An Investigation of Recurrent Neural Network for Daily Activity Recognition using Multi-modal Signals

Akira Tamamori*, Tomoki Hayashi†, Tomoki Toda‡, and Kazuya Takeda†
* Institute of Innovation for Future Society, Nagoya University, Nagoya, Japan
E-mail: tamamori@g.sp.m.is.nagoya-u.ac.jp
† Graduate School of Information Science, Nagoya University, Nagoya, Japan
E-mail: hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp, kazuya.takeda@nagoya-u.jp
† Information Technology Center, Nagoya University, Nagoya, Japan
E-mail: tomoki@icts.nagoya-u.ac.jp

Abstract*—Our aim is to build a daily activity surveillance system for elderly people. In this study, we develop Deep Neural Network (RNN) based approach for human activity recognition task by using multi-modal (acoustic and acceleration) signals. In a recent study, the effectiveness of Feed-Forward Neural Network (FF-NN) has been shown for daily activity recognition (DAR) task. However, the length of temporal context to be considered was limited although an actual daily activity event may span over continuous several seconds or minutes. Moreover, from a perspective of practical use, it will be needed to consider adaptation method to obtain satisfactory recognition performance for multiple users even when the only small amount of training data is available for each user. In this study, we evaluate the effectiveness applying Recurrent Neural Network based on Long Short-Term Memory (LSTM-RNN) to DAR. We also evaluate the effectiveness of applying LSTM-RNN with projection layer (LSTMP-RNN) for subject adaptation: (1) by applying RNN instead of FF-NN, much longer temporal context can be considered in daily activity event spanned over several seconds or minutes, and (2) by introducing LSTMP-RNN, an adaptation method can be realized, which can mitigated over-fit problem while maintaining the recognition performance. The results of experiments on DAR demonstrated that: (1) applying LSTM-RNN is effective compared to FF-NN, and (2) applying LSTMP-RNN is more effective than LSTM-RNN when limited amount data is available.

I. Introduction

In 2016, the percentage of elderly people over 65 years old in population of Japan is 26.7% and will be increased by 0.7% compared to 2015. Aging society in Japan is very rapidly proceeding. By 2035, it is estimated that more than one third of the Japanese population will be over 65 years old [1]. Being faced with such unprecedented super-aging society, it will be required to build a society that the elderly themselves can move safely and securely at anytime and anywhere. Moreover, a society will be also required that sustainable social participation and their activities can be promoted, and they can adopt an active and fresh lifestyle as long as they wish. We expect that the opportunities for the elderly to go out will increase if the above society realize: their ability, activity and memory can be sustained in a personalized style by sensing, recording and understanding their daily activities. Our aim is to develop and utilize an daily activity surveillance system to increase such opportunities.

Figure 1 shows an overview of our target system. Our target system is designed to run in a online manner; The system utilize a smartphone to collect environmental sound signals and user's triaxial acceleration signals continuously. Those signals are send to a server and the system sends the recognition result, user's activity, to their smart-phone. The history of the user's activity can be viewed through a graphical user interface installed on the smart-phone as an application. The users can browse their own activity history. The system will interact with a user through the interface based on the activity history; if a user's recent exercise activities are decreasing with in a term, e.g., a week, then the system will notify the decrease of exercise activities and promote the user to go out by sending a message, like "Today is sunny, then let's jogging in nearby park". The daily activity recognition (DAR) is one of the fundamental techniques to realize such system. In this study, we aim to develop the DAR technique as building blocks for an automatic daily activity surveillance system.

It is necessary to use signals obtained from various sensors in order to recognize the daily activities. Mainly, previous studies for DAR can be divided into two approaches: (i) to recognize user's activities through sensors embedded in indoor environment [2]–[4], or (ii) to recognize user's activities using sensors attached to user's body [5]–[11]. The advantage for the former approach is that composite and highly abstract activities can be recognized. However, the sensor to be embedded is expensive. The advantage for the latter approach is that it does not require embedding sensors into an environment and the cost is cheaper than the former. However, it requires not a few wearing cost and we consider that some users feel obtrusiveness, especially for daily surveillance system. In these previous studies, target activity was performed in a simulated environment and the types of daily activities is also limited. It is difficult to achieve sufficient recognition performance to

identify complicated activities; e.g., cooking, eating, cleaning rooms, etc. Furthermore, large amounts of user-specific data will be required to build a classifier, therefore those recognition performance for highly abstract activities is still not satisfactory.

