PAPER

Daily Activity Recognition with Large-Scaled Real-Life Recording Datasets Based on Deep Neural Network Using Multi-Modal Signals

Tomoki HAYASHI[†a)], ***Nonmember***, **Masafumi NISHIDA**[††b)], **Norihide KITAOKA**[†††c)],
Tomoki TODA[††††d)], ***Members***, ***and*** **Kazuya TAKEDA**[†e)], ***Fellow***

SUMMARY In this study, toward the development of smartphone-based monitoring system for life logging, we collect over 1,400 hours of data by recording including both the outdoor and indoor daily activities of 19 subjects, under practical conditions with a smartphone and a small camera. We then construct a huge human activity database which consists of an environmental sound signal, triaxial acceleration signals and manually annotated activity tags. Using our constructed database, we evaluate the activity recognition performance of deep neural networks (DNNs), which have achieved great performance in various fields, and apply DNN-based adaptation techniques to improve the performance with only a small amount of subject-specific training data. We experimentally demonstrate that; 1) the use of multi-modal signal, including environmental sound and triaxial acceleration signals with a DNN is effective for the improvement of activity recognition performance, 2) the DNN can discriminate specified activities from a mixture of ambiguous activities, and 3) DNN-based adaptation methods are effective even if only a small amount of subject-specific training data is available.

key words: *human activity recognition, activity of daily living, database, deep neural networks, adaptation*

1. Introduction

The goal of human activity recognition (HAR) system is to identify human activities from observed signals. These systems have great potential in various applications such as life logging [1], monitoring the elderly [2], detection of wandering behavior in dementia patients [3], health care [4], [5], control systems in automated “smart homes” (light switches, climate control, etc.) [6], [7], and so on. The mass marketing of electronic devices with sensing capabilities has made it possible to easily acquire various types of signals which can be used to identify the human activity, and as a result, the field of HAR has been attracting more attention.

HAR can be divided into two main categories: environmental augmentation approach and wearable sensing approach. The first approach, environmental augmentation, utilizes information collected with sensors embedded in an environment to recognize subjects’ activities. In the field of computer vision, cameras have been utilized to detect subjects’ physical activities [8], [9], or to understand group activities [10], [11]. On the other hand, in the field of environmental sound understanding, microphones have been utilized to identify sound events such as phone ringing, typing on a keyboard, human speech and so on [12]–[14]. These approaches allow recognition of various types of activities, however, there is a limitation of installation location. Furthermore, the use of cameras may subjects uncomfortable due to lack of privacy. Another approach of environmental augmentation is based on the use of ubiquitous sensors such as radio frequency identifier (RFID) tags and switch sensors [15]–[18]. In these approaches, with the embedding small sensors to all of the objects in a room, the system can not only detect the use of objects such as a knife, spoon, and cups but also recognize complicated human activities such as making coffee, taking a medicine and washing dishes. However, they require the embedding of many sensors, making it very costly. The second approach, wearable sensing, utilizes information collected with wearable sensors attached to a subject’s body to recognize activities, especially which have characteristic motion or sound. Compared to environmental augmentation approach, wearable sensor approaches generally involve much lower costs because it does not require the embedding of many sensors. One of the most typical approaches is based on the acceleration signals recorded with wearable devices to recognize the activities such as walking, running, cycling, going up (or down) the stairs, and on [19]–[22]. However, it is difficult to recognize the complicated activities including the use of objects. To address this issue, some studies have combined various type of sensors such as an acceleration sensor, gyro sensor, microphone, geomagnetic sensor [23]–[25].

Recently, with the improvements in deep learning and the advent of public benchmarking datasets [26], [27], neural network based approaches to HAR in wearable sensing have been proposed. In the study [28], feed-forward neural networks with human-designed features have been utilized, outperforming conventional classification methods such as support vector machine (SVM). Other studies have utilized deep convolutional networks to extract features from observed signals, and long short-term memory (LSTM) recur-

Manuscript received May 4, 2017.
Manuscript revised August 31, 2017.

[†]The authors are with Graduate School of Information Science, Nagoya University, Nagoya-shi, 464-8603 Japan.
[††]The author is with Graduate School of Information, Shizuoka University, Hamamatsu-shi, 432-8011 Japan.
[†††]The author is with Graduate School of Science and Technology, Tokushima University, Tokushima-shi, 770-8506 Japan.
[††††]The author is with Information Technology Center, Nagoya University, Nagoya-shi, 464-8603 Japan.
a) E-mail: hayashi.tomoki@g.sp.m.is.nagoya-u.ac.jp
b) E-mail: nishida@inf.shizuoka.ac.jp
c) E-mail: kitaoka@is.tokushima-u.ac.jp
d) E-mail: tomoki@icts.nagoya-u.ac.jp
e) E-mail: takeda@is.nagoya-u.ac.jp

DOI: 10.1587/transfun.E101.A.199

rent neural networks to capture the temporal dependencies between activities, achieving great performance without the use of human-designed features [29], [30]. However, these approaches were evaluated using a huge database which was recorded in a sensor-rich environment, with subjects who attached many sensors to their bodies. The performance under the practical situation, where only limited sensors or a limited amount of data are available, has not been evaluated. Moreover, a realistic HAR system must be able to handle unknown subjects with only a small amount of subject-specific data. Therefore, it is also necessary to evaluate the performance under these conditions.

In this study, toward the development of monitoring system for life logging (Fig. 1), we construct a human activity database and evaluate it using neural network based methods[†]. We collect over 1,400 hours of data including both the outdoor and indoor daily activities of 19 subjects under practical conditions with a smartphone and a small video camera, and construct huge human activity database, which consists of an environmental sound signal, triaxial acceleration signals and manually annotated activity tags. Using our constructed database, we evaluate the human activity recognition performance of deep neural networks (DNNs), and apply DNN-based adaptation techniques to improve the performance with only a small amount of subject-specific training data. We experimentally demonstrate the following:

1. the use of multi-modal signal, including environmental sound and triaxial acceleration signals with a DNN is effective for the improvement of activity recognition performance,
2. the DNN can discriminate specified activities from a mixture of ambiguous activities,
3. DNN-based adaptation methods are effective even if only a small amount of subject-specific training data is available.

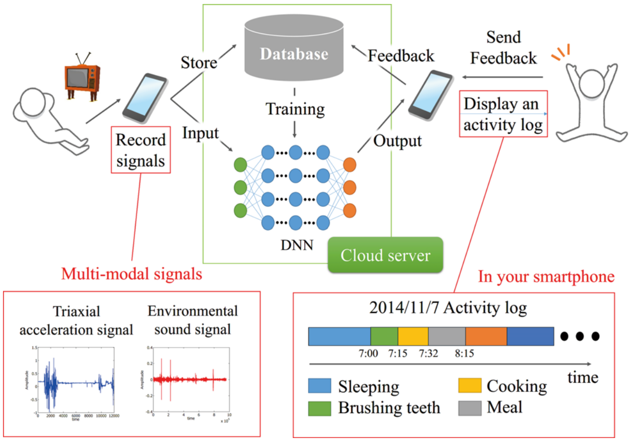


Fig. 1 Overview of our target life-logging system. The system uses a smartphone to continuously record environmental sound and acceleration signals, and sends these signals to a server. In the server, subject's current activity is automatically recognized by an activity recognition model, and then recognition results are then sent to the subject's smartphone. The subjects can not only view their activity history but also send a feedback to improve the recognition performance.

