DAILY ACTIVITY RECOGNITION BASED ON DNN USING ENVIRONMENTAL SOUND AND ACCELERATION SIGNALS

Tomoki Hayashi[*], *Masafumi Nishida*[*], *Norihide Kitaoka*[†], *Kazuya Takeda*[*]

[*]Nagoya Univ., Japan [†]The Univ. of Tokushima, Japan

ABSTRACT

We propose a new method of recognizing daily human activities based on a Deep Neural Network (DNN), using multimodal signals such as environmental sound and subject acceleration. We conduct recognition experiments to compare the proposed method to other methods such as a Support Vector Machine (SVM), using real-world data recorded continuously over 72 hours. Our proposed method achieved a frame accuracy rate of 85.5% and a sample accuracy rate of 91.7% when identifying nine different types of daily activities. Furthermore, the proposed method outperformed the SVM-based method when an additional "Other" activity category was included. Therefore, we demonstrate that DNNs are a robust method of daily activity recognition.

Index Terms— Daily activity recognition, DNN, multimodal, acceleration signal, environmental sound signal

1. INTRODUCTION

An unprecedented aging of the population is occurring in Japan. In 2014, Japan was categorized as a "super-aging society" with more than 21 of the population falling into the category of elderly people over 65 years old. This trend is projected to accelerate, and by 2030 it is estimated that more than one third of the Japanese population will be over 65 years old. Demand for nursing and medical care will increase dramatically, and it will become difficult for society to meet these needs. Hence, it is necessary to develop technology which will allow elderly people to live independently and safely. In this study we propose a system which can monitor and assist elderly people in their daily lives by recognizing their current activities, using sensor signals obtained from a smartphone.

Many researchers have investigated daily activity recognition using sensor signals. Ohishi et al. [1] and Peng et al. [2] proposed an acoustic event detection method using environmental sound signals. Kwapisz et al. [3] conducted activity recognition experiments focusing on simple activities such as walking, running, standing, and so on. Ohuchi et al. [4] proposed a hierarchical daily activity recognition system which used a combination of subject acceleration and environmental sound signals. However, these studies used datasets consisting of simulated activity, which may differ from actual daily activities. Furthermore, the types of daily activities targeted were limited.

In this study we propose a robust daily activity recognition method based on a DNN using environmental sound and an acceleration signal. We conduct recognition experiments and compare the proposed method to other methods such as a Support Vector Machine (SVM), employing a dataset built by Nishida et al. [5]. The dataset consists of sensor signals recorded with a small video camera and a smartphone over 72 continuous hours. In addition to previously determined and programmed target activities, we also attempt to recognize activities which are not targeted, because our system should be able to recognize any kind of daily activity which occurs at any time. For this reason we also evaluate recognition accuracies when there is an additional category for "Other" activities.

2. PROPOSED METHOD

We developed the proposed model as follow:

1. Divide each signal into time windows of equal duration.
2. Extract features from each window.
3. Concatenate the features calculated from environmental sound and those extracted from an acceleration signal.
4. Train the classifier using the concatenated features.

In this study, we used DNN as the classifier, and compared its performance with a SVM-based method, a frequently used classification technique which can achieve good activity recognition performance.

2.1. Feature extraction

We first divide the environmental sound signal and the acceleration signal into synchronous windows of equal duration, and extract the features from each window. Window size and shift size were both 1 sec. Time stamp information from these signals was used for synchronization. We extracted three features from each environmental sound signal window; Mel Frequency Cepstral Coefficients (MFCC) + Power + Δ + $\Delta\Delta$, Zero-Crossing Rate (ZCR) and Root Mean Square (RMS). We obtained 41 dimensional features for each window, and used these features as outlined in [4].

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

We then extracted the following five features from each acceleration signal using the X, Y, and Z axes of each window: mean, variance, energy, entropy in the frequency domain, and correlation coefficients. These features were chosen per [6] and [7]. Here, the mean and variance are defined as the mean and variance of the raw acceleration signal. Energy E represents the sum of the absolute values of FFT components excluding the DC component, as expressed by the following equation:

$$E = \sum_{i=1}^{N-1} |F_i|^2, \tag{1}$$

where F_i indicates the i-th FFT component of the signal of each axis. Entropy in the frequency domain is represented as follows:

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

where $p(i)$ indicates the probability distribution derived from the normalized FFT component using the following equation:

$$p(i) = \frac{|F_i|^2}{\sum_{j=1}^{N-1} |F_j|^2}. \tag{3}$$

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

$$r(\mathbf{x}, \mathbf{y}) = \frac{Cov(\mathbf{x}, \mathbf{y})}{\sigma_{\mathbf{x}} \cdot \sigma_{\mathbf{y}}}, \tag{4}$$

where $Cov(\mathbf{x}, \mathbf{y})$ indicates covariance between two vectors and σ represents a standard deviation of vector components.

