Application of CNN for Human Activity Recognition with FFT Spectrogram of Acceleration and Gyro Sensors

Chihiro Ito
Tokyo Denki University
5 Senju-Asahi-cho, Adachi-ku,
Tokyo 120–8551, Japan

Masaki Shuzo
Tokyo Denki University
5 Senju-Asahi-cho, Adachi-ku,
Tokyo 120–8551, Japan

Xin Cao
Tokyo Denki University
5 Senju-Asahi-cho, Adachi-ku,
Tokyo 120–8551, Japan

Eisaku Maeda
Tokyo Denki University
5 Senju-Asahi-cho, Adachi-ku,
Tokyo 120–8551, Japan
* Corresponding author:
maeda.e@mail.dendai.ac.jp

UbiComp/ISWC'18 Adjunct, October 8–12, 2018, Singapore, Singapore

ACM ISBN 978-1-4503-5966-5/18/10 $15.00
https://doi.org/10.1145/3267305.3267517

Abstract
At the SHL recognition challenge 2018, Team Tesaguri developed a human activity recognition method. First, we obtained the FFT spectrogram from 60-second acceleration and gyro sensor data for each of six axes. A five-second sliding window was used for FFT processing. About 70% of the spectrogram figures from the Sussex-Huawei Locomotion-Transportation dataset were used for training data. Our model was based on CNN using FFT spectrogram images. After training for 50 epochs, F-measure was about 90% for acceleration data and 85% for gyro data. Next, considering the results of each sensor axis, to improve the recognition rate, we combined the information of multiple sensors. Specifically, we synthesized new images by combining the FFT spectrogram figures of two axes and the best combination condition was examined by correlation analysis. The highest score, 93% recognition, came from the vertically arranged images derived from the norm of acceleration and the y-axis gyro.

Author Keywords
Human activity recognition, SHL dataset, CNN, FFT spectrogram, Correlation analysis

ACM Classification Keywords
I.5.4. [Pattern Recognition]: Signal processing.

Introduction

Activity recognition from sensors has been attracting increased attention from various research communities. We participated in the Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge as Team Tesaguri. Our aim was to develop an algorithm for human activity recognition using the SHL dataset [1]. Nowadays, machine learning techniques have become familiar for beginners. From about 2015, advances in deep learning have given birth to powerful tools for human activity recognition [2]. In this paper, we propose recognition method applying the convolutional neural network (CNN) model, which is a well-known technique for image recognition [3]. In prior research on human activity recognition, only a few papers have examined the potential of CNN application with FFT spectrogram images [4, 5]. Here, we report the methodology in detail.

SHL Dataset

The SHL dataset was collected primarily to investigate the recognition of users' modes of locomotion and transportation from sensors in a mobile phone by means of machine learning methods and heuristics [6]. This dataset is a versatile annotated dataset of the modes of locomotion and transportation of mobile users. It was recorded over a period of seven months in 2017 by three participants engaging in eight different modes of transportation in a real-life setting in the United Kingdom. The dataset contains 750 hours of labelled locomotion data: *Car* (88 h), *Bus* (107 h), *Train* (115 h), *Subway* (89 h), *Walk* (127 h), *Run* (21 h), *Bike* (79 h), and *Still* (127 h). The dataset has multi-modal data from a body-worn camera and from four smartphones carried simultaneously at typical body locations.

SHL Recognition Challenge Task

The goal of the SHL recognition challenge (2018) for the machine learning/data science challenge was to recognize eight modes of locomotion and transportation (activities) from the inertial sensor data of a smartphone. The dataset used for this challenge comprised 271 hours of training data (5.5 GB by ZIP archive) and 95 hours of test data (1.9 GB) provided by the organizers.

The training dataset contains 21 plain text files corresponding to various sensor channels and the labels: Accelerometer: Acc_x.txt, Acc_y.txt, Acc_z.txt, Gyroscope: Gyr_x.txt, Gyr_y.txt, Gyr_z.txt, Magnetometer: Mag_x.txt, Mag_y.txt, Mag_z.txt, Linear accelerometer: LAcc_x.txt, LAcc_y.txt, LAcc_z.txt, Gravity: Gra_x.txt, Gra_y.txt, Gra_z.txt, Orientation: Ori_w.txt, Ori_x.txt, Ori_y.txt, Ori_z.txt, Pressure: Pressure.txt, Label: Label.txt, and Order: train_order.txt.

