IndRNN Based Long-term Temporal Recognition in the Spatial and Frequency Domain

Beidi Zhao*
University of Electronic Science and Technology of China
Chengdu, China
beidizhao@hotmail.com

Shuai Li*
Shandong University
Jinan, China
shuaili@sdu.edu.cn

Yanbo Gao
Shandong University
Jinan, China
ybgao@sdu.edu.cn

ABSTRACT

This paper targets the SHL recognition challenge, which focuses on the location-independent and user-independent activity recognition using smartphone sensors. To address this long-range temporal problem with periodic nature, we propose a new approach (team IndRNN), an Independently Recurrent Neural Network (IndRNN) based long-term temporal activity recognition with spatial and frequency domain features. The data is first segmented into one second sliding windows, then temporal and frequency domain features are extracted as short-term temporal features. A deep IndRNN model is used to predict the unknown test dataset location. Under the predicted location, a deep IndRNN model is further used to classify the 8 activities with best performed features. Finally, transfer learning and model fusion are used to improve the result under the user-independence case. The proposed method achieves 86.94% accuracy on the validation set at the predicted location.

CCS CONCEPTS

• Computing methodologies → Artificial intelligence.

KEYWORDS

Activity recognition, IndRNN, SHL dataset, Smartphone

ACM Reference Format:
Beidi Zhao, Shuai Li and Yanbo Gao. 2020. IndRNN Based Long-term Temporal Recognition in the Spatial and Frequency Domain. Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct), September 12–16, 2020, Virtual Event, Mexico, 5 pages. https://doi.org/10.1145/3410530.3414355

* indicates equal contributions.

UbiComp/ISWC '20 Adjunct, September 12–16, 2020, Virtual Event, Mexico

ACM ISBN 978-1-4503-8076-8/20/09…$15.00
https://doi.org/10.1145/3410530.3414355

1 INTRODUCTION

Nowadays, smartphones are widely used to collect users' data with build-in sensors for recognizing human activities because of the extensive use such as unimpressive size and portability. One area is to recognize transposition modes, which can provide users with adaptive contextual services such as more accurate energy consumption, route or park suggestions [11]. In this paper, we focus on using sensor data of a smartphone independent of mobile-phone placement to recognize 8 modes of activities including man-powered and motor-powered transportation.

The SHL recognition challenge provides a dataset [2], which allows to compare methodologies and to systematically advance research in the field. This year's edition uses previously unreleased data. Users' data has been collected from the phone at hand, torso, hips and bag. The training set contains four position data of user 1, whereas the validation data consists of those from user 2 and user 3. The goal is to recognize the test data of user 2 and user 3 from one unknown position. More precisely, the challenge lays emphasis on recognizing modes of transportation in a user-independent manner with an unknown phone position.

Some methods have been proposed for the locomotion and transportation recognition in the literature based on machine learning [11, 14] and deep learning [15, 16]. Convolutional Neural Network (CNN) [12], Extreme Gradient Boosting model (XGB) [11], EmbraceNet [13] have been used as recognition models in the research. In addition to the recognition models, feature extraction as a preprocessing technique has also been widely adopted. Some traditional temporal domain features have been usually used [10]. Due to the periodicity of the smartphone sensor data, FFT is also regarded as an effective feature extraction method. Apart from that, some researches also directly learn useful features [10] from the raw data automatically. After recognition, some methods also employ post-processing to further process the results from different segments. Hidden Markov model [11, 14] is used to smooth the predictions of continuous sequential signals considering that the probability of multiple action change is usually low within a short time window. Cross-location transfer learning [11] have also been studied for location-independence recognition.

