaptive extraction of robust features from sensor data based on RECOGNITION. The overall performance was 94.79%, and the fast Fourier transform recognition accuracy of additional information was 95.75%. Song et al. [9] also used the method of RECOGNITION to identify three kinds of daily activities, with an accuracy rate of 92.71%. However, converting the collected x, y, and z-axis acceleration data into vector amplitude data is an innovation.

Different from the methods mentioned above, we proposed activity recognition based on ResNet. To ensure fairness, we also used accelerometer data, gyroscope based 3d original acceleration sensor data.

3. Proposed Scheme
In this section, we mainly introduce the architecture of ResNet based model for HAR.

3.1. Residual Learning
ResNet applies the concept of residual representation to the construction of CNN model, that is, the block of residual learning. ResNet is characterized by multiple parametric layers to learn the residual representation between inputs and outputs, unlike ordinary CNN networks (such as AlexNet/VGG), which use parametric layers to directly try to learn the mappings between inputs and outputs. Assuming that the input of the neural network is x and the output (the underlying mapping) is H(x), in Eq(1) the learning objective of ResNet is:

\[ F(x) = H(x) - x \] (1)

F(x) is the residual function, which is flexible in form. In this experiment, a two-layer residual function F(x) is used. Meanwhile, we define a building block as:

\[ y = F (x, \{W_i \}) + x \] (2)
X is the input layer vector and y is the output layer vector. F(x, {Wi}) represents the residual mapping of learning. F + x is done by the shortcut connection and adding up the elements one by one. The shortcut connection [10] in Eq(2) does not increase the computational complexity, which is very convincing compared with ordinary recognition.

3.2. Network Architectures
To prevent overfitting and degradation, we improve the internal structure of ResNet by limiting the number of its network layers within a certain range. Since the data used in this study are sequences, the structure of the ResNet is thus been modified to match the input.

Fig. 1 shows the network structure of the modified ResNet. Before entering the first convolutional layer, the x, y, and z-axis acceleration sensor data are normalized. Then data are feed forward to the core part of our proposed model, the residual stages. Finally, there is the pooling layer, and fully connected layer with sigmoid activation function, which ends with the classification layer.

Fig. 2 shows the specific structure of the residual stage. We define the block, and each block consists of two convolution layers. Each square in Fig. 2 represents a layer and marks the number of convolution kernels currently used. There are dotted lines and solid lines. The solid line indicates that when the dimensions of the input vector and the output vector are equal, the two vectors are connected by the formula in Eq(1) as the input vector of the next block. The dotted line represents that when the size changes, the shortcut join still performs the identity mapping, filling in zeros to increase the size, which increases the dimension.

In our experiment, the original ResNet changed the number of layers and the size of the convolution kernel. We designed a total of 5 residual stages. The first residual stage (3 blocks in total, each block has 2 layers, 6 layers in total, the number of convolution cores is 16, that is, the output is 16-bit depth). The second residual stage (3 blocks, each with 2 floors, 6 floors in total, the number of convolution kernels is 32, that is, the output is 32-bit depth.). The third, fourth and fifth residual stage are the same as above. Note that the number of convolution cores defined by each residual stage is 16,32,64,128,256, and the corresponding blocks are 3 at a time.
4. Experiments
In this section, we will first describe the data sets used in the experiment. Second, we introduced the experimental Settings, including the baseline method, evaluation metrics, and network parameters. Finally, our model is compared with the recognition method.

4.1. Dataset
We use the public dataset WISDM [11], which is collected by accelerometer and gyroscope sensors. The dataset is collected on smartphones and smartwatches. The sampling frequency is 20Hz per second.

The dataset collected six kinds of daily human activities, including walking, jogging, upstairs, downstairs, sitting and standing. There are a total of 1,098,207 pieces of data in this data set, but the data distribution of each activity is unbalanced, which increases the difficulty of experiment al results.

In our experiment, we first cut the data, using 90 as the window. Divide the dataset into 24,403 parts. Meanwhile, 25% of the original data was randomly divided into test sets.

4.2. Baseline Method and Evaluation Metrics
We compared the ResNet method with the CNN method after improvement. CNN is a classic way to identify human activity. For evaluation indicators, we calculated the precision, recall, F1-score and so on of each activity. The obfuscation matrix is also made for the two methods, and the overall accuracy is obtained, which is convenient for better observation of experimental results.

4.3. Network Parameters and Training Methodology
In our experiment, we defined the shape of the convolution kernel as 3x1 and the stride as 2. We used 0.001 weight decay, 0.9 momentum, and weight initialization, but no dropout. At the same time, the original ResNet changed the number of layers and the size of the convolution kernel. The number of convolution cores defined by each residual stage was 16,32,64,128,256, and the corresponding blocks were 3 at a time. Set 2 layers in the Block.
4.4. Experimental Results

For evaluation, our model was compared with CNN-based activity recognition method. Table I can be seen that the confusion matrix of activity recognition as well as the precision and recall of each activity. The effect of the activity recognition model based on CNN has been better than other traditional methods such as SVM, achieving more than 90% accuracy. However, our model is better than CNN overall, with an overall accuracy of 95.66%. It is worth mentioning that the activity precision of sit has reached 99.01%, which is an unexpected result. It can be reflected that the model has learned good features and shown good results. The activity precision of Jogging is 96.82%, which may be due to the large acceleration of Jogging activity itself, it is relatively easy to identify.