In this study, we focus on the wearable sensor approach concurrently using a smart-phone, for this is cheap and accessible device. Many researchers also have focused on the smart-phone based approach [12], [13]. Another previous study [14] proposed a technique to recognize environmental sound signals and acceleration signals through Feed-Forward Neural Network (FF-NN) and evaluated it on a dataset built by Nishida et al. [15]. This database contains sensor signals recorded in a a one-room studio apartment, with a small video camera and a smartphone over 48 continuous hours. They showed the effectiveness of FF-NN compared with popular classifier, e.g., Gaussian Mixture Model and Support Vector Machine [16]. However, the recognition performance was not satisfactory in a practical use. From a perspective of practical use, subject adaptation method should also be considered to obtain satisfactory recognition performance in a case that only small amount of data for one subject can be available. The preliminary study [17] confirmed the effectiveness of the subject adaptation for DAR based on FF-NN. However, the recognition performance was also not satisfactory, and we realized that further improvement will be required.

In this study, we first evaluate the effectiveness of applying recurrent neural network based on Long Short Term Memory (LSTM-RNN) for DAR on dataset built by Nishida et al. Next, we apply LSTM-RNN with recurrent projection layer (LSTMP-RNN) [18] for subject adaptation task on the same database. It is expected that LSTM-RNN can capture longer temporal context than FF-NN and LSTMP-RNN can mitigate over-fit problem by reducing the number of network parameters while maintaining the recognition performance. The results of DAR experiments conducted on the dataset showed that applying the LSTM-RNN and BLSTM-RNN was more effective than FF-NN. Moreover, LSTMP-RNN achieved better performance than LSTM-RNN on the adaptation task.

The rest of the paper is organized as follows: Section III describes the framework of DAR in this study. Next, Section IV describes the architectures of recurrent neural network applied to DAR, then in Section V, subject adaptation method for RNN-based DAR is described. Section VI shows results of DAR experiment. Finally, the conclusion and future works are described in Section VII.

II. Nagoya-COI daily activity database

In this section, we describe the Nagoya Center Of Innovation (Nagoya-COI) daily activity database which we used to evaluate the performance of classifier.

A. Recording condition

The outline of recording condition is shown in Table I, and the equipment of subject is shown in Fig2. The recording environment is a one-room studio apartment. Each subject can

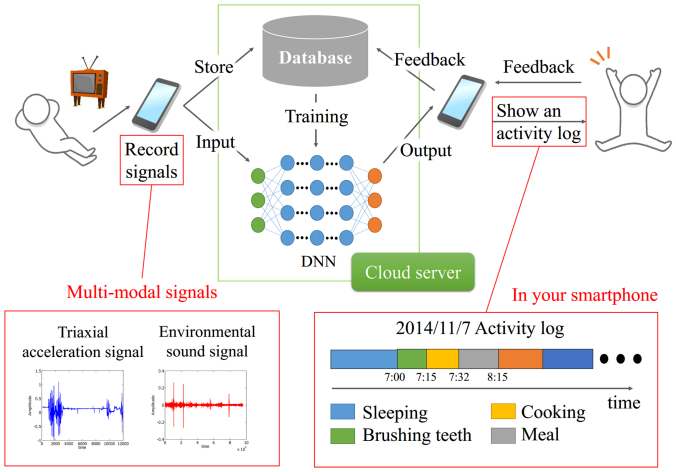

Fig. 1: Overview of target system

Fig. 2: Recording equipment worn by subjects

freely live in the room and go outside with recording staffs, however, to prevent an idle living such as sleeping all day, they are instructed to lead a well-regulated life. An accelerated signal is recorded with a smartphone put in a pocket of rear of subject's trousers, and an environmental sound signal and a video are recorded with a small video camera attached to subject's shoulder.

TABLE I: Data recording conditions

Number of subjects	1 (long-term) + 18 (short-term)
Recording environment	one-room studio apartment
Instructions	Lead well-regulated life
Recorded signals	Acceleration signal
	Environmental sound signal
	Video

B. Recorded data

About 1400 hours data including both indoor and outdoor activities were recorded. 300 hours data of indoor activities were annotated, and two types of dataset were constructed: 1) longterm, single subject data of 48 hours in length, 2) short-term, multiple subject data with a total length of 250 hours. The sampling rates of the recorded acceleration signals and

environmental sound signals were 200 Hz and 16kHz, respectively. The frame rate of the recorded video was 29.97 fps and resolution was 1280 × 720. The video and environmental sound signals were synchronized, but the acceleration signals were not synchronized because a different recording device was used. Therefore, these signals were synchronized using recording time information from the video and the time stamp information of the acceleration signal.