Activity recognition based on smartphones can be considered as a long-term temporal recognition task in view that the high sampling rate of the smartphone sensors (100Hz) results in a large number of data in the temporal domain. To address this task, in this paper, the deep Independently Recurrent Neural Network (IndRNN) [3][4] is used, which can effectively learn long-term temporal features. First, high location-dependent sensor data is

converted to navigation coordinate system. After that, the data recorded in 5 seconds per instance is segmented to 20 overlapping one-second windows. Then, frequency domain features and short-term temporal features in the spatial domain are extracted and combined to pre-process the data of short time windows. Before performing the activity recognition, a location recognition model is first developed with IndRNN to predict the unknown location of testing data. With the predicted location, features that well-behaved on the predicted location are used as the input features, and a deep densely connected IndRNN is constructed for recognition. Finally, two transfer learning models are fused to improve the performance and predict the activity of the testing dataset.

2 SHL DATASET

The SHL dataset contains three parts: train, validation and test dataset. Training set consists of 59 days of user 1, validation set consists of 6 days of user 2 and 3 and testing set contains 40 days of user 2 and 3. The raw sensors data was recorded by Huawei Mate 9 smartphone sensors (magnetometer, acceleration, linear acceleration, gravity, orientation (quaternions), gyroscope and pressure) at four different locations of users' body (bag, hips, torso and hand). The data from one location (unknown to the participants) is used for test, while the data from all the four locations are available for the train and validation sets. Eight kinds of activities (standing still, walking, running, biking, car, bus, train and subway) occurred in the datasets as the to-be-recognized labels. All of the three sets were generated by segmenting the whole data with a non-overlap sliding window of 5 seconds. The data from one phone location (unknown to the participants) is used for test.

3 METHOD

3.1 Pre-Processing and Feature Extraction

Before pre-processing the data, data with the NAN value in the training dataset are first filtered out. Since all of the raw sensors data is collected in the smartphone coordinate and that of acceleration and magnetometer is highly related to the orientation of the phone, we firstly de-rotate the acceleration and magnetometer data to the NED (North-East-Down) coordinate system to eliminate the influence of different orientations of the phone. In the NED coordinate system, local coordinate data from different phone orientations is transformed to the same coordinate, so the influence from the smartphone orientations is eliminated. The transform can be performed by multiplying the raw sensors data with the rotation matrix R_N^B derived from quarternions $[q_w, q_x, q_y, q_z]$ as shown below.

$$R_{NB} = \begin{bmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y-q_wq_z) & 2(q_xq_z+q_wq_y) \\ 2(q_xq_y+q_wq_z) & 1-2(q_x^2+q_z^2) & 2(q_yq_z-q_wq_x) \\ 2(q_xq_z-q_wq_y) & 2(q_yq_z+q_wq_x) & 1-2(q_x^2+q_y^2) \end{bmatrix}$$

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_N = R_{NB} \begin{bmatrix} x \\ y \\ z \end{bmatrix}_B$$

To better learn the short-term features and long-term features, each 500-frame (5 seconds) sample was segmented into 100-frame (1 second) overlapping sliding windows in the same as our last-year competition [5]. Features can be extracted from these windows as short-term temporal features.

In this paper, short-term temporal features in both the spatial and frequency domains are extracted. For the short-term temporal feature in the spatial domain, the maximum, minimum, mean, standard variance, numbers above mean and numbers below mean of segmented windows of accelerometer, gyroscope, magnetometer and per sample normalized pressure are used. On the other hand, since the smartphone sensors data is strongly periodic, FFT spectrum can be used as powerful way to extract the frequency domain features. The power spectrums of the Accelerate (NED coordinate), Gyroscope, and Magnetometer (NED coordinate) and per sample normalized pressure in the segments are used. Moreover, considering that the distributions of the power spectrum may also be different for different activities, the mean and standard variance of the power spectrums are also calculated as features. The feature extraction process is illustrated in Fig. 1.

