Human Activity Recognition Using Multi-input CNN Model with FFT Spectrograms

Kei Yaguchi, Kazukiyo Ikarigawa, Ryo Kawasaki, Wataru Miyazaki, Yuki Morikawa, Chihiro Ito, Masaki Shuzo, Eisaku Maeda
Tokyo Denki University, 5 Senju-Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan
{16jk249,17aj008,17aj045,17aj141,17aj145,20jkm03}@ms.dendai.ac.jp
{shuzo,maeda.e}@mail.dendai.ac.jp

ABSTRACT

An activity recognition method developed by Team DSML-TDU for the Sussex-Huawei Locomotion-Transportation (SHL) recognition challenge was descrived. Since the 2018 challenge, our team has been developing human activity recognition models based on a convolutional neural network (CNN) using Fast Fourier Transform (FFT) spectrograms from mobile sensors. In the 2020 challenge, we developed our model to fit various users equipped with sensors in specific positions. Nine modalities of FFT spectrograms generated from the three axes of the linear accelerometer, gyroscope, and magnetic sensor data were used as input data for our model. First, we created a CNN model to estimate four retention positions (Bag, Hand, Hips, and Torso) from the training data and validation data. The provided test data was expected to from Hips. Next, we created another (pre-trained) CNN model to estimate eight activities from a large amount of user 1 training data (Hips). Then, this model was fine-tuned for different users by using the small amount of validation data for users 2 and 3 (Hips). Finally, an F-measure of 96.7% was obtained as a result of 5-fold-cross validation.

CCS CONCEPTS

• **Human-centered computing** → User models; • **Computing methodologies** → Machine learning.

KEYWORDS

Human activity recognition; SHL dataset; CNN; FFT spectrogram

ACM Reference Format:
Kei Yaguchi, Kazukiyo Ikarigawa, Ryo Kawasaki, Wataru Miyazaki, Yuki Morikawa, Chihiro Ito, Masaki Shuzo, Eisaku Maeda. 2020. Human Activity Recognition Using Multi-input CNN Model with FFT Spectrograms. In *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 '20 Adjunct), September 12–16, 2020, Virtual Event, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3410530.3414342

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/10.1145/3410530.3414342

1 INTRODUCTION

Activity recognition from sensors has been gaining attention from various research communities. Our team DSML-TDU has participated in the SHL Transportation recognition challenge since 2018 [4][5]. Sussex-Huawei Locomotion-Transportation (SHL) dataset includes sensor information such as acceleration, gyro, and magnetic sensors obtained by a smartphone terminal carried by hand, at the hips, on the torso or in a bag during locomotive activities such as still, walking, running, and biking, as well as riding a car, bus, train, and the subway. The objective of the 2020 SHL recognition challenge was to identify eight different movement states of different users.

In this paper, we propose a recognition method that applies a convolutional neural network (CNN) model, a well-known technique for image recognition. In the field of human activity sensing, an example of generating spectrograms from microvibrations propagating through the user's arm and infering the user's context using a CNN model has been reported [7]. In our previous challenge in 2019, we developed a CNN application with Fast Fourier Transform (FFT) spectrogram using acceleration and data from a gyro sensor mounted on a subject's bag, hips, and torso. In this 2020 challenge, our models were developed to fit various users equipped with sensors in specific positions.

2 SHL DATASET AND TASK

The SHL dataset was collected primarily to investigate the recognition of users' means of locomotion and transportation from mobile phone sensors using machine learning methods and heuristics [1][10]. This versatile annotated dataset of mobile users' means of locomotion and transportation was recorded over a seven-month period in 2017. Three participants engaged in eight different modes of transportation in a real-life setting in the United Kingdom. The dataset contains 750 hours of labeled 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). Multi-modal data was captured by a body camera and four smartphones (HUAWEI Mate 9) carried simultaneously at areas of the body where a phone is typically held (in hand, torso, hips, and in a bag). All sensor data consists of the following: accelerometer (x, y, z), gravity (x, y, z), gyroscope (x, y, z), linear accelerometer (x, y, z), magnetometer (x, y, z), orientation (quaternions), and pressure.

The SHL Challenge 2020 was carried out using part of the SHL dataset. The goal of the challenge for machine learning/data science was to recognize eight modes of locomotion and transportation

from a mobile phone's sensor data. The dataset used for this challenge comprised 59 days of training data for one user (user 1), six days of validation data for two other users (users 2 and 3), and 40 days of test data for the latter two users. Both the training and validation datasets had sensor location information, while the test dataset did not.