C. Annotation

Three people independently annotated the recorded signals using the recorded video and the ELAN annotation tool [33]. After that, another person checked the annotation. Activity tags used in the annotation and total duration lengths of individual tag are shown in Table 2. We used 21 tags to represent daily activities and an "Other" tag when a subject's activity could not be determined from the video. There were also situations when subjects conducted multiple activities simultaneously. In these situations we used two types of annotations: a primary tag to represent the main activity and a secondary tag to represent a sub-activity. Finally, to simplify the evaluation experiment, we divided the signals according to their tags, and then cut them into samples of one minute in length.

TABLE II: Recorded daily activities

Activity name	Length (min)	Activity name	Length (min)
Others	3879	Cleaning	188
Sleeping	2731	Writing	150
Note-PC	2252	Cleaning bath	107
Smartphone	1959	Calling	104
Watching TV	1873	Tablet	86
Cooking	1827	Light meal	85
Eating	908	Drying clothes	75
Cleaning table	679	Washing	36
Reading	476	Walking	30
Toilet	310	Monologue	5
Tooth brushing	310	Taking a bath	5

III. Framework of Daily Activity Recognition

In this study, we applied the following procedure of DAR:

1) Feature extraction
   a) Divide each signal into time windows of equal duration.
   b) Extract features from each window.
   c) Concatenate the features extracted from environmental sound and acceleration signal.
2) Train classifier by using the concatenated features.
3) Evaluate the classifier by using test data, which will be classified into one of the known classes.

A. Feature extraction

The environmental sound signal and the acceleration signal were into synchronous frames of equal duration, and extracted the features from each frame. Frame size and shift size were both 1 second. From the windowed environmental sound signals, we extract 41-dimension feature vectors: 13-order Mel-Frequency Cepstrum Coefficients (MFCC) with its 1st and 2nd order derivative coefficients, Zero Crossing Rate and Root Mean Square. MFCC is a feature which reflects human aural characteristics and is often used for speech recognition. RMS and ZCR represent volume and pitch of the sound signal, respectively. From the windowed acceleration signals, we extract 15-dimension feature vectors: the mean, variance, energy and entropy in frequency domain for each axis, and the correlation coefficients between these axis (3 axis × 4 + 3 correlation coefficients). The energy E and entropy S in frequency domain are calculated as follows:

$$E = \sum_{i=1}^{N-1} |F_i|^2, \quad S = -\sum_{i=1}^{N-1} p(i) \log p(i) \tag{1}$$

where F_i indicates the i-th FFT component of the signal of each axis and $p(i)$ is defined as the normalized $|F_i|^2$:

$$p(i) = \frac{|F_i|^2}{\sum_{i=1}^{N-1} |F_i|^2} \tag{2}$$

Correlation coefficient r between two axes is defined for the series data $\boldsymbol{x}$, $\boldsymbol{y}$ of two axis as follows:

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{C(\boldsymbol{x}, \boldsymbol{y})}{\sigma_{\boldsymbol{x}} \sigma_{\boldsymbol{y}}}, \tag{3}$$

In the above procedure, the extracted features are aligned in temporal order and stored by "sample" unit; the aligned features within 60 sec are packed in a file so that only a single activity event appears in each sample. These acceleration features were adopted in accordance with the previous study [19]. The final feature vectors were comprised of 56 dimensions by concatenating these feature vectors. The input of the classifier is also concatenated feature vectors, and the output is activity label(s) which is/are the one(s) of the target activities listed in table III.

IV. Recurrent Neural Network Applied to Daily Activity Recognition

Concatenated feature vectors from adjacent frames can be used as input to a FF-NN for DAR. However, the length of temporal context is limited. In order to consider more longer temporal context, recurrent neural network for DAR is applied in this study. Following sections briefly describe RNN and its variants. We will also describe related works applied to DAR and our motivation in this study.