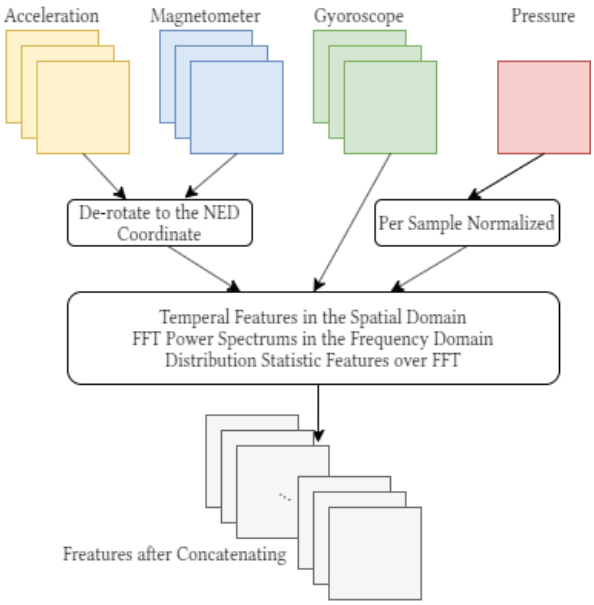

Figure 1: Diagram of the feature extraction process

3.2 Independently Recurrent Neural Network

The IndRNN can easily construct deeper networks compared to the traditional RNN and LSTM, as shown in [3, 4]. It solves the gradient exploding and vanishing problem over time and can keep long-term memory. In this paper, a deep dense IndRNN as in [4] is constructed as the classification model to capture the long-term patterns on top of the short-term temporal features. The form of IndRNN can be represented as:

$$h_t = \sigma(Wx_t + u \odot h_{t-1} + b)$$

Where $x_t \in R^M$ and $h_t \in R^N$ is the input and hidden state at time step t, respectively. $W \in R^{M \times N}, u \in R^N$ and $b \in R^N$ are the weights for the current input and the recurrent input and the bias

of neurons. $\odot$ represent the Hadamard product and σ is the non-linear activation function of neurons. N is the number of neuron of this IndRNN layer. Different from the residual IndRNN we used in the last-years competition [3][5], dense IndRNN concatenates all the features of the previous layers to replace the elementwise added skip-connection [4]. It follows $x_{l,t} = \mathcal{C}\left(x_{l-1,t}, \mathcal{F}_l\left(x_{l-1,t}\right)\right)$, where $\mathcal{C}$ is the concatenation operation, $x_{l,t}$ is a combination of all the features in the previous layers and the non-linear transformation of the features in the previous layer.

3.3 Location Recognition and Activity Recognition

The dataset contains data from four locations (bag, hips, torso and hand). It is found that different features may behave differently for activity recognition at different locations. Therefore, identifying the location of the unknown testing set may assist the final test by extracting the appropriate features according to the location.

First, the data are regrouped to different location labels (1 for bag, 2 for hips, 3 for torso and 4 for hand) regardless of the activity label. An IndRNN model is used to classify the locations. The results are presented in the form of the validation confusion matrix as shown in Fig. 3. It can be seen that while a relatively high recognition performance can be achieved for each location, the confusion among locations are mainly from the hand-bag, and hips-torso. If we further category them into two labels (hand-bag, and hips-torso), the recognition performance can be above 99% in terms of this two-class recognition accuracy. Therefore, the locations are further separated into two location groups (hand-bag, and hips-torso). Performing on the testing set, the results show that the data are from the hips-torso location. Note that in our experiments, we further found that the performance of different features on the hips and torso locations are very similar, thus we did not further separate these two locations.

For the following activity recognition, the features are specified as the FFT power spectrums of the de-rotated Accelerometer, Magnetometer, raw Gyroscope data and its corresponding mean and standard variance, which has performed the best on the hips-torso. The framework of the proposed method for the activity recognition is illustrated in Fig. 4.

