Human Action Recognition on Cellphone Using Compositional Bidir-LSTM-CNN Networks

Jiahao Wang[1, a], Qiuling Long[1, b], PiRah[c], Kexuan Liu and Yingzi Xie
[1]University of Electronic Science and Technology of China, Chengdu 610000, China.
[a]wangjh@uestc.edu.cn, [b]qiuling_long@163.com, [c]pirahmemon3@gmail.com

Abstract. Recently, the multimoal and high dimensional sensor data are prone to problems such as artificial error and time- consuming acquisition processes, especially in supervised human activity recognition. Therefore, this study proposes an activity recognition framework called compositional Bidir-LSTM-CNN Networks, which automatically extracts features from raw data using the optimized Convolutional Neural Network and further capture dynamic temporal features through the Bidirectional Lone Short Term Memory Network. Finally, this study paves the way for accurate recognition of human activities using the proposed framework with significantly improve 8% recognition accuracy along with additional features such as robustness and generalization.

Keywords: Machine learning, Compositional Bidir-LSTM-CNN, Accelerometer Sensors, Human Activity recognition.

1. Introduction

With a widespread of various sensors embedded in mobile devices,human activity recognition ba sed wearable sensor data has been an attractive research topic.Activity recognition is becoming an in tegral part of many mobile applications.

Early studies use machine learning algorithm to do human activity recognition. Chen Y[1] propos ed a cycle detection algorithm and then characterized by time-, frequency-, and wavelet-domain feat ures to classify different action. Wang Z et al[2] proposed a fast and robust human activity recogniti on model which called TransM-RKELM to classify the activity. Münzner, Sebastian et al[3] used co nvolutional neural networks(CNNs)to activity recognition. Similarly,Ignatov A[4] proposed using C NN for local feature extraction together with statistical features that preserved information about the global form of time series.In sharp contrast to the hand-crafted feature extraction, deep learning can feature extraction automatically. Nevertheless, how to build an efficient deep learning model is a fu ndamental yet non-trivial problem.

As a typical time series data,in order to extract the characteristic information,ensure data informat ion flow has a high-speed information channel.In this paper,the combination of CNN and bidirection al LSTM training is proposed to train the model, which can enrich the discriminative feature, so as t o improve the accuracy of human activity recognition.

2. Related Work

At present, machine learning and deep learning can be used to classify human activities. It is pointed out in the article in HaritHa Vellampalli et.al.[5] that Decision Tree, K Nearest Neighbor, Naive Bayes, Multinomial Logistic Regression and Artificial Neural Network algorithms have been used to perform the classification task. Ha S[6] presented CNNs(CNN-pf and CNN-pff)for multi-sensor data which learned from multi-sensor data and were eventually aggregated in upper layers. Chen et.al[7] analyzed sensor readings from accelerometers and gyroscopes using long short-term memory to improve the recognition accuracy. Zhao Y[8] proposed a deep network architecture using residual bidirectional long short-term memory(LSTM)for human activity. Such neural networks as LSTM-RNNs[9], ConvLSTM[10], RNN[11], CNN-LSTM[12], CNN [13] [14], AutoEncoder[15] are also applicable to human activity recognition.

Copyright © 2019, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

3. Human Activity Recognition

3.1 Action Recognition Model.

Convolutional Neural Network has its special structure as a local shared and has unique advantage in speech recognition and image processing which effectively extracts the spatial characteristic data. But for time series of sensitive data, CNN and neural networks are not effective in its classification. LSTM network has a good classification effect for time series data which is sensitive. Therefore, this paper is proposed that combines the Long Short Term Memory Network (LSTM) and the optimized Convolutional Neural Network (CNN). The optimized CNN is used to extract the spatial features and the bidirectional LSTM network is used to extract the spatial features and the bidirectional LSTM network is used to extract the time-series features, combine two ways features and train the model to improve its classification accurately. The neural network framework proposed in this paper is shown in Fig. 1.