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

Our approach was to identify the state of movement using the valid classifier model for the retention position after identifying the retention position of the terminal (section 5). Due to insufficient validation data (users 2 and 3) for predicting the test data (users 2 and 3), we created a pre-learning model using the training dataset and fine-tuned using the validation data in order to create a model more suitable for predicting test data (section 6).

3 PREPROCESSING

First, any data containing gaps or transitional states within the provided five-second segments were omitted. Information from the three sensors (linear acceleration, gyro, and magnetic sensor) were used in our solution. We accounted for the shaking of the smartphone due to the difference in holding position and movement state which appeared in the acceleration and gyro. Wang [10] suggested that combining magnetic sensor can improve performance. We pre-processed the three sensors with the following.

Step 1. We calculated each Euclidean norm from the three axes of the linear acceleration, gyro, and magnetic sensor.

$$m = \sqrt{x^2 + y^2 + z^2} \tag{1}$$

Step 2. We used the mean and standard deviation of user 1 to standardize the training data, and we standardized the validation data and test data using the mean and standard deviation of users 2 and 3.

Step 3. After standardizing the sensor data, all axes of the linear acceleration, gyro, and magnetic sensor data were transformed into an FFT spectrogram. As reported in previous challenges papers [4][5], the spectrogram representation was used as the input for activity recognition models. Spectrogram for the data (5-second durations at 100 Hz) were obtained with a 2.5-second sliding window (256 data points / 10 sampling point overlap). Example images are shown in Figure. 1.

4 CNN MODEL FROM FFT SPECTROGRAM

Next, we describe the overall architecture of our CNN which is shown in Figure. 2. The input data is an FFT spectrogram from the 5-second sensor data. The three convolutional layers are followed by three fully connected layers. The first convolutional layer takes the resized 128 × 25 spectrogram and applies sixteen 5 × 5 filters with a zero-padding of 1. This is followed by a Rectified Linear Unit (ReLU) function and max-pooling, resulting in a 64 × 12 image volume. The second convolutional layer takes the 64 × 12 image volume and applies thirty-two 5 × 5 filters with a zero-padding of 1. This is followed by a ReLU function and max-pooling, resulting in a 32 × 6 image. The third convolutional layer takes the 32 × 6 image volume and applies sixty-four 5 × 5 filters. This is followed by a

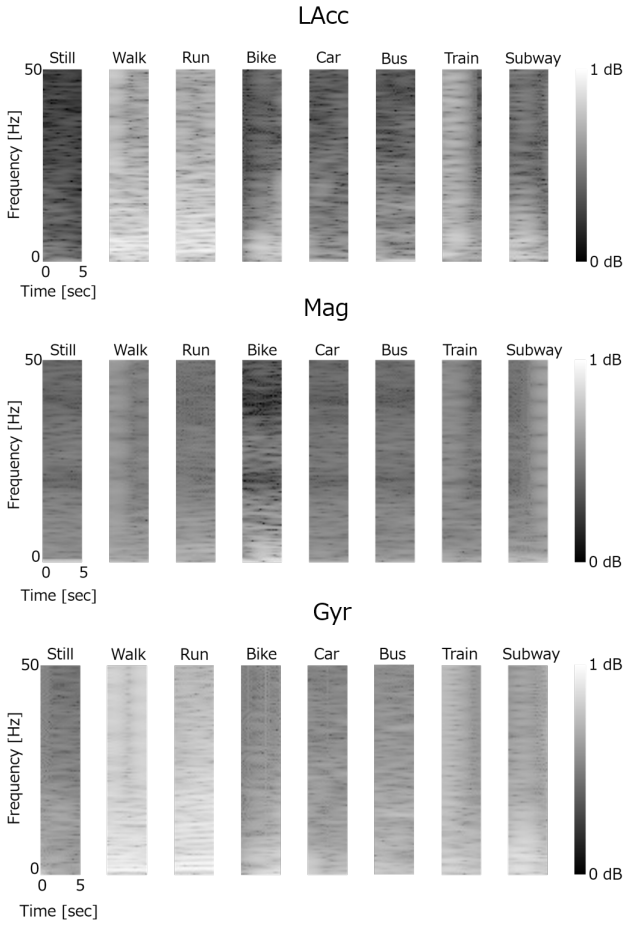

Figure 1: Example of spectrogram

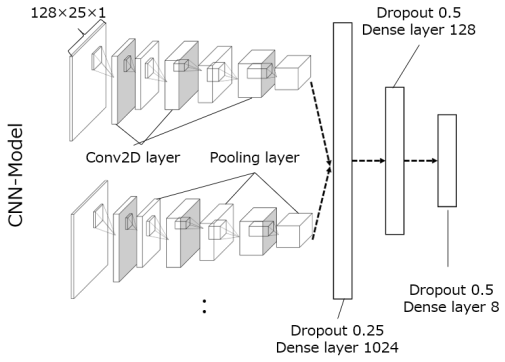

Figure 2: CNN model

ReLU and max-pooling, resulting in a 14 × 1 image volume. The remaining two layers are fully connected layers. The first reduces the size of the image to 1024 and then applies a ReLU. The second reduces the size to 128 and then applies a ReLU. The third reduces the size to 8 and then applies a Softmax function. The conditions