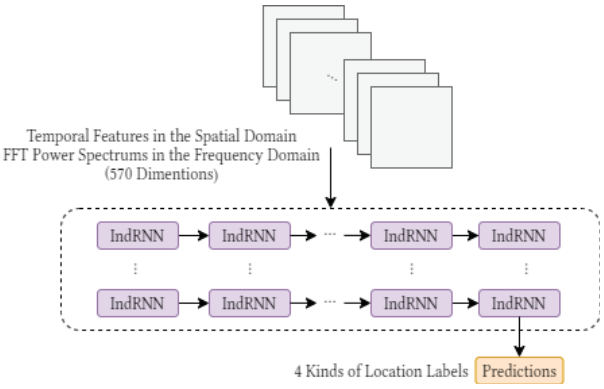

Figure 2: Framework of the proposed IndRNN based location recognition

	Bag	Hips	Torso	Hand
Bag	0.77	0	0	0.23
Hips	0	0.63	0.37	0
Torso	0	0.19	0.81	0
Hand	0.13	0	0	0.87

Figure 3: The confusion matrix of the location recognition on the validation dataset

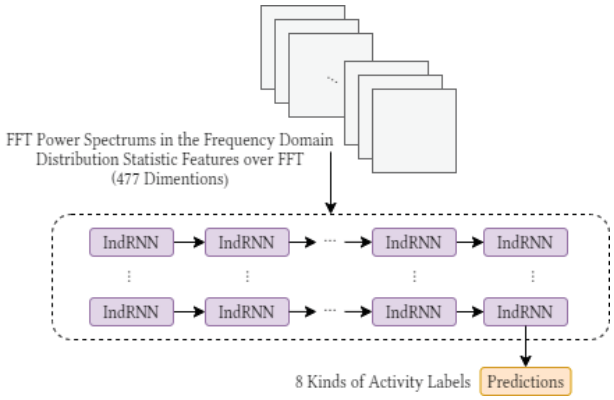

Figure 4: Framework of the proposed IndRNN based activity recognition

3.4 Transfer Learning

The training data is from user 1, while validation and test data are from user 2 and user 3. Therefore, transfer learning is adopted to accommodate such user difference in order to further improve the performance.

Since the validation set is the concatenation of the data from user 2 and user 3, it is firstly roughly split into two users. By observing the labels of validation set, the distribution of activity labels is not balanced over time, and some actions only occur in certain days. To separate roughly similar numbers of actions into different sets, data with the same labels are concatenated first. Then the first two-thirds from user 2 and user 3 with each label are used for training and the other one-third is used for validation in the transfer process, which is denoted by transfer A. Similarly, setting the last two-thirds for training and the other one-third for validation, is denoted by transfer B. Finally, the model trained from transfer A and transfer B are fused together to predict the test data set, where the final output is obtained as the label with the max probability in the summarized output of transfer A and transfer B models.

4 EXPERIMENTS

The preprocessed training data of all locations (bag, hips, torso, hand) is used for training with the dense IndRNN model. Cross-entropy is used as the loss and Adam is used to optimize during the training process. The learning rate of our model is set to 5×10^{-5}. To restrain the slightly larger fluctuation at the beginning of the training process, it is set to 5×10^{-6} at the first 10 epochs. The learning rate drops 10 times once the validation

accuracy does not increase (over a patience 100). Mini-batch with 128 batch size is used to train our model. The dense block configuration is set to (8, 6, 4), where in the first, second and third dense block, 8, 6 and 4 dense layers are used, respectively. This keeps a relatively similar number of neurons in each dense block. The growth rate is set to 48. In order to reduce overfitting, dropout is applied after the input (0.5), each dense layer (0.5), each bottleneck layer (0.1) and each transition layer expect (0.1 for the last and 0.3 for the others). The results on the validation dataset are shown in Figure 5 with the proposed dense IndRNN model compared with the previous residual IndRNN model [5]. It can be seen that the proposed dense IndRNN model further improves the performance by around 4%.