A. Long Short-Term Memory

To cope with vanishing gradient / gradient explosion problem in RNN and bidirectional RNN, RNN with Long Short-Term Memory architecture (LSTM-RNN) has been proposed [20]. The LSTM-RNN has the special architecture, LSTM memory blocks. The hidden units in the vanilla RNN are replaced with it. The LSTM memory block contains

TABLE III: Target activitiy

Activities	Cleaning	Cooking	Meal	Note-PC	Reading	Sleeping	Smart-phone	Toilet	Watching-TV	Others
Num. of samples	679	1827	908	2252	476	2731	1959	310	1873	2319

memory cell which stores past information of the state, and gates which controls the duration of storing.

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \quad (4)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \quad (5)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \quad (6)$$

$$\tilde{\boldsymbol{c}}_t = \phi(\boldsymbol{W}_c[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c) \quad (7)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \quad (8)$$

$$\boldsymbol{h}_t = \phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t \quad (9)$$

where $\boldsymbol{c}_t$ is the state of memory cell. $\boldsymbol{i}_t$, $\boldsymbol{f}_t$, and $\boldsymbol{o}_t$ represent input gate, forget gate, and output gate for memory cell, respectively. The input gate controls the flow of input activations into the memory cell. The output gate controls the flow of the output activation from the cell into the rest of the network. The forget gate has the functionality to reset the content of memory cell.

LSTM-RNN can capture long term context by representing information from past inputs as hidden vector and propagating it to future direction. Furthermore, it is bidirectional LSTM-RNN which additional hidden vectors propagating information from future to past are used together. Bidirectional LSTM-RNN (BLSTM-RNN) has two types of the hidden layer: a forward hidden layer has connections to the direction from past to future and a backward hidden layer has connections in reverse order of forward hidden layer. However, because of introducing the memory blocks, the number of parameters of LSTM-RNN/BLSTM-RNN increases compared with the vanilla RNN.

B. Variants of LSTM-RNN

Many variants of LSTM-RNN have been already proposed, e.g., Gated Recurrent Unit (GRU) [21], and Minimal Gated Unit (MGU) [22]. They have simpler architecture and therefore smaller parameters than LSTM-RNN. Among these variants, LSTM with Recurrent Projection Layer (LSTMP-RNN) has been proposed to reduce the number of parameters in LSTM-RNN [18], where a linear transform (projection layer) is inserted after an LSTM layer. The outputs of projection layer then feedback to the LSTM layer. In the LSTMP-RNN, Eq. (9) can be replaced with following equation:

$$\boldsymbol{h}_t = \boldsymbol{W}_p(\phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t), \quad (10)$$

where $\boldsymbol{W}_p$ is the projection matrix with the size of $P \times N$, N is the dimension of the output vector of LSTM $\phi(\boldsymbol{c}_t) \odot \boldsymbol{o}_t$ and P is the dimension after linear transformation. If the projected dimension, P, is set to satisfy $P < N$, the number of parameters in LSTM block can be reduced significantly. It is expected that this parameter reduction leads to a similar effect to L1-regularization.

C. Related works

Previous studies [23], [24] applied one-dimentional convolutional neural network (CNN) for HAR. Especially, in [24], the combination of CNNs and LSTM-RNNs was proposed for HAR, where a stack of CNNs was utilized as feature extractor and a stack of LSTM-RNNs followed. They used the OPPORTUNITY dataset [25], where the activity data was collected in a daily living scenario by using many wearable sensors (inertial sensors and triaxial accelerometers). Those sensors were attached to many part of the human body, e.g., arm, elbow, show, back, hand, hip, knee. It is true that such many sensors can capture naturalistic activities. However the cost of those devices is not so cheap, that is, using many wearable sensors requires not a few wearing and equipment cost. We consider that users may feel obtrusiveness from those and therefore it is unsuitable for many users to use automatic daily surveillance system. Another previous study [26] compared the recognition performance of FF-NN, LSTM-RNN, and CNN for HAR. In accordance with the author's report, best recognition performance when using the OPPORTUNITY database was given by a bidiretional LSTM-RNN. Therefore the effectiveness of LSTM-RNN has been shown and we applied it for DAR task.

V. Subject Adaptation for RNN-based Daily Activity Recognition

To achieve better activity recognition performance, we need to build a customized classifier for each user. However, data collection cost for each user is often high and it is difficult to collect sufficient data for them. Furthermore, classifier is trained in a supervised manner. In other words, the annotated data is always required. Therefore, we consider utilize an adaptation technique used in the field of speech recognition, which enables to fit a trained model for a specific user using a small amount of the user-specific data.