Finally, we concatenated these features extracted from the sound signal and acceleration signals and used a total 56 dimensional features as classifier inputs.

2.2. Activity Classifier

In this study, we used the DNN shown in Fig.1 as our classier. The number of layers and hidden nodes in each layer are 5 and 2,048, respectively. The number of nodes of the input layer corresponds to the dimensions of the input features, and the number of nodes of the output layer corresponds to the number of target activity classes.

We trained the DNN using the following procedure. First, we concatenated the features of 11 frames, which included the center frame, the 5 preceding frames, and the 5 succeeding frames by utilizing a key property of DNNs, which are the ability to deal with large numbers of dimensional feature vectors and time series data. In total, we used a $616(56 \times 11)$ dimensional feature vector as our DNN input. Second, we normalized the concatenated features as the mean and the variance of each dimension, so that they became 0 and 1, respectively, using all of the training data. Third, we pre-trained the DNN using a restricted Boltzmann machine (RBM) [8, 9]

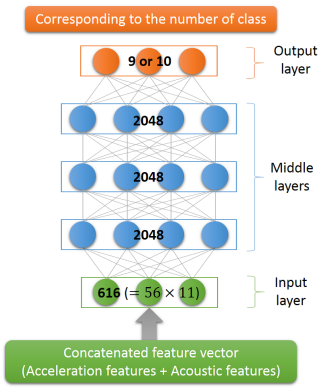

Fig. 1. Activity classifier using DNN

to set the appropriate initial parameters of the DNN using the normalized concatenated features. Finally, we trained the DNN by fine-tuning with back-propagation [10] using labeled data. In the fine-tuning phase we adopted two approaches. In the first approach, we trained the DNN at a fixed learning rate. After 5 epochs, we began to halve the learning rate every half epoch. In the second approach, we used the Dropout [11] method with a fixed learning rate.

3. EXPERIMENTS

We conducted a daily activity recognition experiment to confirm the performance of the proposed model, using the dataset built by Nishida et al. [5], which recorded real-world human activity continuously for 72 hours. This dataset includes environmental sound signals recorded with a Go-pro video camera attached to the subject's shoulder and an acceleration signal recorded with a smart-phone in the back pocket of the subject's trousers. The Subject is a 20 year-old, male undergraduate student, living in a one-room studio apartment. [i.e., NOT a one bedroom apartment (an apartment with a kitchen/dining area and a separate bedroom).]

3.1. Experimental conditions

A list of target activities is shown in Table 1, with the numbers in parentheses representing the number of samples of each

Table 1. Target activity

A	Cleaning [39]	F	Sleeping [1257]
B	Cooking [108]	G	Smart-phone [198]
C	Meal [120]	H	Toilet [61]
D	Note-PC [141]	I	Watching-TV [109]
E	Reading [164]	J	Other [582]

Table 2. Experimental conditions

Environmental Sound signal sampling rate	16000 Hz
Acceleration signal sampling rate	128 Hz
Window size	1.0 sec
Shift size	1.0 sec
#Sample	2779
Validation method	Hold-out validation
#K of KNN	5
#Mixture of GMM	10
SVM	libsvm-3.18
SVM Kernel	RBF kernel
SVM Type	One-Versus-One

Table 3. DNN Fine-tuning conditions

Number of layer	5
Middle later nodes	2048
Learning rate	0.006 (w/o dropout) 0.06 (w/ dropout)
momentum	0.0
L2	0.0 (w/o dropout) 0.00001 (w/ dropout)
epoch	20 (w/o dropout) 400 (w/ dropout)
Droprate of input layer	0.2
Droprate of middle layers	0.5

activity. Experimental conditions and DNN training conditions are shown in Tables 2 and Table 3 respectively. In Table 2, the "sample" means a data segment whose length is 60 sec. In these experiments we adopted the Hold-out validation method, since the number of samples for each activity class is different. For Hold-out validation we chose 10 samples randomly from each class, and used these samples as test data. The remaining data was used as training data.