	still	walk	run	bike	car	bus	train	subway
still	0.88	0.01	0	0.01	0	0.01	0.06	0.03
walk	0.07	0.88	0	0.03	0	0.01	0.01	0.01
run	0	0.05	0.54	0.40	0	0.01	0.02	0.10
bike	0.02	0.05	0	0.72	0.02	0.07	0.02	0.10
car	0.03	0	0	0	0.86	0.07	0.03	0.01
bus	0.01	0.01	0	0.03	0.03	0.85	0.07	0
train	0.05	0	0	0	0.02	0.04	0.79	0.10
subway	0.02	0	0	0	0.03	0.01	0.29	0.65

Figure 5: The confusion matrix of the proposed model on the validation set

Table 1: Accuracy of the proposed model on the validation set

Inputs	Sensors	Model	Accurac
FFT & FFT based features	Acc, Gyr, Mag	Residual IndRNN	77.32%
FFT & FFT based features	Acc, Gyr, Mag	Dense IndRNN	81.58%

The trained dense IndRNN model is used for the further transfer learning. The learning rate is set to 2×10^{-5} to finetune the model and the validation confusion matrix of transfer A and transfer B are shown in Figure 6 and Figure 7. After transfer learning, the accuracy of validation set increases to 86.94%, which means that cross-user transfer learning is useful for testing on the data from different users.

5 COMPUTATIONAL RESOURCES

- Program language: Python 3.7.0
- Framework: PyTorch 1.1
- GTX Titan Xp GPU
- Computing platform: Ubuntu18.04 with Intel Core E5-2640 V4 Processor (up to 2.40GHz)
- Model size: 43502KB
- Training time: about 18 hours
- Testing time: about 2540s

	still	walk	run	bike	car	bus	train	subway
still	0.87	0.03	0	0.02	0.01	0	0.02	0.05
walk	0.11	0.86	0	0.01	0.01	0	0	0.01
run	0	0.02	0.97	0	0	0	0	0
bike	0.05	0.07	0	0.83	0.01	0.02	0.02	0.01
car	0.03	0	0	0	0.93	0.01	0	0.01
bus	0.01	0.01	0	0	0.13	0.84	0.02	0
train	0.08	0.02	0	0.01	0.02	0.01	0.75	0.10
subway	0.01	0	0	0	0.01	0	0.09	0.89

Figure 6: The confusion matrix after transfer A

	still	walk	run	bike	car	bus	train	subway
still	0.90	0.08	0	0	0	0	0.01	0.01
walk	0.02	0.98	0	0	0	0	0	0
run	0	0.03	0.97	0	0	0	0	0
bike	0.09	0.04	0	0.86	0	0	0	0.01
car	0.08	0.01	0	0	0.89	0	0.01	0.01
bus	0.02	0.02	0	0.05	0.06	0.66	0.16	0.01
train	0.01	0	0	0	0.01	0	0.83	0.15
subway	0.05	0	0	0	0.03	0	0.12	0.82

Figure 7: The confusion matrix after transfer B

Table 2: Accuracy of the proposed model on the validation set with transfer learning and model fusion

Process	Accuracy	Final accuracy
Transfer A	86.28%	86.94%
Transfer B	87.60%	

6 CONCLUSION

This paper presents a long-term temporal recognition method with features extracted in both spatial and frequency domains. The features include the short-term temporal features in the spatial domain such as mean, maximum, minimum, and the short-term temporal features such as the FFT spectrums and the distribution statistic feature over the FFT spectrums. The long-term temporal features are further learned with the dense IndRNN, which is able to capture long-range patterns and keep long-term memory. Considering the behavior on different locations are usually different, the location of test data is further predicted to better utilize the features. With the recognition model trained on the training set with data from only one user, transfer learning is

adopted in the experiments, which further brings about 5.5% increase of the validation accuracy. Different data partition strategies are used to learn different models and model fusion is applied to, make the most of the validation data for transfer learning. As a result, our model achieved validation accuracy of 81.58% before transfer learning and 86.94% after that. The recognition result for the testing dataset will be presented in the summary paper of the challenge [8].

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2018YFE0203900) and the National Natural Science Foundation of China (No. 61901083).