Hayashi et al [17] evaluated the recognition performance in the subject adaptation framework using FF-NN. In fact, they applied three types of adaptation methods; (i) re-training of all of the layers, (ii) re-training of only a specific layer with a special type of regularization terms, and (iii) embedding the linear transformation network [27]. From the results, (i) and (iii) were almost the same performance and better than (ii). When 25 samples for each class were used for adaptation, the recognition rate of (8) was about 81%, where 1 sample corresponds 60 sec segment and the number of target activity was 9. However, we consider that the performance is still not satisfactory in practical use. The one of the possible reasons of such results is the excessive number of parameters even though the proper regularization technique is adopted. In other words, the FF-NN was still over-fitted to the adaptation data. Therefore, we consider to reduce the number of parameters

while maintaining the recognition performance. In order to reduce the number of parameters, LSTM with Recurrent Projection Layer (LSTMP-RNN) [18] is adopted in this paper. LSTMP-RNN has a linear transform (projection) layer and the projection layer is applied to the output of LSTM cell. By settings the size of the projection layers properly, LSTMP-RNN can reduce the number of parameters in LSTM-RNN. Therefore it is expected that the over-fitting problem can be mitigated.

In order to confirm the effectiveness of LSTMP-RNN, following initialization methods were applied in this study:

- Initialization by subject independent network
- Initialization by random values

We refer to the former as "SI-Init" and the later as "Random-Init" in the following. The procedure of adaptation with "SI-Init" is as follows:

1) Train subject-independent network.
2) Select samples for adaptation randomly from dataset for adaptation.
3) Train network by using the adaptation samples with the subject-independent network.
4) Evaluate the resultant network by using test data.
5) Repeat 3 and 4 with increasing the number of adaptation data.

For "Random-Init", the above procedures except '1' are applied. In "SI-Init", firstly a network is trained by data which includes many subjects. Adopting the parameters of the resultant network as initial values, then the network is re-trained by using the data for adaptation, which is not included in the training data. In "Ramdom-Init", the network parameters are initialized by random values, then the network is trained by using the adaptation data only.

VI. Experimental Evaluation

A. Experimental Conditions

For feature extraction, sound and acceleration signals were windowed with a 1.0 sec size and a 1.0 sec shift. The nineteen subjects are contained in the database. The recording duration of the one specific subject is 48 hours. The total duration of the other eighteen subjects is about 250 hours and the duration per one subject is within 24 continuous hours.

We conducted two experiments: **subject-closed** and **subject-adaptation**. In order to evaluate the effectiveness of LSTM-RNN for DAR, the former was conducted. The latter was for evaluation LSTMP-RNN in subject adaptation task. For subject-closed experiment, we used 48 hours recording data of for the above one subject and refer to this data as **dataset-closed**. For subject adaptation experiment, we used 250 hours recording data of the eighteen subject, in order to train subject-independent network for "SI-Init". We refer to this data as **dataset-adapt**. The feature vectors were extracted from both dataset by the same manner as described in Section III.

1) Subject-Closed Experiment: From our preliminary results, the number of hidden layers was set to 2, and the number of units per one hidden layer was set to 256. In LSTM-RNN, the dimension of memory cell in hidden layer equals to the number of units, 256. The corresponding dimension of memory cell in bidirectional LSTM-RNN was set to a half of the number of units, 128, because bidirectional LSTM-RNN has both forward and backward hidden layer. The length of unfold of both LSTM-RNNs was set to 60 frames. The activation function was Rectified Linear Unit (ReLU). The optimization algorithm was back-propagation through time via Adam [28] and the learning rate was fixed to 0.001. The network parameters were initialized by uniform distributed random numbers between $[-0.001, 0.001]$. The dropout rate was set to 0.5 and minibatch size was 128.

We adopted k-nearest neighbors (KNN), Gaussian Mixture Model (GMM), decision tree, Support Vector Machine (SVM), and FF-NN as comparative methods. The number of k neighbors was set to 5 and the number of mixture weights of GMM was set to 10. The kernel function of SVM was a RBF kernel. These hyper-parameters were determined through preliminary experiments, and all of these models were trained using the same feature vector, which consisted of both acceleration and acoustic features. The number of hidden layers of FF-NN was set to 3 and the number of units per one hidden layer was set to 2048, which were the same as the previous study [14]. The activation function for FF-NN was also ReLU.