3.2. Investigation of effectiveness of acceleration features

First, we built a DNN and other models (KNN, GMM, Decision tree, SVM) to classify nine activity classes, without an "Other" class to represent non-target activities. To confirm the effectiveness of the acceleration features, we compared performance using only acoustic features with performance using both acoustic and acceleration features. Results are shown in Fig. 2. "Frame" indicates a frame accuracy, which is the recognition accuracy for the frame unit. "Sample" indicates a sample accuracy, which is the recognition accuracy obtained using the majority of the frame recognition results in each sample. The results for all of the models show that recognition accuracy, especially frame accuracy, improved when we added acceleration features to the input features. Therefore, we can confirm the effectiveness of using acceleration features for activity classification.

Next, we focused on the difference in effectiveness between DNNs and other methods. Our results, shown in Fig. 2, revealed that the DNN outperformed other methods, especially when using a DNN with a dropout. Comparing performance with and without a dropout, the DNN with a dropout achieved higher frame and sample accuracies. The reason for this dramatic improvement may be related to the variety of signals in the same class. Since we used data recorded in a real environment and there were many signals in the same class, we assume the effect of over-fitting became more apparent, hence, generalization methods such as dropout became more effective. Whatever the case may be, we confirmed that the DNN achieved better performance.

3.3. Recognition of non-target activity

Next, we added an "Other" category for non-targeted activity, and built a DNN and other models for a ten activity classification problem. Experimental results are shown in Fig. 3. Our results show that the performance of other methods decreased dramatically in both frame and sample accuracy when a non-target activity category was added. The confusion matrix of SVM in Table 5 shows that adding an "Other" class influences the other classes significantly. Classification of the "Cleaning" activity class was most drastically affected, with more than 3/4 of the samples misclassified into the "Other" class. This may be caused by the encroachment of the "Other" class into the rest of the classes affected the decisions made by the discriminative hyper-plane in the feature domain. To confirm this, we plotted input features in a 2-D space through dimensional reduction using PCA, as shown in Fig. 4. We can see that the "Other" class encroached into the rest of the classes.

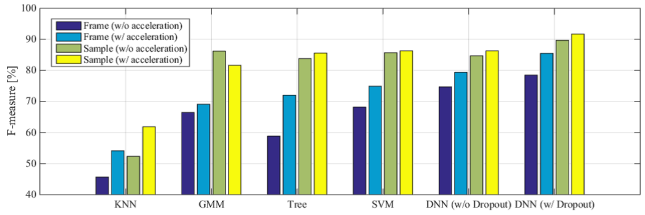

Fig. 2. Classification performance with/without acc. features

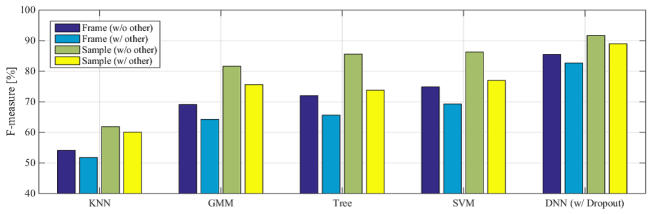

Fig. 3. Classification performance with/without "Other"

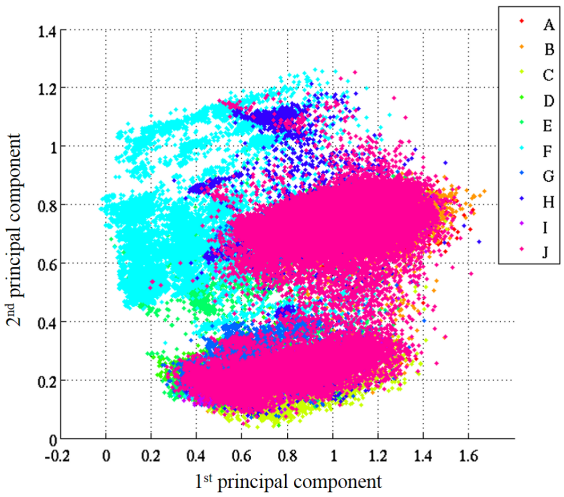

Fig. 4. Input feature of each classes in the feature domain

On the other hand, the classification performance of the DNN did not decrease as sharply as it did with other methods after an "Other" class was added. Comparing the confusion matrices in Tables 5 and 6 confirms that the effect of the "Other" class was reduced when using the DNN, especially regarding the misclassification of the "Cleaning" activity samples. These results show that the DNN did not construct a simple hyper-plane for discrimination, which indicates that DNNs are likely more robust to outliers.

3.4. N-best result

In actual application, improvement of the recognition rate occurs during re-scoring, using an N-best list. Hence, it is crucial important whether or not the correct label appears in the N-best list. The N-best results for the DNN with dropout are shown in Table 4. Even with the existence of an "Other" class, our results confirm that the DNN with dropout achieved high performance, with a recognition rate accuracy of over 90% when $N = 2$.