for learning are as follows, epoch: 37, batch size: 128, iteration: 17.9. We used TensorFlow's Keras API as a backend in the Python environment for training. Most of the components of our network model are available in recent deep learning frameworks.

Although we also considered at deeper-structured CNN models, such as ResNet [2][3] and VGG [8], they did not show any improvement in accuracy despite the long training time.

5 RETENTION POSITION ESTIMATION

Since the retention position of the challenge test data is unknow, we first created the estimation model for retention position using our CNN model. In this section, we did not use Mag_norm; we only used spectrograms generated from Acc_norm and Gyr_norm as input. We considered that magnetic sensors, which are influenced by the surrounding environment rather than the user's movements, help classify vehicles but do not contribute to classifying retention positions.

The retention position was labeled for the data measured from each position (Bag, Hand, Hips, and Torso). Classification was performed using this label as an objective variable.

Then we trained on user 1's training data using the CNN model in (section 4). The estimation results for the validation data of user 2 and 3 are shown in Figure 3 and Table 1 respectively.

Finally, we estimated the retention position for challenge test data from the same two users and verified that 99% of the data was measured from Hips.

6 STATE ESTIMATION

Given the results of the estimated retention position in the previous section, we decided to use only Hips data for training and validation data to create the model for estimating the state of movement.

In this section, three sensors (LAcc, Gyro, and Mag) were used for discrimination. Here, by our preliminarily testing, the recognition rate with using all single axes of sensors were better than one with Euclidean norm data described in section 3 (data not shown). Also, we have known the North-East-Down (NED) coordination system of sensor axes works well on activity recognition based on the report on previous challenge result [6]. So, the transformed each sensor axes data was pre-processed according to Steps 2 and 3 in section 3. The coordination conversion to NED system was performed by multiplying the smartphone coordinate system data with the rotation matrix R_{NB}.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix}_N = R_{NB} \begin{bmatrix} x \\ y \\ z \end{bmatrix}_B \tag{2}$$

Here, R_{NB} was given with quaternions $[q_w, q_x, q_y, q_z]$ as followed,

$$R_{NB} = \begin{bmatrix} 1-2(q_y^2+q_z^2) & 2(q_xq_y-q_wq_x) & 2(q_xq_z+q_wq_y) \\ 2(q_xq_y+q_wq_z) & 1-2(q_x^2+q_z^2) & 2(q_yq_z-q_wq_x) \\ 2(q_xq_z-q_wq_y) & 2(q_yq_z+q_wq_x) & 1-2(q_x^2+q_y^2) \end{bmatrix}. \tag{3}$$

Training was performed by inputting 9-axis FFT spectrograms into the CNN model described in Section 4. First, we trained the model on the user 1 training data. The average F-measure was 67.9% when this model was evaluated using validation which comprised data for users 2 and 3. The confusion matrix and classification report are shown in Figure 4 and Table 2 respectively.

Next, this pre-learning model was fine-tuned using the data on user 2 and 3 (validation dataset), resulting in an average F-measure of 96.7%. Cross validation was performed by dividing the Validation data into five parts, the results of which are shown in Figure 5 and Table 3 respectively.

Ground Truth \ Predicted	Bag	Hand	Hips	Torso
Bag	20394	96	4	2455
Hand	274	22671	0	3
Hips	0	6	22942	0
Torso	7132	0	0	15816

Figure 3: Confusion matrix of sensor location (The colored bar shows the ratio when the sum of each line is 1.)

Table 1: Estimation of retention position

Retention position	F-measure	Precision	Recall
Bag	0.8037	0.7336	0.8887
Hand	0.9917	0.9955	0.9879
Hips	0.9997	0.9998	0.9997
Torso	0.7674	0.8655	0.6892

Ground Truth \ Predicted	Still	Walk	Run	Bike	Car	Bus	Train	Subway
Still	5539	33	0	2	5	13	241	103
Walk	533	3864	1	304	0	2	350	136
Run	4	28	340	178	0	0	1	0
Bike	229	81	0	926	14	25	844	281
Car	545	4	0	11	844	765	1705	215
Bus	17	10	0	3	46	1335	362	57
Train	144	24	0	2	42	149	3449	543
Subway	248	13	0	1	11	21	1782	2260

Figure 4: Confusion matrix of pre-trained model (The colored bar shows the ratio when the sum of each line is 1.)