We adopted a hold-out validation method for evaluation because the number of samples for each activity class is different. In this validation method, 10 test samples are randomly selected for each class and the rest is used for training, and this procedure was repeated by 10 times. For evaluation, the following averaged F-measure was adopted in this study:

$$F = \frac{1}{10}\sum_{r=1}^{10}\left(\frac{2}{C}\sum_{c=1}^{C}\frac{precision(c,r)\times recall(c,r)}{precision(c,r)+recall(c,r)}\right), \quad (11)$$

where C is the number of classes to be recognized and the summation index r suggests "repetition".

2) Subject Adaptation Experiment: On the basis of our preliminary results [29], the number of hidden layer L was set to 1 and the number of hidden units N was set to 512 for LSTM-RNN and LSTMP-RNN. Especially, the projected dimension was set to 32 for LSTMP-RNN. For FF-NN, L and N were set to the same values as those of **subject-closed** experiment. In this experiment, the same number of samples are selected from each class for adaptation and 10 samples are selected from each class for test. From each class, we select 1, 5, 10, 15, 20, or 25 adaptation samples. We also adopted the same hold-out validation method as subject-closed experiment and F-measure of Eq (11).

B. Results of Experiment

1) Results on Subject-Closed Experiment: Figure 3 shows the performance of daily activity recognition. The "Frame" represents a frame level accuracy and "Sample" represents a sample level accuracy, which is the recognition accuracy

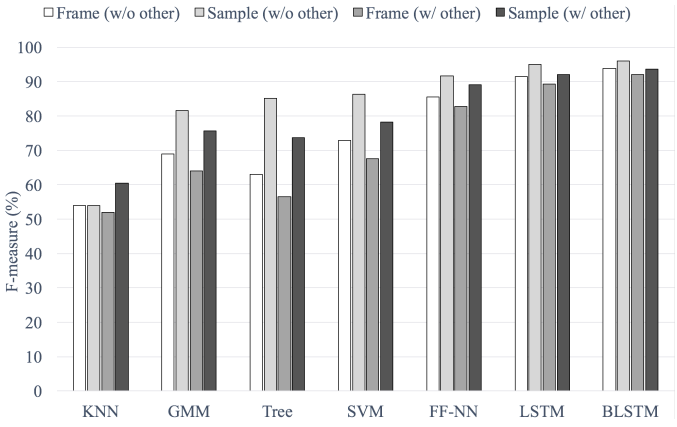

Fig. 3: Performance of Daily Activity Recognition

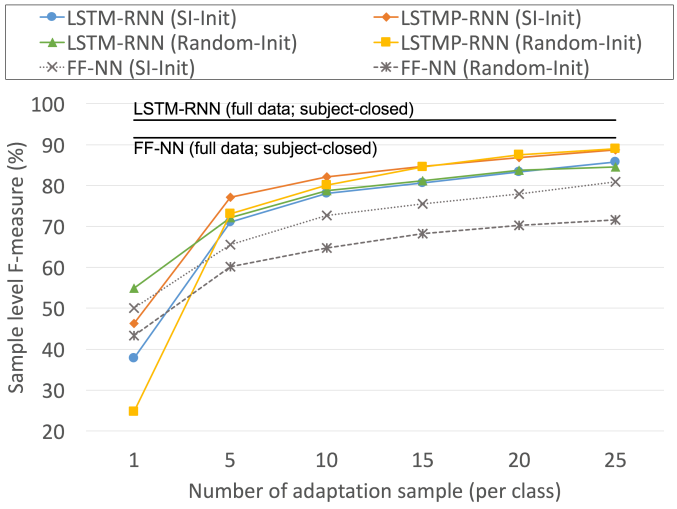

Fig. 4: Recognition performance of subject adaptation

obtained using the majority vote of the frame recognition results in each sample. The "w/o other" indicates the case that the "Others" class was not included as target class, and the "w/ other" indicates the case that the "others" class was included as target class. From this figure, "Sample" shows better results than "Frame" for each method. This can be explained via the effect of taking the majority from predicted results for each frame. Throughout the results, it can be seen that LSTM-RNN obtained better performance than FF-NN, and BLSTM-RNN surpassed LSTM-RNN.

2) Results on Subject Adaptation Experiment: Figure 4 shows recognition performance of subject adaptation by the sample-level F-measure. The horizontal solid lines, "LSTM-RNN (full data; subject-closed)" and "FF-NN (full data; subject-closed)" represent the F-measures evaluated on the dataset-closed for LSTM-RNN and FF-NN, respectively.