The SHL recognition challenge participants were requested to develop an algorithm pipeline to process the sensor data, create models, and output the recognized activities.

Preprocessing

First, we used only acceleration and gyro sensor data for recognition in this paper. The training dataset provided by organizers was separated into our training data for model creation and our testing data. We omitted 650 files that had multi labeled data. We randomly assigned 70% of the single labeled files

(15,660) for our training and the remaining 30% (4,695) for our validation.

Subsequently, the spectrogram of the accelerometer and gyro sensor data is a three dimensional representation of changes in the energy content of a signal as a function of frequency and time. We use the spectrogram representation as the input of activity recognition models, since spectrograms of speech waveforms are always used as informative features in acoustic modeling. All axes of the acceleration and gyro sensor data were transformed into a fast Fourier transform (FFT) spectrogram by using the Matplotlib as a Python calculation option.

Spectrogram images for 60 seconds with a five-second window were obtained. The FFT window overlapped every 100 ms of the sensor sampling rate. Images with the parula colormap were saved as JPEG format with the size of 1400 $\times$ 1106 pixels. A logarithmic axis for frequency was used so as to distinguish lower frequency signals. These images were resized to 48 $\times$ 48 pixels before input to CNN.

At the stage of figure preparation, as we already know, most human activity signals are in the range of less than 20 Hz. We confirmed that the logarithmic scale of the frequency axis showed better results than the linear scale axis.

CNN Model from FFT Spectrogram

Next, we describe the overall architecture of our CNN, as shown in Fig. 1. The three convolutional layers are followed by two fully connected layers.

The first convolutional layer takes the 48 $\times$ 48 spectrogram image and applies 16 3 $\times$ 3 filters with a zero-padding of 1. This is followed by a ReLU and max-pooling resulting in a 24 $\times$ 24 image volume. The second convolutional layer takes the 24 $\times$ 24 image volume and applies 32 3 $\times$ 3 filters with a zero-padding of 1. This is followed by a ReLU and max-pooling resulting in a 12 $\times$ 12 image. The third convolutional layer takes the 12 $\times$ 12 image volume and applies 64 3 $\times$ 3 filters. This is followed by a ReLU and max-pooling resulting in a 5 $\times$ 5 image volume.

The remaining two layers are fully connected layers. The first reduces the size to 128 and then applies a ReLU. The second reduces to 8 and then applies a ReLU.

We used with the TensorFlow Keras as a backend in the Python environment for training. Most of the components of our network model are available in recent deep learning frameworks. Here, we tried to manually program under Keras.