Table 4. N-best results of DNN with dropout

	w/o other		w/ other	
	Frame	Sample	Frame	Sample
1-Best [%]	85.5	91.7	82.7	89.0
2-Best [%]	92.4	96.3	92.4	96.3
3-Best [%]	95.6	97.9	96.4	97.8

4. CONCLUSION

We proposed a new method for the recognition of daily human activities using a DNN with environmental sound and subject acceleration signals. We conducted recognition experiments and compared our method with other methods such as an SVM using a real-world dataset recorded over 72 continuous hours. Our results showed that acceleration features are effective for recognizing daily activities. The proposed method demonstrated its effectiveness by achieving a frame accuracy rate of 85.5% and a sample accuracy rate of 91.7% when categorizing nine different types of daily activities. Furthermore, our proposed method outperformed an SVM-based classification method when using nine activity categories and a tenth "Other" (out-of-target) category, achieving a frame accuracy of 82.7% and a sample accuracy of 89.0%. Regarding future work, we will investigate differences in activity classification rates when using multiple subjects, as well as the effect on performance when additional activities are targeted.

Table 5. Confusion matrix of activity classification by SVM

Answer label \ Predicted label	A	B	C	D	E	F	G	H	I	J
A	26	0	0	0	0	0	0	0	0	74
B	0	64	0	0	0	0	0	0	0	36
C	0	1	83	0	0	0	1	0	0	15
D	0	0	0	93	0	0	3	0	0	4
E	0	0	0	0	79	10	2	0	1	8
F	0	0	0	0	0	100	0	0	0	0
G	0	0	0	2	4	0	75	0	4	15
H	0	0	0	0	0	2	3	62	1	32
I	0	0	1	0	8	0	5	0	67	19
J	0	0	0	0	0	0	2	0	0	98

Table 6. Confusion matrix of activity classification by DNN

Answer label \ Predicted label	A	B	C	D	E	F	G	H	I	J
A	78	3	0	0	0	0	0	0	0	19
B	1	88	0	3	0	0	0	0	0	8
C	0	0	93	0	0	0	1	0	0	6
D	0	0	0	96	0	0	2	0	2	0
E	0	0	0	0	88	6	3	0	2	1
F	0	0	0	0	0	100	0	0	0	0
G	0	0	1	2	6	0	87	0	0	4
H	0	0	0	0	1	1	1	82	0	15
I	0	0	2	0	5	0	1	0	78	14
J	0	0	0	0	0	0	0	0	0	100

REFERENCES

[1] Y. Ohishi, D. Mochihashi, T. Matsui, M. Nakano, H. Kameoka, T. Izumitani, and K. Kashino, "Bayesian semi-supervised audio event transcription based on Markov Indian buffet process," in *Proc. ICASSP*, 2013.

[2] 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 *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, June 2009, pp. 1218–1221.

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

[4] Kazushige Ouchi and Miwako Doi, "Living activity recognition using off-the-shelf sensors on mobile phones," *Annals of Telecommunications*, vol. 67, no. 7-8, pp. 387–395, 2012.

[5] M. Nishida, N. Kitaoka, and K. Takeda, "Development and preliminary analysis of sensor signal database of continuous daily living activity over the long term," in *Proc. APSIPA*, 2014.

[6] Ling Bao and StephenS. Intille, "Activity recognition from user-annotated acceleration data," in *Pervasive Computing*, Alois Ferscha and Friedemann Mattern, Eds., vol. 3001 of *Lecture Notes in Computer Science*, pp. 1–17. Springer Berlin Heidelberg, 2004.

[7] Nishkam Ravi, D. Nikhil, Preetham Mysore, and Michael L. Littman, "Activity recognition from accelerometer data," in *In Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence(IAAI)*. 2005, pp. 1541–1546, AAAI Press.

[8] GeoffreyE. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural Networks: Tricks of the Trade*, Gregoire Montavon, GenevieveB. Orr, and Klaus-Robert Muller, Eds., vol. 7700 of *Lecture Notes in Computer Science*, pp. 599–619. Springer Berlin Heidelberg, 2012.

[9] Vinod Nair and Geoffrey E. Hinton, "Implicit mixtures of restricted Boltzmann machines," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., pp. 1145–1152. Curran Associates, Inc., 2009.

[10] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, "Learning representations by back-propagating errors," in *Neurocomputing: Foundations of Research*, James A. Anderson and Edward Rosenfeld, Eds., pp. 696–699. MIT Press, Cambridge, MA, USA, 1988.

[11] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *CoRR*, vol. abs/1207.0580, 2012.