2. Nagoya-COI daily Activity Database

In this section we describe the construction of Nagoya Center Of Innovation (Nagoya-COI) daily activity database.

2.1 Recording Condition

The outline of recording condition of daily activity is shown in Table 1, and the equipment of subject is shown in Fig. 2. 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. The recording environment is a one-room studio apartment. Note that it is an apartment with a kitchen/dining area and a separate bedroom. Subjects can freely live in the room, however, they were asked to lead a well-regularized life so that we could obtain a variety of activity data from each subject. In other words, subjects were encouraged to avoid sleeping all day, watching excessive amounts of television, etc. And in order to prevent recording errors, subjects were asked not to go outside alone, but to let an assistant accompany them to help record their outdoor activities.

2.2 Recorded Data

We recorded 1,400 hours data including both indoor and outdoor activities. We then annotated 300 hours data of indoor

Table 1 Data recording conditions.

Number of subject	1 (long-term) + 18 (short-term)
Recoding environment	one-room studio apartment
Instruction	Lead well-regulated life
Recorded signals	Triaxial acceleration signals (200 Hz)
	Environmental sound signal (16k Hz)
	Video (1280×720, 29.97 fps)

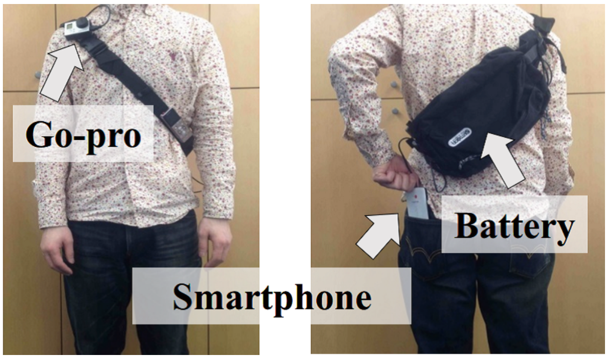


Fig. 2 Recording equipment worn by subjects. Note that the video camera is only for data annotation purposes and is not part of the target system.

[†] In comparison to our previous work [31], we here further investigate the performance of our method under various conditions, and apply adaptation techniques to handle unknown subjects with only a small amount of data.

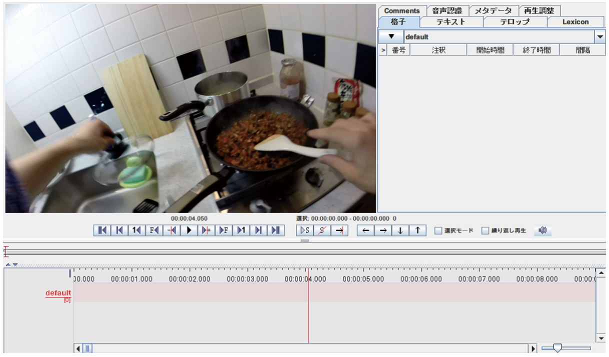


Fig. 3 Annotation with ELAN.

Table 2 Recorded daily activities.

Activity name	Length [min]	Activity name	Length [min]
Others	3,879	Cleaning	188
Sleeping	2,731	Writing	150
Note-PC	2,252	Cleaning bath	107
Smartphone	1,959	Calling	104
Watching TV	1,873	Tablet	86
Cooking	1,827	Light meal	85
Eating	908	Drying clothes	75
Clearing table	679	Washing	36
Reading	476	Waking	30
Toilet	310	Monologue	5
Tooth brushing	214	Taking a bath	958

activities, and constructed two types of dataset: 1) long-term, 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 16,000 Hz, respectively. The frame rate of the recorded video was 29.97 fps and resolution was $1,280 \times 720$. The video and environmental sound signals were synchronized, but the acceleration signals were not synchronized because a different recording device was used. Therefore, we synchronized these signals using recording time information from the video and the time stamp information of the acceleration signal. Note that the time stamp information was recorded every sampling, and therefore, it has enough time resolution to synchronize.

2.3 Annotation

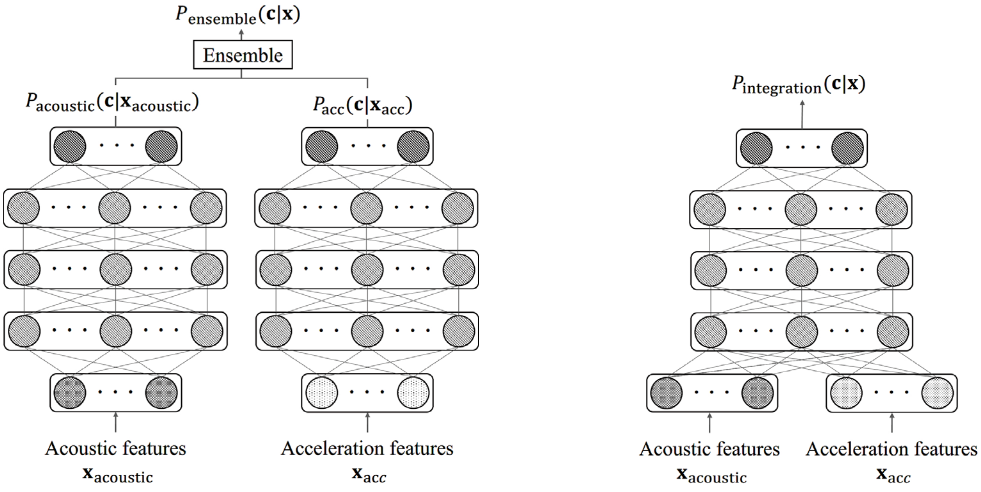
Three people independently annotated the recorded signals using the recorded video and the ELAN annotation tool [32]. After that, another person checked the annotation. Activity tags used in the annotation and total duration lengths of the individual tag are shown in Table 2. Total 21 tags are used to represent daily activities, and an "Other" tag is used to represent when a subject's activity could not be determined from the video. We tagged all of the activities of our subjects which could be determined from the recorded video. When recording the activity of "Sleeping", the subject wore only the smartphone and the camera placed on the bedside desk. Similarly, when recording the activity of "Taking a bath", the subject removed all equipment but the equipment was placed in the bathroom and continued recording. There were also situations when subjects conducted multiple activities simultaneously (e.g., eating lunch while watching TV). 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. In this study, we assumed that the activity started first was the primary, and that simultaneous activities initiated later were secondary. 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.

3. Daily Activity Recognition Model

In this section, we describe our Deep Neural Network (DNN)-based daily activity recognition model and its adaptation methods.