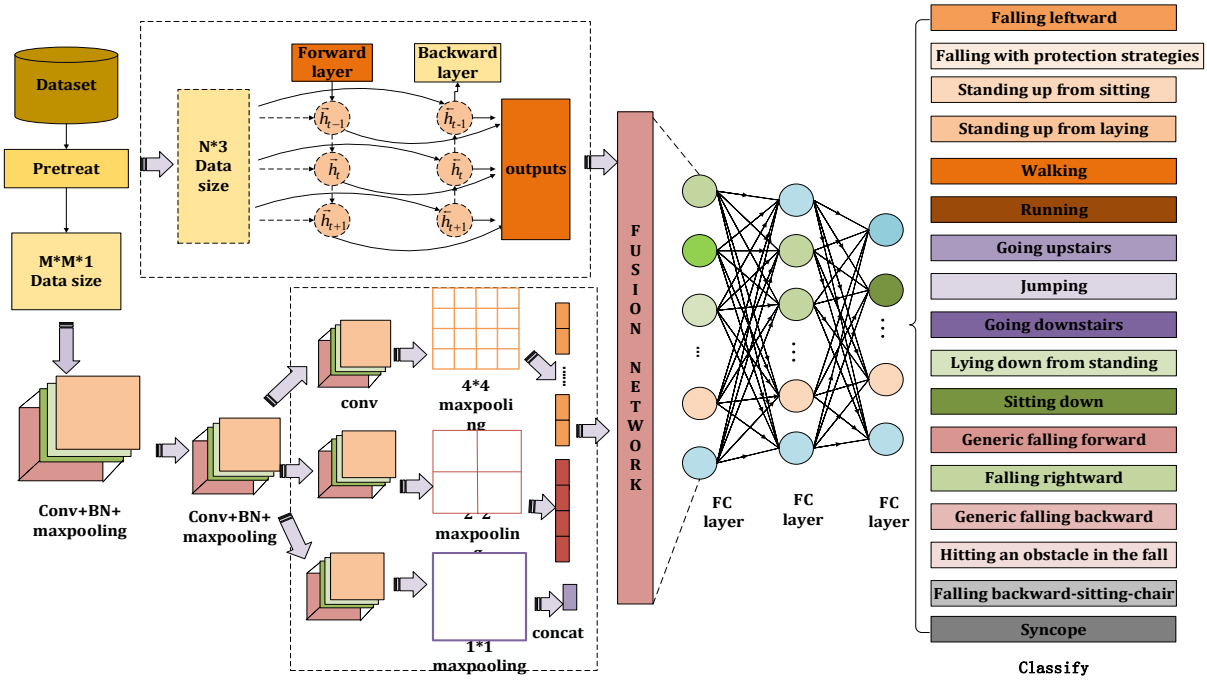

Fig. 1 Bidir-LSTM-CNN action recognition

The raw data collected on the cellphone is littered with a lot of useless information (incomplete data, duplicate data, etc.).Mean-var normalization is adopted to normalize the data to the same range and accelerate the convergence speed of model classification. Meanwhile, mobile average filter is adopted to reduce the noise of the data.

The pre-processed data will be fed into dual-channel neural network, on the one hand for time series data to distinguish the characteristic information extraction, using bi-dirdectional LSTM neural network training data. On the other hand is mainly used for extracting data space characteristics.

3.2 Bidirectional LSTM Network.

LSTM improve the learning ability of the model, control the discarding or adding information through the"gate", to realize the function of forgetting or remembering, which is suitable for complex large-scale human activity recognition. The Cell layer, as a transport belt of information, holds the important information in the data. The forget gate can decide whether to retain the information of the previous sequence or not. For the most part, the network uses sigmoid as the activation function.

The bidirectional LSTM, as a time recursive neural network, ensures the integrity of the neural network's deep flow information through bidirectional communication. Under the condition that the model parameters are fixed, the scale of integration at different moments be changed dynamically, which avoids the problems of vanishing or gradient exploding. The output gate is calculated as follows

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (1)$$

where, o_t is the output threshold, h_{t-1} is the output of the unit at the previous moment, x_t is the current input data, σ is the activation function and usually uses sigmoid, W_o and b_o are the weight and bias of the output gate, respectively.

Data collectedData collected based on the sensor is N*3 data samples(include three axes from x,y,z; each axis contains N samples).We distinguish the triaxial data to avoid the negative influence of triaxial data confusion on feature extraction, and set the dimension of the data to[N,3]before sending it into the network. Meanwhile, the dropout layer is added to the bidirectional LSTM to avoid the phenomenon of over-fitting. In the paper, we set the number of neurons in each layer to 20.

3.3 Optimized Convolutional Neural Network.

The optimized CNN adopts a single channel CNN, adding several zeros to the data of each axis, which not only avoids the information loss caused by data clipping, but also avoids the confusion of triaxial data. At the same time, standard normalization layer is added after each layer. It can not only accelerate the optimization speed of the network, but also solve the phenomenon that the network layer number is too much forward transmission effectively, prevent the occurrence of over-fitting and improve the generalization ability of the network. The normalization layer is calculated as follows

$$y_i = \gamma \hat{x}_i + \beta = BN_{\gamma,\beta}(x_i) \quad (2)$$

where $\hat{x}_i$ is the standardized sample value, γ and β is the learnable reconstruction parameter of BN standardized layer.