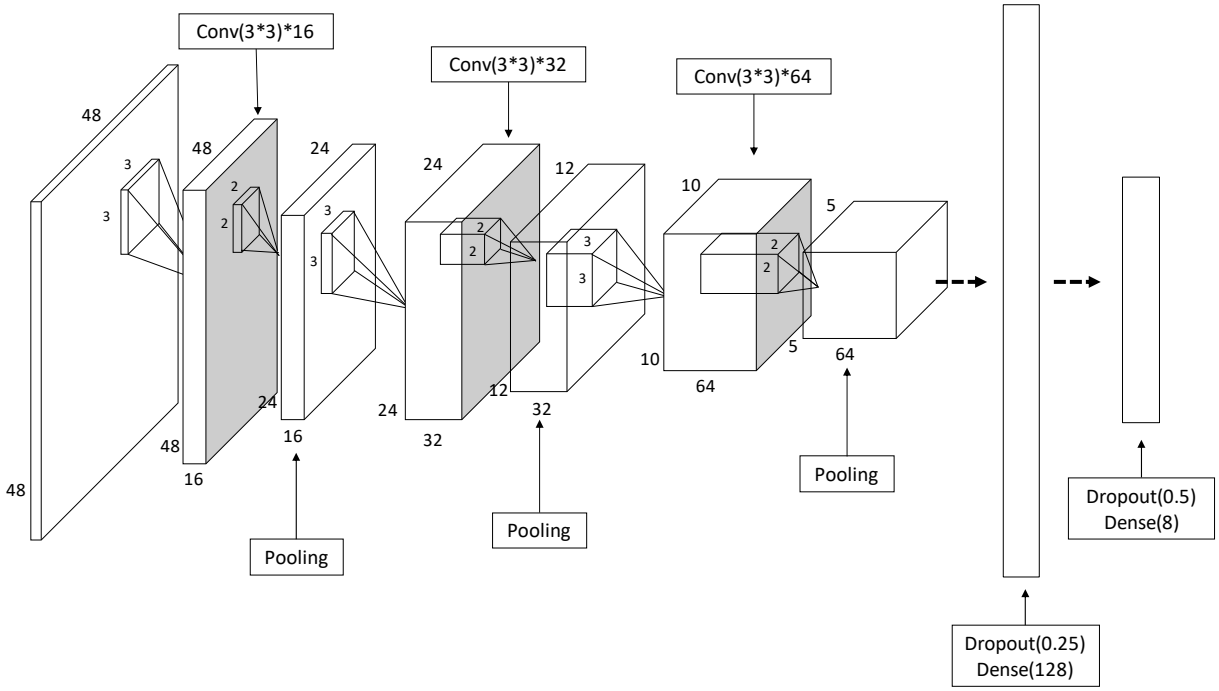

Figure 1: Proposed CNN model for human activity recognition using FFT spectrogram images.

Training Results on Single Axis Dataset

We obtained training results from each of the three axes of acceleration and gyro sensor data by the above method. As an example, a typical learning curve for the Acc_x dataset is shown in Fig. 2. The loss functions dropped (left) while the accuracy rose (right). The closeness of the validation curves to the training curves indicates that the network was generalizing well to the validation data.

Finally, the accuracy increased every epoch without over-fitting and resulted in 0.974 at epoch 50 in the case of only Acc_x. In other words, F-measure was 0.905.

We also show the confusion matrix by our proposed CNN method on our validation dataset, which was extracted from the provided training data (Table 1). We found a few miss-recognitions in *Car-Bus* and *Train-*

Subway. We assume it was difficult to recognize these clearly by using motion sensors because such locomotion data have very similar modalities.

The F-measure results for the y-axis and for the z-axis were almost the same: 0.911 and 0.901, respectively. As we do not care greatly about the sensor direction, we also tested the norm of acceleration (Acc_norm); results showed that the accuracy (0.980) and F-measure (0.928) were both higher than each single axis results.

Next, gyro data was examined in the same way. The F-measure of each axis was 0.847 for Gyr_x, 0.878 for Gyr_y, and 0.837 for Gyr_z. We were not able to determine why the Gyr_y result was slightly higher than the others.

Correlation Analysis of Each Axis

In order to improve the recognition result, we tried to find more effective images for our CNN model. The input image of an FFT spectrogram is able to reconstruct itself as multiple-axis images. In this paper, considering the computational resources, a double size image derived from two FFT images would be suitable for us. We utilized correlation analysis to determine the best combination of two images.

Although the modalities of each axis are the same in a single sensor, those of hybrid sensors may be different. Therefore, first we checked the correlation of the acceleration and gyro sensor data of each axis. Table 2 lists the correlation analysis results obtained using the *corrcoef* function of the NumPy module in the Python environment. The correlation coefficient between each axis in the acceleration data ranged from 91.4 to

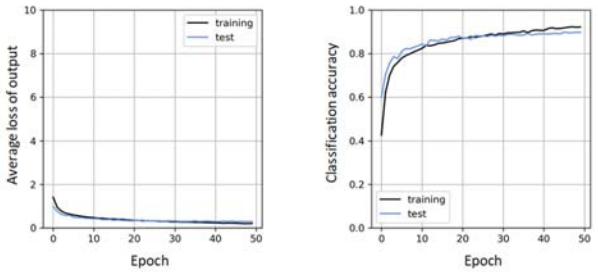

Figure 2: Learning curve in case of Acc_x.