Table 1. Compare the results with CNN.

|         | CNN | Downstairs | Jog | Sit | Stand | Upstairs | Walk | Recall(%) |
|---------|-----|------------|-----|-----|-------|----------|------|-----------|
| Downstairs | 472 | 18 | 0 | 0 | 55 | 48 | 79.60 |
| Jog | 5 | 1847 | 3 | 0 | 2 | 25 | 98.14 |
| Sit | 2 | 0 | 290 | 41 | 4 | 4 | 85.04 |
| Stand | 1 | 0 | 3 | 233 | 5 | 7 | 93.57 |
| Upstairs | 72 | 21 | 4 | 2 | 526 | 67 | 76.01 |
| Walk | 27 | 35 | 2 | 0 | 23 | 2257 | 96.29 |
| Precision(%) | 81.52 | 96.15 | 96.02 | 84.42 | 85.53 | 93.73 | 92.19 |

|         | ResNet | Downstairs | Jog | Sit | Stand | Upstairs | Walk | Recall(%) |
|---------|--------|------------|-----|-----|-------|----------|------|-----------|
| Downstairs | 499 | 13 | 0 | 0 | 30 | 12 | 90.07 |
| Jog | 3 | 1860 | 1 | 0 | 2 | 12 | 99.04 |
| Sit | 0 | 0 | 299 | 33 | 2 | 2 | 88.99 |
| Stand | 0 | 0 | 0 | 243 | 1 | 0 | 99.59 |
| Upstairs | 65 | 19 | 1 | 0 | 563 | 10 | 85.56 |
| Walk | 12 | 29 | 1 | 0 | 17 | 2372 | 97.57 |
| Precision(%) | 86.18 | 96.82 | 99.01 | 88.04 | 91.54 | 98.50 | 95.66 |

At the same time, it is not difficult to see in table 1 that the Downstairs has slightly lower precision than the other activities. The possible factor is that the activity data on the Downstairs often contains ambiguous signals, which make the accuracy low. At the same time, the result of standing is second only to downstairs, which may be caused by the ambiguous judgment of downstairs, upstairs and standing in the learning process of the model, which reduces the recognition performance. For this reason, follow-up work will pay more attention to this aspect.

5. Conclusion

We proposed a ResNet based human motion recognition method, using triaxial accelerometer data collected from smartphones and smart bracelets. Our method is superior to CNN in human activity classification, and we have partially modified the network structure of the ResNet.

To sum up, we can see that the performance of activity recognition will be improved if the depth is increased under the condition that no fitting is guaranteed. Therefore, we might consider that recognition performance might be improved if features learned in activity recognition are deeper.

6. References

[1] Holte, M. B., Tran, C., Trivedi, M. M., & Moeslund, T. B. (2012). Human Pose Estimation and Activity Recognition From Multi-View Videos: Comparative Explorations of Recent
Developments. IEEE Journal of Selected Topics in Signal Processing, 6(5), 538–552. doi:10.1109/jstsp.2012.2196975.

[2] Popoola, O. P., & Kejun Wang. (2012). Video-Based Abnormal Human Behavior Recognition—A Review. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), 865–878. doi:10.1109/tsmcc.2011.2178594.

[3] Ding, G., Tian, J., Wu, J., Zhao, Q., & Xie, L. (2018). Energy efficient human activity recognition using wearable sensors. 2018 IEEE Wireless Communications and Networking Conference Workshops (WCNCW). doi:10.1109/wcncw.2018.8368980.

[4] Karantonis, D. M., Narayanan, M. R., Mathie, M., Lovell, N. H., & Celler, B. G. (2006). Implementation of a Real-Time Human Movement Classifier Using a Triaxial Accelerometer for Ambulatory Monitoring. IEEE Transactions on Information Technology in Biomedicine, 10(1), 156–167. doi:10.1109/titb.2005.856864.

[5] Pham, C., Diep, N. N., & Phuong, T. M. (2017). e-Shoes: Smart shoes for unobtrusive human activity recognition. 2017 9th International Conference on Knowledge and Systems Engineering (KSE). doi:10.1109/kse.2017.8119470.

[6] Murad, A., & Pyun, J.-Y. (2017). Deep Recurrent Neural Networks for Human Activity Recognition. Sensors, 17(11), 2556. doi:10.3390/s17112556.

[7] Chen, W.-H., Betancourt Baca, C. A., & Tou, C.-H. (2017). LSTM-RNNs combined with scene information for human activity recognition. 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom). doi:10.1109/healthcom.2017.8210846.

[8] Ronao, C. A., & Cho, S.-B. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. Expert Systems with Applications, 59, 235–244. doi:10.1016/j.eswa.2016.04.032.

[9] Song-Mi Lee, Sang Min Yoon, & Heeryon Cho. (2017). Human activity recognition from accelerometer data using Convolutional Neural Network. 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). doi:10.1109/bigcomp.2017.7881728.

[10] Raiko, Tapani, Valpola, Harri, & LeCun, Yann. (2012). Deep learning made easier by linear transformations in perceptrons. doi:http://dx.doi.org.

[11] Kwapisz, J. R., Weiss, G. M., & Moore, S. A. (2011). Activity recognition using cell phone accelerometers. ACM SIGKDD Explorations Newsletter, 12(2), 74. doi:10.1145/1964897.1964918.