3.1 Pre-Processing and Feature Extraction

The acceleration signals recorded using a smartphone included pulsive noise signal which were not related to actual movement, and sometimes the signals lacked consecutive samples, which were likely caused by inadequate smartphone processor performance. These factors had a negative influence on the analysis of our data, therefore, we applied a median filter to remove pulsive noise signal and conducted spline interpolation to project signals which were missing during sampling as pre-processing procedures. After pre-processing, we divided the environmental sound signal and the acceleration signal into synchronous frames of equal duration, and extracted the features from each frame. Frame size and shift size were both 1 second. We extracted three features from each environmental sound signal frame: 1) Mel Frequency Cepstral Coefficients (MFCC) + Power + Δ + $\Delta\Delta$, 2) Root Mean Square (RMS) and 3) Zero-Crossing Rate (ZCR). We obtained 41-dimensional acoustic features for each frame. MFCC is a feature which reflects human aural characteristics and is often used for speech recognition, and its effectiveness has also been confirmed in acoustic event detection [33]. RMS and ZCR represent volume and pitch, respectively. We then extracted the following five features from each acceleration signal using the X, Y, and Z axes of each frame: mean, variance, energy, entropy in the frequency domain, and correlation coefficients, where mean and variance are defined as the mean and variance of the raw acceleration signal. The mean represents the orientation of the smartphone, and is closely related to the user's posture. For example, suppose that the smartphone is put in a rear pocket of the user's trousers. When the user is standing, the Y axis acceleration component includes acceleration forces due to the earth's gravity. However, when the user is sitting, the Y axis acceleration component omits the effect of gravity. The variance represents the intensity of a user movement, and is effective for detecting user movement. Energy E represents the sum of the absolute values of the fast Fourier transform (FFT) components excluding the DC component,


$P_{\text{ensemble}}(\mathbf{c}|\mathbf{x})$
Ensemble
$P_{\text{acoustic}}(\mathbf{c}|\mathbf{x}_{\text{acoustic}})$
$P_{\text{acc}}(\mathbf{c}|\mathbf{x}_{\text{acc}})$
Acoustic features
$\mathbf{x}_{\text{acoustic}}$
Acceleration features
$\mathbf{x}_{acc}$
$P_{\text{integration}}(\mathbf{c}|\mathbf{x})$
Acoustic features
$\mathbf{x}_{\text{acoustic}}$
Acceleration features
$\mathbf{x}_{\text{acc}}$

(a) *Ensemble model.* (b) *Integrated model.*

Fig. 4 Proposed DNN activity classifier.

as expressed by the following equation:

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

where F_i represents the i-th FFT component of the signal of each axis. This feature is also effective for detecting user movement. Entropy in the frequency domain is represented as follows:

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

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

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

This entropy value enables us to discriminate between different activities which have the same intensity. Correlation coefficient r between two axis is defined for the series data $\mathbf{s}_1, \mathbf{s}_2$ of two axis as follows:

$$r(\mathbf{s}_1, \mathbf{s}_2) = \frac{\text{Cov}(\mathbf{s}_1, \mathbf{s}_2)}{\sigma_{\mathbf{s}_1} \cdot \sigma_{\mathbf{s}_2}}, \tag{4}$$

where $\text{Cov}(\mathbf{s}_1, \mathbf{s}_2)$ represents covariance between two vectors and σ represents a standard deviation of vector components. The correlation coefficients represent the direction of movement of the smartphone, which is related to user movement.

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

3.2 Activity Classifier

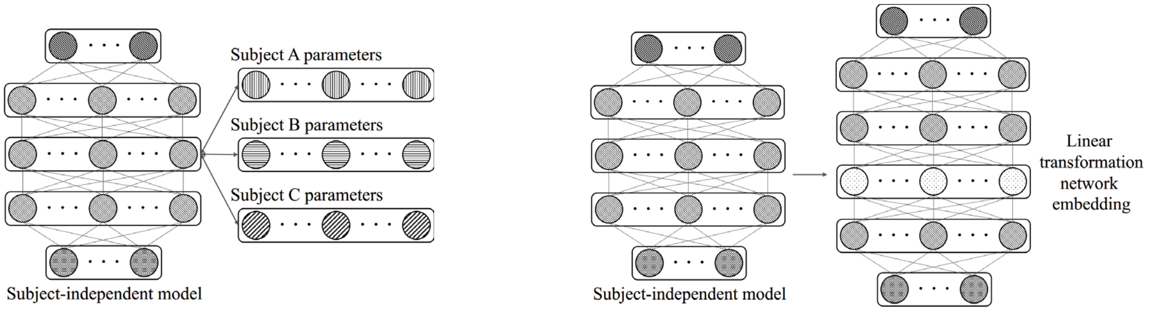
In this study, we use DNNs as an activity classier. DNNs can not only deal with high dimensional feature vectors reflecting a time sequence, but can also be trained to automatically convert themselves into discriminative feature vectors through lamination of their hidden layers.

In this study, we evaluate two types of DNNs: 1) an ensemble model which integrates the outputs of the acoustic and acceleration feature models, and 2) an integrated model which utilizes acoustic and acceleration features as an input feature vector. The structures of these DNNs are shown in Fig. 4. The outputs of ensemble model $P_{\text{ensemble}}(c|\mathbf{x})$ are calculated as follows:

$$P_{\text{ensemble}}(c \mid \mathbf{x}) = P_1(c \mid \mathbf{x}_{\text{acoustic}})^w P_2(c \mid \mathbf{x}_{\text{acc}})^{1-w}, \tag{5}$$

where $P_1(c \mid \mathbf{x}_{\text{acoustic}})$ and $P_2(c \mid \mathbf{x}_{\text{acc}})$ are outputs of the acoustic feature model and the acceleration feature model, respectively, c is an index of activity class, w is a weight coefficient between the acoustic feature model and the acceleration feature model, and $\mathbf{x}_{\text{acoustic}}$, $\mathbf{x}_{\text{acc}}$, and $\mathbf{x}$ are the acoustic feature vector, the acceleration feature vector, and the concatenated feature vector, respectively. The networks consist of 3 hidden layers with 2,048 hidden nodes, and a sigmoid function is used as an activation function. The number of nodes in the input layer corresponds to the dimensions of the input feature vector, and the number of nodes of the output layer corresponds to the number of target activity classes.

The training procedure is as follows. First, we concatenate the features of 11 frames, which included a center frame, the 5 preceding frames and the 5 succeeding frames, utilizing a key property of DNNs which is the ability to deal with large numbers of dimensional feature vectors. Second, we normalize the concatenated features using all of the training data, making the mean and the variance of each dimension 0 and 1, respectively. Third, we pre-train the DNN using greedy learning with a denoising auto encoder (DAE), in order to appropriately set the initial parameters of the DNN using the normalized, concatenated features. When training


Subject A parameters
Subject B parameters
Subject C parameters
Subject-independent model
Subject-independent model
Linear transformation network embedding

(a) *Re-training only a specific layer.*

(b) *Linear transformation network embedding.*

Fig. 5 Outline of adaptation methods.

the DAE, we add Gaussian noise with a variance of 0.1 to the input vectors. Finally, we train the DNN by fine-tuning with back-propagation using annotation data. During the fine-tuning phase, we use the Adam optimization method [34] and dropout [35] with a fixed learning rate of $5e-4$.

3.3 Adaptation Methods

To achieve better activity recognition performance, we need to build a customized model for each user. However, it is difficult to collect sufficient data for each user, and building the model requires annotation of all the data. Therefore, we utilize an adaptation technique used in the field of speech recognition, which enables to fit a trained model for a specific user even if only a small amount of subject-specific training data is available.

We use three types of adaptation methods whose effectiveness has been confirmed in the field of speech recognition [36], [37]. The first adaptation method is to train all of the layers of the DNN. When using this approach, we use the parameters of the subject independent model as the initial parameters, and then re-train the network using subject-specific data for adaptation. If the amount of subject-specific data used for adaptation is small, the network will tend to overfit, therefore we need to determine a suitable regularization coefficient.