	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)
P(1)	584	0	0	1	5	4	54	32
P(2)	1	592	0	2	0	0	0	0
P(3)	0	1	196	2	0	0	0	0
P(4)	1	8	0	595	1	2	0	0
P(5)	7	1	0	1	710	12	8	11
P(6)	2	1	0	2	9	552	10	15
P(7)	17	2	0	0	2	16	567	97
P(8)	26	2	0	0	5	22	99	418

Table 1: Confusion matrix in case of Acc_x. A(*i*) and P(*i*) stand for accuracy and precision, respectively. The number *i* refers to activity: 1: *Still*, 2: *Walk*, 3: *Run*, 4: *Bike*, 5: *Car*, 6: *Bus*, 7: *Train*, 8: *Subway*.

	Acc_x	Acc_y	Acc_z	Gyr_x	Gyr_y	Gyr_z	Acc_norm
Acc_x	1.000						
Acc_y	0.920	1.000					
Acc_z	0.921	0.915	1.000				
Gyr_x	0.753	0.781	0.742	1.000			
Gyr_y	0.827	0.819	0.809	0.923	1.000		
Gyr_z	0.728	0.735	0.692	0.929	0.906	1.000	
Acc_norm	0.922	0.935	0.944	0.729	0.786	0.684	1.000

Table 2: Correlation coefficient.

92.1%. Similarly, the correlations in the gyro data ranged from 90.6 to 92.9%. Combinations with higher results would not be useful for recognition improvement, but the lower combination would be beneficial for our model. Therefore, for combinations of axes between Acc and Gyr, we confirmed that the correlations ranged between 69.2 and 82.7%.

According to the confusion matrix of the CNN training in Gyr_x and Gyr_z, these results had weak recognition for *Train-Subway*. If Acc data is combined with the data of Gyr_x or Gyr_z, we cannot expect much improvement. Finally, on the basis of the correlation analysis and confusion matrix results, we decided to combine the norm of the acceleration data with the Gyr_y data. The two resultant FFT spectrogram images were arranged vertically (Fig. 3).

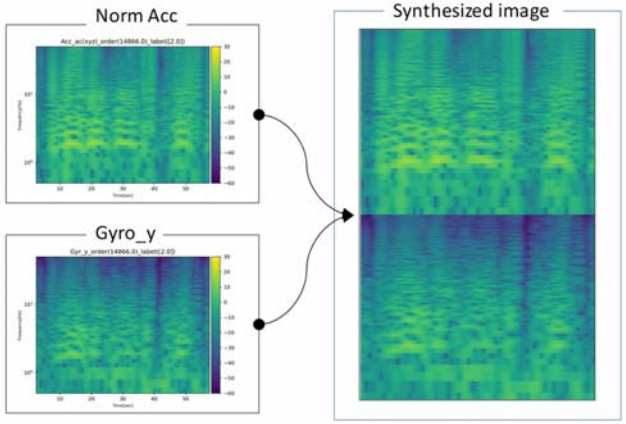

Figure 3: Example of synthesized image (right). FFT spectrogram images of norm Acc (left-above) and Gyr_y (left-below) were vertically arranged.

Results of Synthesized Image Input

Next, we trained our CNN model again using the synthesized image data discussed above. The results of the learning curve and confusion matrix are shown in Fig. 4 and Table 3. The F-measure was 0.934 after 50 epochs. As the basic score before synthesizing was high enough, it is difficult to clarify exactly how this was effective. All we can say for certain is that the score definitely increased. In previous methods, the recognition of *Train* and *Subway* was relatively worse than the other activity recognitions. From the confusion matrix, we found that the results of *Train* and *Subway* were improved.

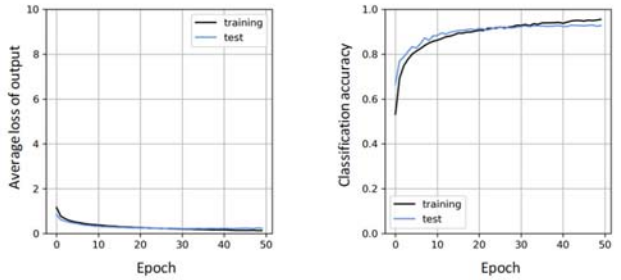

Figure 4: Learning curve in case of synthesized image.