Table 2: F-measure for each activity determined by pre-trained model

Activity	F-measure
Still	0.8418±0.0043
Walk	0.8524±0.0167
Run	0.7220±0.0253
Bike	0.6717±0.0979
Car	0.4686±0.0772
Bus	0.6385±0.0302
Train	0.5918±0.0415
Subway	0.6419±0.0380
macro avg	0.6786±0.0272

7 SUBMISSION RESULTS

The result of our retention position estimation indicated that 99.9% of the test data was estimated to be measured from Hips. Only the data measured from Hips was used for training for model creation. The model that learned the training data had an F-measure of 67.9%

		Predicted							
		Still	Walk	Run	Bike	Car	Bus	Train	Subway
Ground Truth	Still	5669	152	0	18	20	8	23	46
	Walk	145	5014		19	2	1	7	2
	Run	0	11	540	0	0	0	0	0
	Bike	43	89	0	2259	2	3	3	1
	Car	51	6	0	2	4024	3	0	3
	Bus	5	22	0	10	9	1771	8	5
	Train	57	32	0	13	3	6	4164	78
	Subway	64	14	0	7	4	4	78	4165

Figure 5: Confusion matrix of transfer learning (The colored bar shows the ratio when the sum of each line is 1.)

Table 3: F-measure for each activity determined by transfer learning

Activity	F-measure
Still	0.9473±0.0031
Walk	0.9524±0.0063
Run	0.9899±0.0101
Bike	0.9556±0.0030
Car	0.9872±0.0022
Bus	0.9769±0.0069
Train	0.9644±0.0021
Subway	0.9646±0.0029
macro avg	0.9673±0.0029

when predicting the validation data. The validation data and test data may have been less reliable than the expected results due to individual differences, as the user measured was different from that in the training data. To address this shortcoming and improve test data predictions, we fine-tuned this model using the validation data. DSML teams transfer model for activity recognition is expected to produce a high F-measure of more than 96.7%. The recognition result for the test data will be presented in the summary paper for the challenge [9].

COMPUTER RESOURCES

In this challenge, we used a Supermicro GPU computer (CPU:Intel Xeon CPU E5-1620 v4 @ 3.50 GHz × 8 / GPU: GeForce GTX 1080 Ti × 1 / MEM: 125.8GB). Our CNN model was developed by Keras in a Python environment. It took about two hours to create a predictive model (including 10 minutes fine tuning), and then it took five minutes to evaluate the test dataset.

REFERENCES

[1] H. Gjoreski, M. Ciliberto, L. Wang, F. J. Ordonez Morales, S. Mekki, S. Valentin, and D. Roggen. 2018. The University of Sussex-Huawei Locomotion and Transportation Dataset for Multimodal Analytics With Mobile Devices. *IEEE Access* 6 (2018), 42592–42604.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity Mappings in Deep Residual Networks. arXiv:1603.05027 [cs.CV]

[4] Chihiro Ito, Xin Cao, Masaki Shuzo, and Eisaku Maeda. 2018. Application of CNN for Human Activity Recognition with FFT Spectrogram of Acceleration and Gyro Sensors. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3267305.3267517

[5] Chihiro Ito, Masaki Shuzo, and Eisaku Maeda. 2019. CNN for Human Activity Recognition on Small Datasets of Acceleration and Gyro Sensors Using Transfer Learning. 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*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3341162.3344868

[6] Vito Janko, Nina Rešçiç, Miha Mlakar, Vid Drobnič, Matjaž Gams, Gašper Slapničar, Martin Gjoreski, Jani Bizjak, Matej Marinko, and Mitja Luštrek. 2018. A New Frontier for Activity Recognition: The 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*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3267305.3267518

[7] Gierad Laput and Chris Harrison. 2019. Sensing Fine-Grained Hand Activity with Smartwatches. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3290605.3300568

[8] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV]

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

[10] L. Wang, H. Gjoreski, M. Ciliberto, S. Mekki, S. Valentin, and D. Roggen. 2019. Enabling Reproducible Research in Sensor-Based Transportation Mode Recognition With the Sussex-Huawei Dataset. *IEEE Access* 7 (2019), 10870–10891.