As the number of convolutional layers increases, the features extracted from the model will become more and more abstract, which will probably lead to the decrease of accuracy. For the last convolution layer, as shown in Fig 2, we will use the convolution kernel of three dimensions to extract the feature information simultaneously, extract the features of the previous layer from different perspectives and reduce the dimension through the pooling layer, construct the description of local feature information.

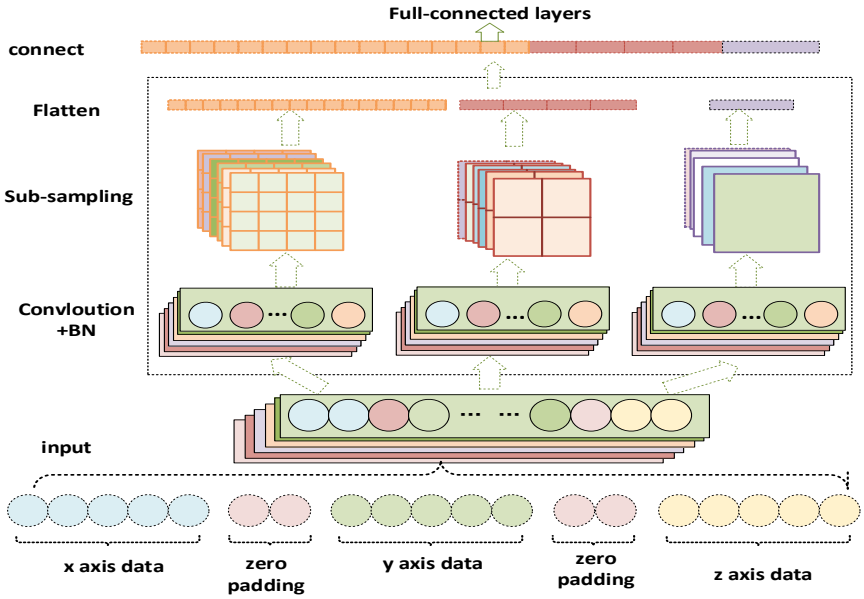

Fig. 2 Multi-Dimensional convolution diagram

3.4 The Fusion of Network.

The training of the network needs good information flow. If a layer in the middle of the network is closed to the input layer, it will have a high-speed channel for information flow. The linear training mode will make the information flow not have high-speed channel characteristics, and will lead to the risk of vanishing gradient and loss of information. This study abandoned the linear training process of convolution-pooling-LSTM-full connection-classification, using parallel training and re-fusion to

avoid interference caused by temporal features when collecting spatial features. The bidirectional LSTM and the optimized CNN are combined to form a hybrid optimization design method, which makes the bidirectional LSTM and CNN complement their shortcomings.

Dimension obtained by bidirectional LSTM training in the network is not consistent with that obtained by optimized CNN training, so dimension splicing can not be carried out directly. We feed it into the flatten layer respectively, and then combine the multidimensional data into one dimension. The feature information obtained from the training in LSTM is aggregated with the optimized CNN feature information, and it is fed into the multi-layer full connection layer, sent it into the multi-layer full connection layer. Through this method can effectively improve the accuracy of classification and recognition.

4. Experiments and Results

4.1 Dataset Description.

The UniMiB SHAR dataset is an activity recognition based on smartphone accelerometer data, the smartphone used in the experiments was a Samsung Galaxy Nexus I9250 with the Android OS version 5.1.1 and equipped with a Bosh BMA220 acceleration sensor. Fig 3 shows that the dataset containing 17 fine grained classes of human activities grouped in two coarse grained activities classes :(ADLs and falls). The dataset contains a total of 11,771 (7,759 ADLs and 4,192 falls) activities performed by 30 subjects.