The second adaptation method is to re-train only a specific layer with subject-specific data, which is selected as a subject adaptation layer [36]. A diagram of this method is shown in Fig. 5(a). The adaptation is performed as follows:

$$\underset{(\hat{\mathbf{W}}^{(l)},\hat{\mathbf{b}}^{(l)})}{\text{argmin}} \; E(\mathbf{\Lambda},\hat{\mathbf{W}}^{(l)},\hat{\mathbf{b}}^{(l)})+\frac{\beta}{2}\left(||\hat{\mathbf{W}}^{(l)}-\mathbf{W}^{(l)}||^2+||\hat{\mathbf{b}}^{(l)}-\mathbf{b}^{(l)}||^2\right), \qquad (6)$$

where l is the index of the adaptation layer, $\mathbf{W}$ and $\mathbf{b}$ represent the weight and bias parameters before re-training, respectively, $\hat{\mathbf{W}}$ and $\hat{\mathbf{b}}$ represent the weight and bias parameters after re-training, respectively, $\mathbf{\Lambda}$ represents all of the network parameters, and β is a regularization coefficient. The first term represents the error function of the network, and the second term represents a regularization term which prevents leaving too much in common with the original parameters.

The third adaptation method is embedding the linear transformation network (LTN embedding) [37]. A diagram of this method is shown in Fig. 5(b). A linear transformation layer is inserted before a specific layer and only the linear transformation layer is re-trained. When inserting the linear transformation layer, its weight parameters $\mathbf{A}$ and its bias parameters $\mathbf{a}$ are initialized as an identity matrix and a zero vector, respectively. The optimization is conducted based on the following equation:

$$\underset{(\mathbf{A},\mathbf{a})}{\text{argmin}} \; E(\mathbf{\Lambda},\mathbf{A},\mathbf{a})+\frac{\beta}{2}\left(||\mathbf{A}-\mathbf{I}||^2+||\mathbf{a}-\mathbf{0}||^2\right), \qquad (7)$$

where $\mathbf{\Lambda}$ represents all of the network parameters, and β is a regularization coefficient. The first term represents the error function of the network, and the second term represents a regularization term which prevents leaving too much data from the identity matrix and zero vector. Note that the third adaptation method, LTN embedding, has a strong restriction compared to the second adaptation method.

4. Experimental Evaluation

We conducted experiments to evaluate the performance of our proposed activity recognition model using the Nagoya-COI daily activity database described in Sect. 2.

4.1 Subject-Closed Experiment

First, we conducted a subject-closed experiment in which the same subject's data was used in both the training and test phases, the results of which will represent the performance under the condition where a large amount of subject-specific data can be prepared in advance. The target activities are shown in Table 3, where the numbers in brackets represent the length of the recorded data in minutes. For this experiment we used a long-term, single person dataset, and the most frequently observed nine activities were used as the target activities, while all of the remaining activities were used as non-target activity. When multiple activities were occurring simultaneously, we only focused on the primary tag for that sample, i.e., the main activity. A data segment of 60 seconds in length was regarded as one sample, and data segments of less than 60 seconds in length were not used for the experiment.

The experiment was conducted as follows: 1) randomly

Table 3 Target activities in subject-closed experiment.

Tag	Length [min]	Tag	Length [min]
Cleaning	39	Sleeping	1,257
Cooking	108	Smartphone	198
Meal	120	Toilet	61
Note-PC	141	Watching-TV	109
Reading	164	Other	582

select 10 samples from each activity data set as test data; 2) train the network using the remaining data as training data; 3) evaluate performance for the model using the selected test data; 4) repeat steps 1-3 ten times. In order to evaluate several different models fairly, the test data selected to evaluate the first model was also used to evaluate the other models. We used the average F measure for an activity tag as the evaluation criterion, and all of the DNNs were trained using the open source toolkit Torch7 [38] with a single GPU (Nvidia GTX 980).

4.1.1 Effectiveness of Multi-Modal Signals

To confirm the effectiveness of using multi-modal signals, we compared the performance of the following four models using nine types of target activities:

1. Acceleration feature model (Only Acceleration)
2. Acoustic feature model (Only Acoustic)
3. Ensemble model (Ensemble)
4. Integrated model (Integration)

All of these models have the same DNN structure with the exception of the input layer. Weight coefficient w in Eq. (5) was set at 0.75, which was determined experimentally in order to maximize performance.

The experimental result is shown in Fig. 6. *Frame* represents frame F measure at the level of the frame unit. *Sample* represents sample F measure obtained using the majority of the frame recognition results in each sample. When we compare the performance of the single signal models and the multi-modal signal models, we can see that the multi-modal signal models achieved better performance. Therefore, we can say that the use of multi-modal signals is more effective for activity recognition. Second, when we compare the performance of the ensemble and integrated models, we can see that the integrated model achieved better performance. This is because the DNN could extract discriminative features using both acceleration features and acoustic features as inputs.

Based on these results, we used the integrated model as DNN in subsequent experiments.

4.1.2 Comparison with Conventional Models

Next, we compared our integrated DNN model with the following four conventional methods:

1. k-Nearest Neighbor (KNN)
2. Gaussian Mixture Model (GMM)

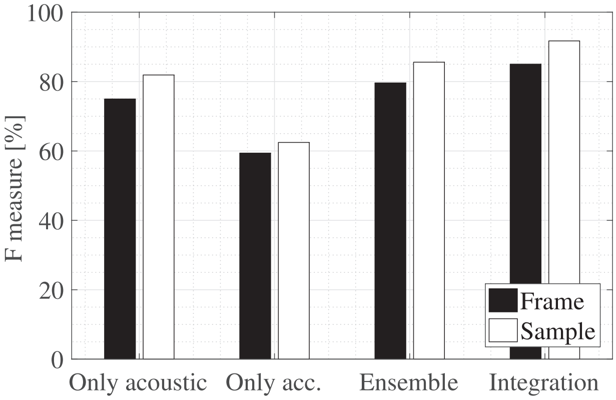


Fig. 6 Comparison of four DNN models.

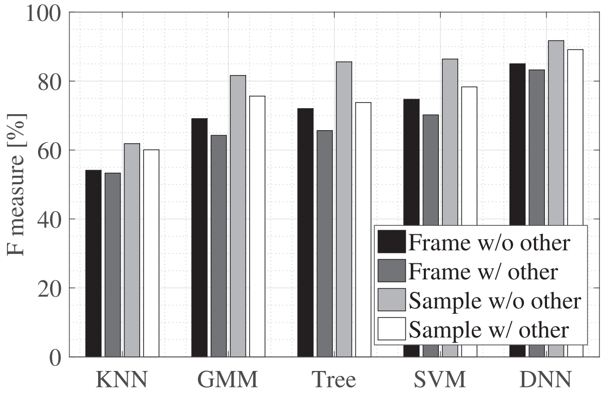


Fig. 7 Comparison of proposed DNN model with other conventional methods.