	A(1)	A(2)	A(3)	A(4)	A(5)	A(6)	A(7)	A(8)
P(1)	627	6	0	2	6	10	63	26
P(2)	1	599	0	2	0	0	0	0
P(3)	0	0	196	0	0	0	0	0
P(4)	0	1	0	598	0	0	0	0
P(5)	2	0	0	0	705	10	1	4
P(6)	2	0	0	0	9	553	4	7
P(7)	3	0	0	0	6	10	621	77
P(8)	3	1	0	1	6	25	49	459

Table 3: Confusion matrix in case of synthesized image.

Additional Work

In this paper, we mainly focused on preprocessing the data and clarifying the data characteristics. We also concentrated on the creation of the learning model. The model used in this paper is not a deep learning model. Here, we examined how the accuracy rate was converted to a deeper neural network. The model used was AlexNet and VGG16. Input was Acc_(x, y, z) and Gyr_(x, y, z) simultaneously, 6-channel images were parallel to the fully connected layer, and 6-channel neurons were merged into the fully connected layer. This brought the accuracy of the randomly selected test data to 92%. In this study, deep learning did not yield satisfactory results. We feel this is linked to the loss of features stemming from the number of learnings and the compression of the image pixel.

Conclusion

In the SHL recognition challenge, we developed a CNN application for Human Activity Recognition with an FFT spectrogram of acceleration and gyro sensors. After training for 50 epochs, the F-measure was about 90% for acceleration data and 85% for gyro data. Considering the validation results of each sensor axis, to improve the recognition rate, we combined the information from multiple sensors. Specifically, we created a new figure by combining two axes of FFT spectrogram figures and then investigated the best combination condition. We found that combining the absolute value of acceleration and the y-axis gyro resulted in a 93% recognition score. This recognition result for the testing dataset will be presented in the summary paper of the challenge [8].

Computational Resources

For this SHL recognition challenge, we used two entry models of Windows laptop PCs: HP ProBook 430G5 (Core i5-7200U 2.5–3.1GHz, 8GB, SSD) and HP ProBook 450G3 (Core i7-6500U 2.5–3.1GHz, 8GB, HDD). 12 machines were used in total. Our CNN model was developed by Keras on the Pycharm Python environment. Using this environment, it took about 90 hours to create the model.

Acknowledgements

This work was supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). Prof. M. Kawakatsu gave us special lectures and environments of machine learning technology.

References

1. H. Gjoreski, M. Ciliberto, F. J. Ordoñez Morales, D. Roggen, S. Mekki, and S. Valentin. "A versatile annotated dataset for multimodal locomotion analytics with mobile devices." In Proc. ACM Conference on Embedded Networked Sensor Systems. 2017.
2. H. Gjoreski, M. Ciliberto, L. Wang, F. J. Ordonez 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, 2018, [In Print], DOI: 10.1109/ACCESS.2018.2858933.
3. J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu. "A survey: Deep learning for sensor-based activity recognition." Pattern Recognition Letters. 2017.
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks." In Proc. 25th International Conference on Neural Information Processing Systems. Vol. 1, pp. 1097–1105. 2012.

5. Y. Mohammad, K. Matsumoto, and K. Hoashi. "Primitive activity recognition from short sequences of sensory data." Applied Intelligence. 2018.
6. M. A. Alsheikh, A. Selim, D. Niyato, L. Doyle, S. Lin, and H-P. Tan. "Deep activity recognition models with triaxial accelerometers." In Proc. 2016 AAAI Workshop. 2016.
7. M. Ciliberto, F. J. Ordoñez Morales, H. Gjoreski, D. Roggen, S. Mekki, and S. Valentin. "High reliability Android application for multidevice multimodal mobile data acquisition and annotation." In Proc. ACM Conference on Embedded Networked Sensor Systems. 2017.
8. L. Wang, H. Gjoreski, K. Murao, T. Okita, and D. Roggen. "Summary of the Sussex-Huawei Locomotion-Transportation recognition challenge." Proc. 6th International Workshop on Human Activity Sensing Corpus and Applications (HASCA2018). Singapore, Oct. 2018.