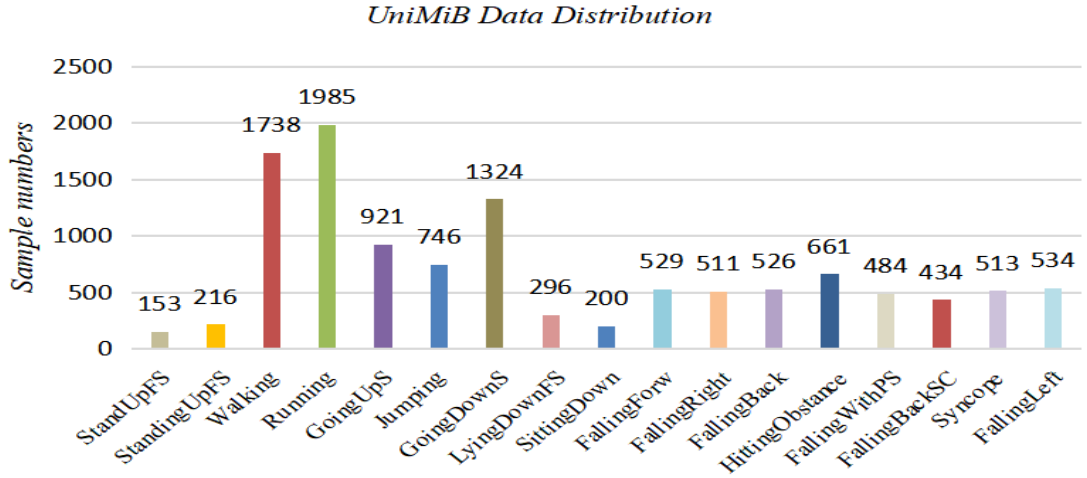

Fig. 3 Activity samples distribution

4.2 Experimental Results.

Based on results in Fig 4 and Table 1, we can observe that the human activity classification based on Bidir-LSTM-CNN 92.5% accuracy which is superior to machine learning and CNN and LSTM networks. In the machine learning algorithm, the accuracy of the random forest can reached 85%,which further expresses the influence of artificial feature extraction on the classification and the self-learning feature based on deep learning can well classify the activities. The classification accuracy of CNN-based, LSTM-based is 89.97%, 86.32% respectively and bidir-LSTM-CNN's classification accuracy is 92.53%.we can also observe that the greater the human activities similarity, the more difficult it is to distinguish. The accuracy of activities such as walking and jumping is 97%and 99%, respectively, while the accuracy of similar activities (falling left and syncope) is slightly reduced, and falling left is often recognized as generic falling forward. The proposed fusion network and its associated activity recognition algorithm can be regarded as an effective method for human daily activity recognition tasks.

Table 1 Accuracies of activity recognition

Classification results for UniMiB dataset									
	StandingUpFS	StandingUpFL	Walking	Running	GoingUpS	Jumping	GoingDownS	LyingDownFS	SittingDown
bidri-LSTM-CNN	0.97	0.81	0.97	1	0.97	0.99	0.95	0.92	0.86
SVM	0.71	0.59	0.75	0.71	0.66	0.44	0.46	0.45	0.44
J48	0.53	0.45	0.8	0.78	0.61	0.67	0.65	0.48	0.54
Random Forest	0.86	0.85	0.89	0.86	0.81	0.87	0.84	0.76	0.84
CNN	0.84	0.81	0.95	0.98	0.93	0.95	0.81	0.66	0.84
LSTM	0.8	0.75	0.91	0.92	0.87	0.86	0.75	0.76	0.83
	FallingForw	FallingRight	FallingBack	HittingObstacle	FallingWithPS	FallingBackSC	Syncope	FallingLeft	Average Acc
bidri-LSTM-CNN	0.92	0.8	0.82	0.95	0.85	0.79	0.83	0.88	92.53%
SVM	0.44	0.69	0.52	0.47	0.41	0.56	0.48	0.4	62.52%
J48	0.71	0.66	0.64	0.83	0.68	0.65	0.63	0.67	69.65%
Random Forest	0.83	0.8	0.84	0.86	0.87	0.84	0.87	0.85	85.36%
CNN	0.89	0.85	0.84	0.95	0.76	0.76	0.68	0.84	89.97%
LSTM	0.78	0.82	0.76	0.88	0.83	0.8	0.82	0.79	86.32%