3. Decision Tree (Tree)
4. Support Vector Machine (SVM)

K in the KNN method was 5, the number of mixtures in the GMM was 10, the kernel function of the SVM was an RBF kernel, and the type of SVM is one-versus-one. The SVM was trained using libSVM [39]. 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. We assumed that our target system also receives signals of an ambiguous activity which is difficult to determine, therefore, we evaluated performance under two conditions: 1) the activity "Other" is not added to target activities (w/o other), and 2) the activity "Other" is added to target activities (w/ other).

Experimental results are shown in Fig. 7. We can see that the DNN-based integrated model performed significantly better than the conventional methods, especially when the activity "Ohter" is added to target activities. The performance of conventional methods such as decision tree or SVM decreased by about 10 points when the activity "Other" was added. This is because the "Other" data is widely distributed and significantly overlapped with the other data in the feature space, and therefore, it is basically difficult to determine the complicated hyperplane. On the other hand, the proposed

DNN-based model maintained its recognition performance when the “Other” was added. This result implies that the DNN can relatively well model such a complicated hyperplane. A visualization of third hidden layer outputs using t-SNE [40] is shown in Fig. 8. We can see that each activity data is distributed separately in the manifold space. This result supports our hypothesis that DNN can model the complicated hyperplane through the conversion to discriminative features using multiple hidden layers.

4.2 Subject-Open Experiment

Next, we conducted a subject-open experiment where the data of different subjects was used in the training and test phases. This experiment evaluates performance when we cannot prepare subject-specific data in advance, and is an important indicator of viability in practical use. For this experiment we used a short-term, multiple-subject dataset, and the target activities are shown in Table 4. The experiment was conducted using leave-one-subject-out validation, where one subject’s data is used as test data, and the remaining data of the other subjects is used as training data.

The experimental results are shown in Fig. 9 and Table 5. From these results we can see that performance in the subject-open evaluation is much lower than in the subject-closed experiment, especially for the activities of “Reading”, “Note-PC” and “Smartphone”, for which the model achieved recognition performance of less than 20 points. There are two reasons for the poor recognition performance for these activities. First, each subject has a very different way of performing these tasks, i.e., there are large differences in subject behavior, even when they are performing the same activity. Examples are shown in Fig. 10, where signals for the activity “Smartphone” are shown for two different subjects. From the figures, we can see that there is a big difference in the recorded signals for each subject as they perform the same activity. Another examples are shown in Fig. 11. In this comparison, one subject’s manner of “Reading” is similar to another subject’s manner of “Sleeping”, since the first sub-

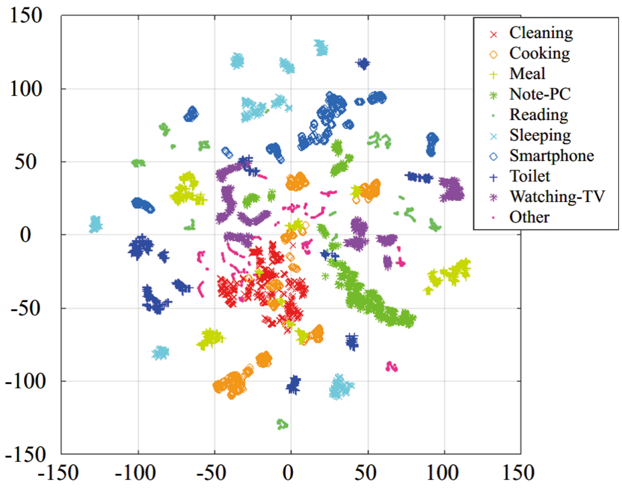


Fig. 8 Visualization of third hidden layer outputs using t-SNE.

Table 4 Target activities in subject-open experiment.

Tag	Length [min]	Tag	Length [min]
Cleaning	679	Sleeping	2,731
Cooking	1,826	Smartphone	1,959
Meal	908	Toilet	310
Note-PC	2,252	Watching-TV	1,873
Reading	476		

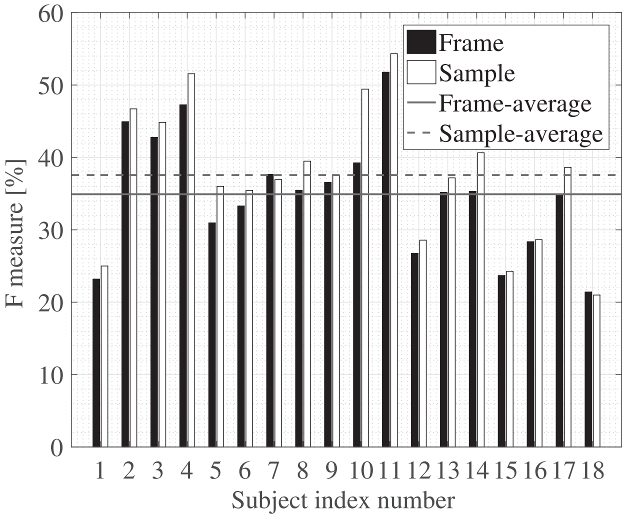
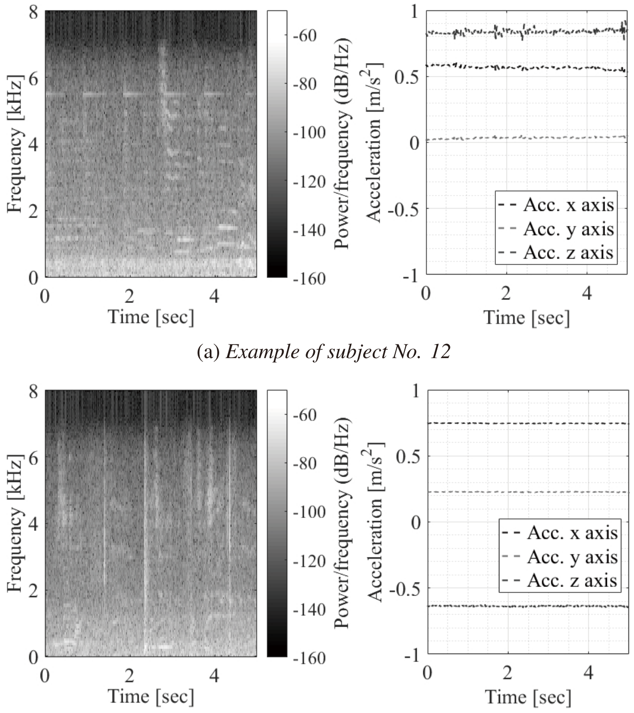


Fig. 9 Leave-one-out result.

Table 5 Confusion matrix of subject-open experiment. Diagonal elements represent recall, that of the right end column represent precision, and that of the lower end row represent F measure.