First, from figure 4, it can be seen that LSTM-RNN outperformed FF-NN even in adaptation task. Comparing "SI-Init" with "Random-Init", "SI-Init" showed better recognition performance than "Random-Init". It is suggested that "SI-Init" could help utilize subject independent information. Moreover,

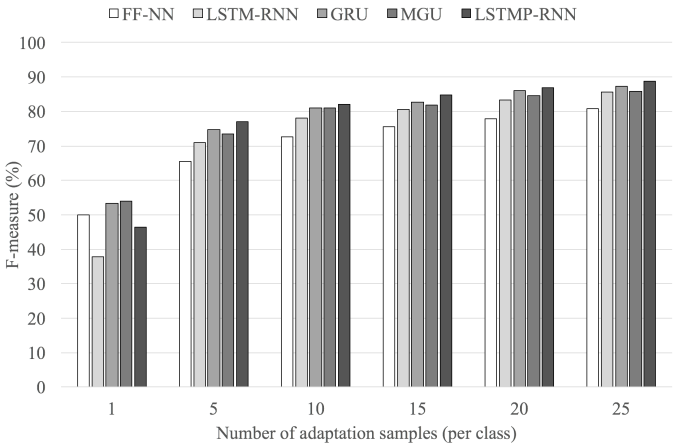

Fig. 5: Comparison of recognition performance with variants of LSTM-RNN in subject adaptation

TABLE IV: Number of parameters in each architecture on the subject adaptation experiment

Network	Number of Parameters
FF-NN	1.27×10^7
LSTM-RNN	1.17×10^6
GRU	0.88×10^6
MGU	0.59×10^6
LSTMP-RNN	1.81×10^5

it can be observed that results from LSTMP-RNN were better than LSTM-RNN. This means LSTMP-RNN could successively mitigate over-fit problem. In order to further evaluate the performance of LSTMP-RNN, we also applied variants of LSTM-RNN described in IV-B, i.e., GRU, and MGU.

Figure 5 shows frame-level recognition performance when applying each variants, where "LSTMP" represents LSTMP-RNN with "SI-Init". It can be observed that LSTMP-RNN is effective among them, especially when a few adaptation samples is available. Table IV lists the number of parameters in each architecture on the subject adaption experiment. From this table, the reduction of the parameters was effective, especially when a few adaptation samples per class are available.

VII. Conclusion

In this study, we investigated the effectiveness of LSTM-RNN for DAR. We also investigated LSTM-RNN with recurrent projection layer (LSTMP-RNN) for adaptation task. From the experimental results, it was shown that better recognition performance could be obtained compared to FF-NN, by using LSTM-RNN and bidirectional LSTM-RNN. Moreover, it was also shown that LSTMP-RNN was effective for adaptation task where limited amounts of data can be available, compared to other variants of LSTM-RNN architectures. The results showed that the reduction of network parameters by LSTMP-RNN could mitigate over-fit problem. In future, we will investigate suitable network architecture of RNN for adaptation task. We are also interested in developing 1D-CNN based architecture for DAR.

ACKNOWLEDGMENT

This research is partially supported by the Center of Innovation Program (Nagoya-COI) from Japan Science and Technology Agency (JST).

REFERENCES

[1] "Aging society white paper in fiscal 2015," Cabinet Office, Government of Japan, 2015.

[2] Kenichi Nakagawa, Taro Sugihara, Hitoshi Koshiba, Ryozo Takatsuka, Naotaka Kato, and Susumu Kunifuji, "Development of cooperative care support system for people with dementia by society oriented approach," *Trans.IPS.Japan*, vol. 49, no. 1, pp. 2–10, 2008.

[3] M. Philipose, K. P. Fishkin, M. Perkowitz, D. J. Patterson, D. Fox, H. Kautz, and D. Hahnel, "Inferring activities from interactions with objects," *IEEE Pervasive Computing*, vol. 3, no. 4, pp. 50–57, Oct 2004.

[4] Anthony Fleury, Michel Vacher, and Norbert Noury, "Svm-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 274–283, March 2010.

[5] Ya-Ti Peng, Ching-Yung Lin, Ming-Ting Sun, and Kun-Cheng Tsai, "Healthcare audio event classification using hidden markov models and hierarchical hidden markov models," in *2009 IEEE International Conference on Multimedia and Expo*, June 2009, pp. 1218–1221.

[6] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore, "Activity recognition using cell phone accelerometers," *SIGKDD Explor. Newsl.*, vol. 12, no. 2, pp. 74–82, Mar. 2011.