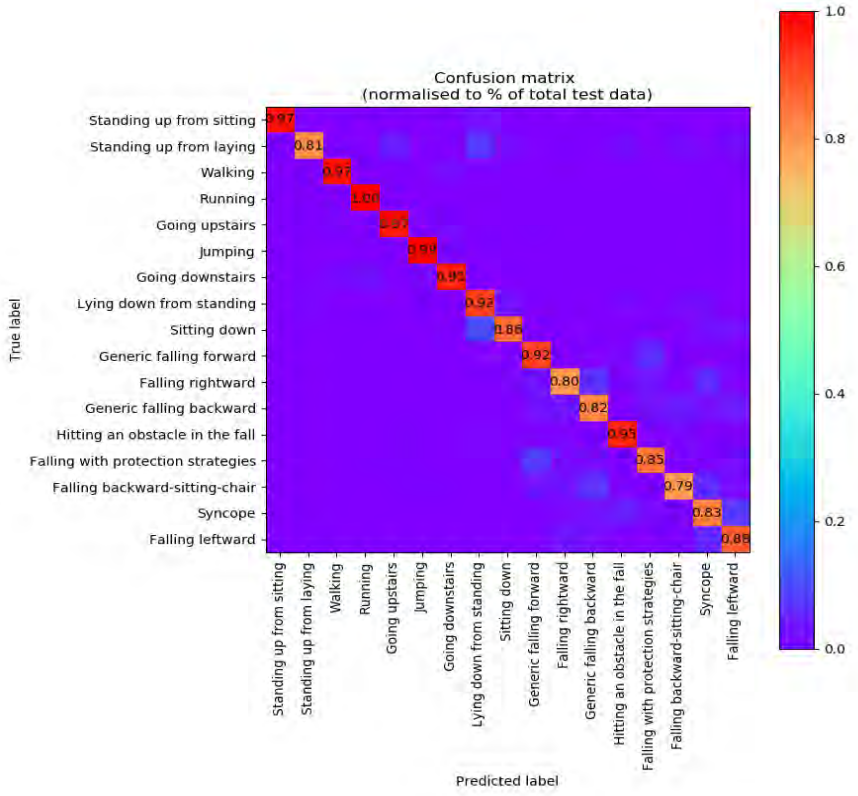

Fig. 4 Confusion matrix for compositional Bidir-LSTM-CNN

5. Conclusion

In this paper, we proposed a new fusion method to automate feature extraction for human activity recognition task. The proposed method mainly employed the convolution and pooling operations to capture the salient feature information of sensor signals at different time scales. Instead of manually extracting features to distinguish activity categories, the method adopted automatic feature extraction to improve efficiency and accuracy 92.5%.Experimental results showed that the approach offers a feasible solution to human activity recognition.

References

[1]. Chen Y, Shen C. Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition[J]. IEEE ACCESS, 2017, 5(99):3095-3110.

[2]. Wang Z, Wu D, Gravina R, et al. Kernel fusion based extreme learning machine for cross-location activity recognition[J]. Information Fusion, 2017, 37(C):1-9.

[3]. Münzner, Sebastian, Schmidt P , Reiss A , et al. [ACM Press the 2017 ACM International Symposium-Maui, Hawaii (2017.09.11-2017.09.15)] Proceedings of the 2017 ACM International Symposium on Wearable Computers, - ISWC \"17 - CNN-based sensor fusion techniques for multimodal human activity recognition[J]. 2017:158-165.

[4]. Ignatov A. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks[J]. Applied Soft Computing, 2018, 62: 915-922.

[5]. Vellampalli,H.(2017)Physical Human Activity Recognition Using Machine Learning Algorithm s Masters thesis, DIT,2017.

[6]. Ha S, Choi S. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors[C]//Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE, 2016: 381-388.

[7]. Chen W H, Baca C A B, Tou C H. LSTM-RNNs combined with scene information for human activity recognition[C]// IEEE, International Conference on E-Health Networking, Applications and Services. IEEE, 2017.

[8]. Zhao Y, Yang R, Chevalier G, et al. Deep Residual Bidir-LSTM for Human Activity Recognition Using Wearable Sensors[J]. 2017.

[9]. Chen W H, Carlos Andrés Betancourt Baca, Tou C H. LSTM-RNNs combined with scene information for human activity recognition[C]// IEEE International Conference on E-health Networking. IEEE, 2017.

[10]. Shi X, Chen Z, Wang H, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[J]. 2015.

[11]. Fang H, Si H, Chen L. Recurrent Neural Network for Human Activity Recognition in Smart Home[M]// Proceedings of 2013 Chinese Intelligent Automation Conference. Springer Berlin Heidelberg, 2013:341-348.

[12]. OKITA T, INOUE S. Recognition of multiple overlapping activities using compositional CNN-LSTM model[C]//Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers. ACM, 2017: 165-168.

[13]. Münzner S, Schmidt P, Reiss A, et al. CNN-based sensor fusion techniques for multimodal human activity recognition[C]// Acm International Symposium. 2017.

[14]. Jiang X, Lu Y, Lu Z, et al. Smartphone-Based Human Activity Recognition Using CNN in Frequency Domain[J]. 2018.

[15]. Badem H , Caliskan A , Basturk A , et al. Classification of human activity by using a Stacked Autoencoder[C]// Medical Technologies Nationa.