Recall	Cleaning	Cooking	Meal	Note-PC	Reading	Sleeping	Smartphone	Toilet	Watching-TV	Precision
Cleaning	69.2	28.4	0.9	0.1	0.0	0.0	0.6	0.7	0.0	41.9
Cooking	31.3	64.0	0.5	0.5	0.2	0.8	0.2	2.2	0.3	74.6
Meal	1.0	6.1	55.9	12.6	1.2	3.1	4.2	3.2	12.8	44.3
Note-PC	0.3	1.7	10.7	22.2	15.4	16.4	5.6	3.2	24.5	40.7
Reading	0.4	1.9	9.0	13.9	6.7	19.1	8.0	21.8	19.1	3.9
Sleeping	0.0	0.0	0.2	8.1	7.3	66.8	7.9	4.8	4.9	64.9
Smartphone	1.0	1.5	8.2	10.5	3.9	17.8	16.7	4.7	35.7	32.6
Toilet	10.3	15.5	1.3	2.6	3.9	8.7	0.3	54.8	2.6	24.4
Watching-TV	0.5	1.2	9.1	5.6	7.3	5.8	13.2	2.8	54.4	38.8
F measure	52.2	68.9	49.5	28.8	4.9	65.8	22.1	33.7	45.3	41.2



(a) *Example of subject No. 12*

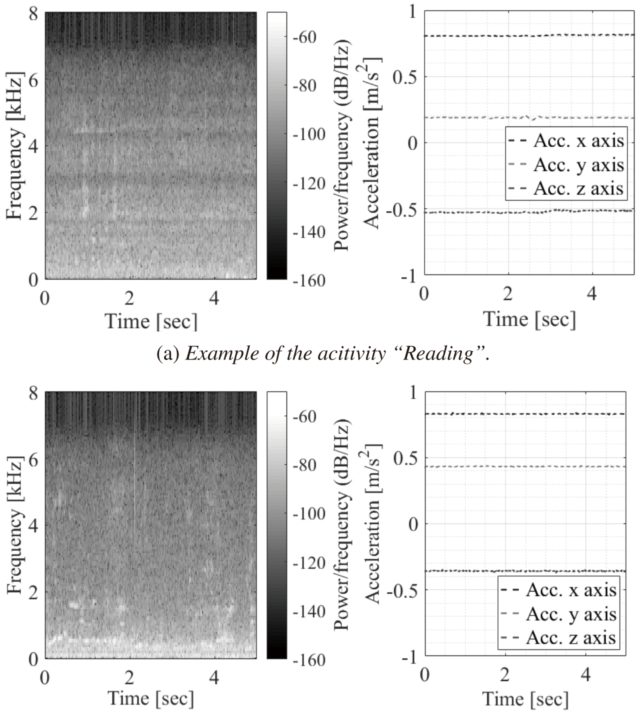
(b) *Example of subject No. 15*

Fig. 10 Example of recorded signals of the activity "Smartphone" for two different subjects. The left figure represents a spectrogram of sound signal, and the right figure represents the recorded acceleration signals.

ject reads while lying on a bed. In addition, such activities do not tend to emit frequent, characteristic sounds, making them harder to detect. A second reason is the lack of uniform orientation of the smartphone in the rear pockets of the subjects' trousers. Subjects were instructed how to attach the smartphone in their pockets, but some subjects were not careful about the orientation of the smartphone. To investigate this problem in more detail, we extracted data when the subject was standing without making any movements, and divided the rotation of smartphone in the rear pocket into four patterns with the mean of acceleration signals. The results are shown in Fig. 12, where each axis represents the axis of the noted acceleration signals as recorded by the smartphone in each of the possible orientations, and where the percentages represent the proportion of subjects who positioned their phone in each orientation. From these results, we can confirm that subjects were not careful about the orientation of the smartphone, and that even if subjects perform the same activity in the same manner, recorded acceleration signals will be very different if the smartphone is not in the same position. Therefore, it is necessary to perform a pre-processing to estimate the actual orientation of smartphone, or extract the new feature which is independent of the orientation of the smartphone.

4.3 Adaptation Experiment

Finally, we conducted an adaptation experiment to see if we could improve activity recognition performance with only a small amount of subject-specific data. To confirm the effectiveness of the adaptations, we compared performance after adaptation with performance when the model is constructed using a random initialization. It is expected that performance when using the adaptations will be better than when using a random initialization if the adaptation methods are effective. The adaptation experiment was conducted as follows: 1) build the subject-independent model using a short-term, multiple-subject dataset as training data; 2) randomly select adaptation samples for each activity class from a long-term, single-subject dataset; 3) apply an adaptation method to the



(a) *Example of the acitivity "Reading".*

(b) *Example of the acitivity "Sleeping".*

Fig. 11 Example of recorded signals of the activities "Reading" and "Sleeping" for two different subjects. The left figure represents a spectrogram of sound signal, and the right figure represents the recorded acceleration signals.

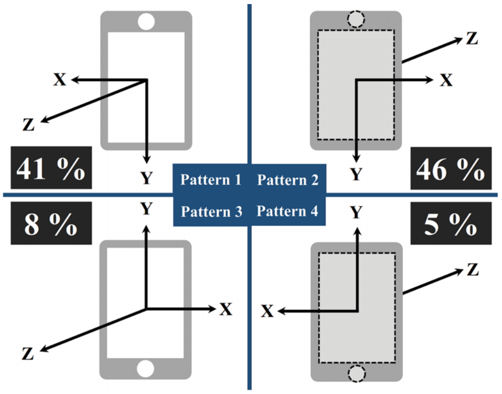


Fig. 12 Analysis of smartphone rotation in the rear pocket.

subject-independent model using selected samples; 4) evaluate performance using test data selected in the subject-closed experiment; 5) repeat steps 2-4 while increasing the number of adaptation samples. In this experiment, we selected $N = \{1, 2, 3, ..., 25\}$ samples from each activity class for adaptation, 10 samples from each activity class for test data, and repeated these steps ten times. When we applied the second adaptation method, re-training only a specific layer, we selected the second hidden layer as the adaptation layer, and set the regularization coefficient in Eq. (6) to $1e - 6$. When we applied the third adaptation method, LTN embedding, we inserted a linear transformation layer before the second hidden layer, and set the regularization coefficient in Eq. (7) to $5e - 6$. These adaptation layer and regularization coefficients were determined through preliminary experiments.

Experimental results are shown in Fig. 13. We can see that performance when using the adaptation methods is better than when using random initialization. This result shows that all of the adaptation methods are effective even if only a small amount of subject-specific training data is available. In Fig. 13, when the number of adaptation samples is from one to ten, there is no difference in performance between these adaptation methods. However, when the number of adaptation samples is more than ten, performance with re-training only a specific layer tend to become saturated. Hence, it is better to use other adaptation methods if we can prepare adaptation data more than ten. There is no significant difference between performance using all layer re-training and when using LTN embedding, however, the number of parameters to keep for each subject are significantly fewer when using LTN embedding than when using all layer re-training. Therefore, we can say that LTN embedding is a suitable adaptation method in terms of not only improving performance, but also requiring limited computational resources.