[7] Ling Bao and Stephen S. Intille, *Activity Recognition from User-Annotated Acceleration Data*, pp. 1–17, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[8] Tam Huynh and Bernt Schiele, "Towards less supervision in activity recognition from wearable sensors," in *2006 10th IEEE International Symposium on Wearable Computers*, Oct 2006, pp. 3–10.

[9] Quan Kong and Takuya Maekawa, "Reusing training data with generative/discriminative hybrid model for practical acceleration-based activity recognition," *Computing*, vol. 96, no. 9, pp. 875–895, 2014.

[10] Takuya Maekawa, Yutaka Yanagisawa, Yasue Kishino, Katsuhiko Ishiguro, Koji Kamei, Yasushi Sakurai, and Takeshi Okadome, "Object-based activity recognition with heterogeneous sensors on wrist," in *Pervasive Computing: 8th International Conference, Pervasive 2010, Helsinki, Finland, May 17-20, 2010. Proceedings*, Berlin, Heidelberg, 2010, pp. 246–264.

[11] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[12] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul J.M. Havinga, "A survey of online activity recognition using mobile phones," *Sensors*, vol. 15, no. 1, pp. 2059–2085, 2015.

[13] A. Wang, G. Chen, J. Yang, S. Zhao, and C. Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, June 2016.

[14] Tomoki Hayashi, Masafumi Nishida, Norihide Kitaoka, and Kazuya Takeda, "Daily activity recognition based on DNN using environmental sound and acceleration signals," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, Aug 2015, pp. 2306–2310.

[15] Masafumi Nishida, Norihide Kitaoka, and Kazuya Takeda, "Development and preliminary analysis of sensor signal database of continuous daily living activity over the long term," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, Dec 2014, pp. 1–6.

[16] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[17] Tomoki Hayashi, Norihide Kitaoka, Tomoki Toda, and Kazuya Takeda, "Adaptation Methods for Daily Activity Recognition Based on Deep Neural Network," *Technical Report of The Institute of Electronics, Information and Communication Engineers (in Japanese)*, vol. 116, no. 189, pp. 2306–2310, Aug 2016.

[18] Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Sep 2014, pp. 338–342.

[19] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L. Littman, "Activity recognition from accelerometer data," in *Proceedings of the 17th Conference on Innovative Applications of Artificial Intelligence - Volume 3*. 2005, IAAI'05, pp. 1541–1546, AAAI Press.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov 1997.

[21] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Ed., Doha, Qatar, 2016, pp. 1724–1734.

[22] Guo-Bing Zhou, Jianxin Wu, Chen-Lin Zhang, and Zhi-Hua Zhou, "Minimal gated unit for recurrent neural networks," *International Journal of Automation and Computing*, vol. 13, no. 3, pp. 226–234, 2016.

[23] Jian Bo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proceedings of the 24th International Conference on Artificial Intelligence*. 2015, IJCAI'15, pp. 3995–4001, AAAI Press.

[24] Francisco Javier Ordóñez and Daniel Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, 2016.

[25] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Gerhard Trster, Paul Lukowicz, Gerald Pirkl, David Bannach, Alois Ferscha, Jakob Doppler, Clemens Holzmann, Marc Kurz, Gerald Holl, Ricardo Chavarriaga, Hesam Sagha, Hamidreza Bayati, , and Jos del R. Milln, "Collecting complex activity data sets in highly rich networked sensor environments," in *Seventh International Conference on Networked Sensing Systems (INSS' 10)*, June 2010, pp. 5140–5144.

[26] Nils Y. Hammerla, Shane Halloran, and Thomas Ploetz, "Deep, Convolutional, and Recurrent Models for Human Activity Recognition using Wearables," in *Proceedings of 25-th International Joint Conference on Artificial Intelligence*, 2016, pp. 1533–1544.

[27] T. Ochiai, S. Matsuda, H. Watanabe, X. Lu, C. Hori, and S. Katagiri, "Speaker adaptive training for deep neural networks embedding linear transformation networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 4605–4609.

[28] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," Proceedings of the 3rd International Conference on Learning Representations (ICLR), 2014.

[29] A. Tamamori, T. Hayashi, T. Toda, and K. Takeda, "Daily Activity Recognition based on Recurrent Neural Network," *Technical Report of The Institute of Electronics, Information and Communication Engineers (in Japanese)*, vol. 116, no. 189, pp. 2306–2310, Aug 2016.