REFERENCES

[1] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen,"Enabling reproducible research in sensor-based transportation mode recognition with the Sussex-Huawei dataset," IEEE Access 7 (2019): 10870-10891.

[2] H. Gjoreski, M. Ciliberto, L. Wang, F.J.O. Morales, S. Mekki, S. Valentin, and D. Roggen, "The University of Sussex-Huawei locomotion and transportation dataset for multimodal analytics with mobile devices," IEEE Access 6 (2018): 42592-42604.

[3] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2018. Independently recurrent neural network (indrnn): Building a longer and deeper rnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5457–5466.

[4] Shuai Li and Wanqing Li and Chris Cook and Yanbo Gao and Ce Zhu. 2019. Deep Independently Recurrent Neural Network (IndRNN).rXiv.cs.CV 1910.06251.

[5] L. Zheng, S. Li, Y. Gao, "Application of IndRNN for Human Activity Recognition - The Sussex-Huawei Locomotion-Transportation Challenge," In Proceedings of the 2019 ACM International Joint Conference and 2019 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, 869–872.

[6] L. Wang, H. Gjoreski, K. Murao, T. Okita, and D.l Roggen, "Summary of the Sussex-Huawei locomotion-transportation recognition challenge," in Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 1521-1530, 2018.

[7] L. Wang, H. Gjoreski, M. Ciliberto, P. Lago, K. Murao, T. Okita, and D. Roggen, "Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2019," in Proceedings of the 2019 ACM International Joint Conference and 2019 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, pp. 849-856, 2019.

[8] L. Wang, H. Gjoreski, M. Ciliberto, P. Lago, K. Murao, T. Okita,and D. Roggen. "Summary of the Sussex-Huawei locomotion-transportation recognition challenge 2020", Proceedings of the 2020 ACM International Joint Conference and 2020 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, 2020.

[9] Martin Gjoreski, Vito Janko, Nina Reščič, Miha Mlakar, Mitja Luštrek, Jani Bizjak, Gašper Slapničar, Matej Marinko, Vid Drobnič, and Matjaž Gams. 2018. Applying multiple knowledge to Sussex-Huawei locomotion challenge. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, 1488–1496.

[10] Zhuo S, Sherlock L, Dobbie G, Koh YS, Russello G, Lottridge D. REAL-Time Smartphone Activity Classification Using Inertial Sensors-Recognition of Scrolling, Typing, and Watching Videos While Sitting or Walking. Sensors (Basel). 2020;20(3):655. Published 2020 Jan 24. doi:10.3390/s20030655

[11] Vito Janko, Martin Gjoreski, Carlo Maria De Masi, Nina Reščič, Mitja Luštrek, and Matjaž Gams. 2019. Cross-location transfer learning for the sussex-huawei locomotion recognition challenge. In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct). Association for Computing Machinery, New York, NY, USA, 730–735.

[12] Yida Zhu, Fang Zhao, and Runze Chen. 2019. Applying 1D sensor DenseNet to Sussex-Huawei locomotion-transportation recognition challenge. In Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct). Association for Computing Machinery, New York, NY, USA, 873–877.

[13] Choi, Jun-Ho & Lee, Jong-Seok. (2019). EmbraceNet for activity: a deep multimodal fusion architecture for activity recognition. 693-698. 10.1145/3341162.3344871.

[14] Adriana Wilde, Robert Streeting, and Ed Zaluska. 2013. Unobstrusive human activity recognition using smartphones and Hidden Markov Models. Journal of Ambient Intelligence and Humanized Computing (2013).

[15] Martin Gjoreski, Vito Janko, Nina Reščič, Miha Mlakar, Mitja Luštrek, Jani Bizjak, Gašper Slapničar, Matej Marinko, Vid Drobnič, and Matjaž Gams. 2018. Applying multiple knowledge to Sussex-Huawei locomotion challenge. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers. ACM, 1488–1496.

[16] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. 2019. A fully trainable network with rnn-based pooling. Neurocomputing (2019).