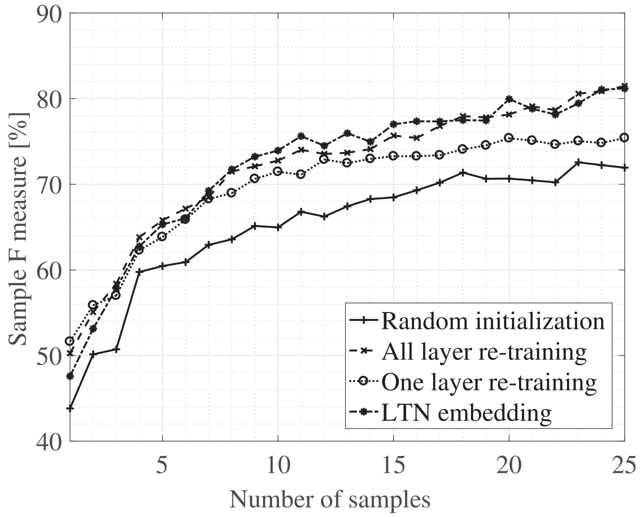
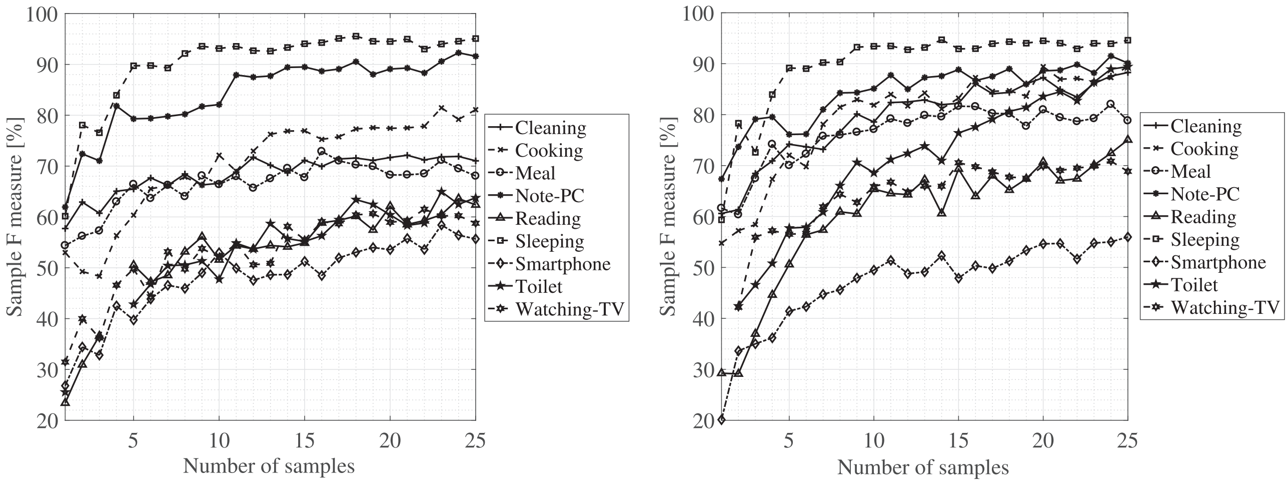


Fig. 13 Class average performance using various adaptation methods and different numbers of samples.

Finally, the change in classwise performance is shown in Fig. 14, where the graph on the left represents change in the activity recognition rate when using random initialization, and the graph on the right represents that when using LTN embedding as an adaptation method. By comparing these results, we can see that the adaptation method is effective for improving accuracy of most of the activities, but not for activities such as "Sleeping", "Note-PC", and "Smartphone". For "Sleeping" and "Note-PC", even if we use a random initialization the performance is nearly 90 points, therefore there is little room for improvement using an adaptation method. Recognition performance for "Smartphone" was poor even when an adaptation method was used. One reason for this poor performance is that there was a great deal of inconsistency in the manner in which subjects used their smartphones (See the example in Fig. 10). Therefore, using other subject's data have no advantage to recognize the activity "Smartphone".



(a) Random initialization

(b) LTN embedding

Fig. 14 Change in performance for various activities when using different number of samples w/o and w/ adaptation.

5. Conclusion

In this study, toward the development of smartphone-based monitoring system for life logging, we collected over 1,400 hours data including both outdoor and indoor daily activities of 19 subjects in a practical situation with a smartphone and small camera, and then constructed huge daily activity database which consists of an environmental sound signal, triaxial acceleration signals, and manually annotated tags. We evaluated the activity recognition performance of DNN-based methods using our constructed database, and apply DNN-based adaptation methods to improve the performance even if only a small amount of subject-specific data is available. We experimentally demonstrated that the use of multimodal including an environmental sound and triaxial acceleration signals with DNN is effective for the improvement of performance, and DNN can discriminate specified activities from a mixture of unspecified activities. Furthermore, we confirmed that DNN-based adaptation method is effective even if only a small amount of subject-specific training data is available.

In future works, we will expand the focus of the proposed model to recognize not only indoor but also outdoor activities. We will also investigate a method of pre-processing which would allow us to compensate for differences in the orientation of the smartphone used to monitor acceleration.

Acknowledgements

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

References

[1] C. Gurrin, A.F. Smeaton, and A.R. Doherty, "Lifelogging: Personal big data," Foundations and Trends® in Information Retrieval, vol.8, no.1, pp.1–125, 2014.

[2] M.P. Rajasekaran, S. Radhakrishnan, and P. Subbaraj, "Elderly patient monitoring system using a wireless sensor network," Telemedicine and e-Health, vol.15, no.1, pp.73–79, 2009.

[3] Q. Lin, D. Zhang, X. Huang, H. Ni, and X. Zhou, "Detecting wandering behavior based on GPS traces for elders with dementia," Control Automation Robotics Vision, pp.672–677, IEEE, 2012.

[4] Y.T. Peng, C.Y. Lin, M.T. Sun, and K.C. Tsai, "Healthcare audio event classification using hidden Markov models and hierarchical hidden Markov models," IEEE International Conference on Multimedia and Expo, pp.1218–1221, IEEE, 2009.

[5] Y. Liang, X. Zhou, Z. Yu, and B. Guo, "Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare," Mobile Networks and Applications, vol.19, no.3, pp.303–317, 2014.

[6] D. Valtchev and I. Frankov, "Service gateway architecture for a smart home," IEEE Commun. Mag., vol.40, no.4, pp.126–132, 2002.

[7] P. Rashidi and D.J. Cook, "Keeping the resident in the loop: Adapting the smart home to the user," IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, vol.39, no.5, pp.949–959, 2009.

[8] C. Zhang and Y. Tian, "RGB-D camera-based daily living activity recognition," J. Computer Vision and Image Processing, vol.2, no.4, p.12, 2012.

[9] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," IEEE Trans. Neural Netw. Learning Syst., vol.23, no.3, pp.412–424, 2012.

[10] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp.4772–4781, IEEE, 2016.

[11] M.S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp.1971–1980, IEEE, 2016.

[12] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," EURASIP Journal on Audio, Speech, and Music Processing, vol.2013, no.1, pp.1–13, 2013.

[13] T. Komatsu, T. Toizumi, R. Kondo, and Y. Senda, "Acoustic event detection method using semi-supervised non-negative matrix factorization with mixtures of local dictionaries," Proc. Detection and Classification of Acoustic Scenes and Events 2016 Workshop, pp.45–49, September 2016.

[14] T. Hayashi, S. Watanabe, T. Toda, T. Hori, J. Le Roux, and K. Takeda, "BLSTM-HMM hybrid system combined with sound activity detection network for polyphonic sound event detection," IEEE International Conference on Acoustics, Speech and Signal Processing, pp.766–770, IEEE, 2017.

[15] D.J. Patterson, D. Fox, H. Kautz, and M. Philipose, "Fine-grained activity recognition by aggregating abstract object usage," IEEE International Symposium on Wearable Computers, pp.44–51, IEEE, 2005.

[16] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J.M. Rehg, "A scalable approach to activity recognition based on object use," IEEE International Conference on Computer Vision, pp.1–8, IEEE, 2007.

[17] 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 Comput., vol.3, no.4, pp.50–57, 2004.

[18] A. Fleury, M. Vacher, and N. Noury, "SVM-based multimodal classification of activities of daily living in health smart homes: Sensors, algorithms, and first experimental results," IEEE Trans. Inf. Technol. Biomed., vol.14, no.2, pp.274–283, 2010.

[19] J.R. Kwapisz, G.M. Weiss, and S.A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, vol.12, no.2, pp.74–82, 2011.

[20] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford, "A hybrid discriminative/generative approach for modeling human activities," 2005.

[21] L. Bao and S.S. Intille, "Activity recognition from user-annotated acceleration data," International Conference on Pervasive Computing, pp.1–17, Springer, 2004.

[22] T. Huynh and B. Schiele, "Towards less supervision in activity recognition from wearable sensors," IEEE International Symposium on Wearable Computers, pp.3–10, IEEE, 2006.

[23] P. Lukowicz, J.A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner, "Recognizing workshop activity using body worn microphones and accelerometers," International Conference on Pervasive Computing, pp.18–32, Springer, 2004.

[24] K. Ouchi and M. Doi, "Smartphone-based monitoring system for activities of daily living for elderly people and their relatives etc.," ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication, pp.103–106, ACM, 2013.

[25] T. Maekawa, Y. Yanagisawa, Y. Kishino, K. Ishiguro, K. Kamei, Y. Sakurai, and T. Okadome, "Object-based activity recognition with heterogeneous sensors on wrist," International Conference on Pervasive Computing, pp.246–264, Springer, 2010.

[26] A. Reiss and D. Stricker, "Introducing a new benchmarked dataset for activity monitoring," International Symposium on Wearable Computers, pp.108–109, IEEE, 2012.

[27] R. Chavarriaga, H. Sagha, A. Calatroni, S.T. Digumarti, G. Tröster,

J.d.R. Millán, and D. Roggen, “The opportunity challenge: A benchmark database for on-body sensor-based activity recognition,” Pattern Recog. Lett., vol.34, no.15, pp.2033–2042, 2013.

[28] H. Guo, L. Chen, L. Peng, and G. Chen, “Wearable sensor based multimodal human activity recognition exploiting the diversity of classifier ensemble,” ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp.1112–1123, ACM, 2016.

[29] F.J. Ordóñez and D. Roggen, “Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition,” Sensors, vol.16, no.1, p.115, 2016.

[30] N.Y. Hammerla, S. Halloran, and T. Ploetz, “Deep, convolutional, and recurrent models for human activity recognition using wearables,” arXiv preprint arXiv:1604.08880, 2016.

[31] T. Hayashi, M. Nishida, N. Kitaoka, and K. Takeda, “Daily activity recognition based on dnn using environmental sound and acceleration signals,” European Signal Processing Conference, pp.2306–2310, IEEE, 2015.

[32] “ELAN-Linguistic Annotator,” http://www.mpi.nl/corpus/html/elan

[33] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, “Context-dependent sound event detection,” EURASIP Journal on Audio, Speech, and Music Processing, vol.2013, no.1, pp.1–13, 2013.

[34] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.

[35] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv preprint arXiv:1207.0580, 2012.

[36] T. Ochiai, S. Matsuda, X. Lu, C. Hori, and S. Katagiri, “Speaker adaptive training using deep neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp.6349–6353, IEEE, 2014.

[37] T. Ochiai, S. Matsuda, H. Watanabe, X. Lu, C. Hori, and S. Katagiri, “Speaker adaptive training for deep neural networks embedding linear transformation networks,” IEEE International Conference on Acoustics, Speech and Signal Processing, pp.4605–4609, IEEE, 2015.

[38] “Torch 7|A Scientific Computing Framework for Luajit,” http://torch.ch/

[39] C.C. Chang and C.J. Lin, “LIBSVM: A library for support vector machines,” ACM Trans. Intell. Syst. Technol., vol.2, no.3, p.27, 2011.

[40] L.v.d. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Machine Learning Research, vol.9, no.Nov, pp.2579–2605, 2008.

Tomoki Hayashi received his B.E. degree in engineering and M.E. degree in information science from Nagoya University, Japan, in 2013 and 2015, respectively. He is currently a Ph.D. student at the Nagoya University. His research interest include statistical speech and audio signal processing. He received the Acoustical Society of Japan 2014 Student Presentation Award. He is a student member of the Acoustical Society of Japan, and a student member of the IEEE.

Maasfumi Nishida Masafumi Nishida received a B.E. in 1997, an M.E. in 1999, and a Ph.D. in 2002, all in Electronics and Informatics, and all from Ryukoku University, Shiga, Japan. From 2002 to 2003, he was a post-doctoral researcher at PRESTO, Japan Science and Technology Corporation. In 2003, he became a Research Associate at the Department of Information Science, Chiba University. From 2007 to 2008, he was an Assistant Professor at the Graduate School of Advanced Integration Science, Chiba University. From 2009 to 2013, he was an Associate Professor at the Department of Information Systems Design, Doshisha University. In 2014, he was an Designated Associate Professor at the Institute of Innovation for Future Society, Nagoya University. Currently, he is an Associate Professor at the Department of Informatics, Shizuoka University. His research interests include speech recognition, speaker recognition, spoken dialogue systems, well-being information technology, and behavior signal processing. He received the 2011 Yamashita SIG Research Award from IPSJ. He is a member of the IEICE, the IPSJ, the ASJ and the JSAI.

Norihide Kitaoka received his B.S. and M.S. degrees from Kyoto University. In 1994, he joined DENSO CORPORATION. In 2000, he received his Ph.D. degree from Toyohashi University of Technology (TUT). He joined TUT as a research associate in 2001 and was a lecturer from 2003 to 2006. He became an associate professor in Nagoya University in 2006. Since 2015 he has been a professor in Tokushima University.

Tomoki Toda received his B.E. degree from Nagoya University, Japan, in 1999 and his M.E. and D.E. degrees from Nara Institute of Science and Technology (NAIST), Japan, in 2001 and 2003, respectively. He was a Research Fellow of the Japan Society for the Promotion of Science from 2003 to 2005. He was then an Assistant Professor (2005–2011) and an Associate Professor (2011–2015) at NAIST. From 2015, he has been a Professor in the Information Technology Center at Nagoya University. His research interests include statistical approaches to speech and audio processing. He received more than 10 paper/achievement awards including the IEEE SPS 2009 Young Author Best Paper Award and the 2013 EURASIP-ISCA Best Paper Award (Speech Communication Journal).

Kazuya Takeda received his B.E.E., M.E.E. and Ph.D. from Nagoya University. After graduating from Nagoya University in 1985, he worked at Advanced Telecommunication Research Laboratories and at KDD R&D Laboratories, Japan, mostly in the area of speech signal processing. He was a Visiting Scientist at the Massachusetts Institute of Technology from October 1988 to April 1989. In 1995, Dr. Takeda moved to Nagoya University, where he started a research group for signal processing applications. Since then he has been working on a wide range of research topics, including acoustics and speech, as well as driving behavior. He is the co-author of more than 100 journal articles and five books. Dr. Takeda is currently a Professor at the Graduate School of Informatics and the Green Mobility Collaborative Research Center, Nagoya University, Japan.