International Journal of u- and e- Service, Science and Technology
Vol.9, No. 6 (2016), pp.409-420
http://dx.doi.org/10.14257/ijunesst.2016.9.6.37

Object Learning Through Interaction with Person in the Context of Finding Lost Objects

Junji Takahashi, Ryuji Suzuki, and Yoshito Tobe

Department of Integrated Information Technology
Aoyama Gakuin University, Sagamihara, Japan
takahashi@it.aoyama.ac.jp, liuji@rcl-aoyama.jp, yoshito-tobe@rcl-aoyama.jp

Abstract

This paper deals with machine learning through interaction with a person. We developed an interactive learning system (LTIP) by concentrating on a specific situation where a user and a robot cooperatively find a lost object in real world. The LTIP consists of a gathering candidate images process, a self-categorizing process based on SOM, and a cooperative and repeatable narrow down process so that the robotic system effectively learns the appearance of the target object that the user is looking for. We conducted several experiments to evaluate and analyze the process of interactive learning. We confirmed that preliminarily determining the effective features set for categorization is difficult. So our proposed LTIP scheme is useful for such a practical processes such as finding a lost object.

Keywords*: interactive learning; Self-Organizing Map*

1. Introduction

In many countries, people suffer from the shortage of labor force due to low birth rates. A solution to this problem is using service robots for assembly lines, and medical and nursing care. In particular, service robots have been used more often to assist nursing care in Japan. With this increase in the importance of nursing care service robots, they can be further divided into several types depending on the type of activity that they can help with: assistance of impaired walking, helping the elderly relax by making conversation with them, and substitution of daily-life chores such as room cleaning.

In this paper, we focus on service robots that assist conversation with the elderly to accelerate maintaining the mental health of the elderly. Furthermore, we investigate the function of interaction with people itself in robots instead of their mechanical appearance; they do not need to be humanoids.

Currently, service robots are not being used practically. We can see them in a booth as a research demonstration, but not at home. This results from the fact that they do not have sufficient ability of recognizing objects and their environments and understanding unknown objects in addition to conversation.

There has been a lot of research conducted to solve the above problems. Lim *et al.*[2] proposed a scheme to visualize the contour of an object for recognizing objects and environments. Kawanishi *et al.* [3] proposed a method to use color features for detecting fast object recognition. Masuta *et al.* [5] established a method to identify an unknown object by having created a model of the object using its image with depth. Additionally, Akgun1 *et al.* [6] investigated how the interaction between humans and robots influenced communications. Osawa *et al.* [7] observed that humans changed their ways of communicating unconsciously when interacting with a robot.

Unlike these approaches, we propose a method of recognizing unknown objects by a service robot using interactions between the robots and humans. Our method relies on iterative use of Self-Organizing Map (SOM) to categorize the objects obtained from the

ISSN: 2005-4246 IJUNESST

World Wide Web (WWW), using a search word given by a human until the intended object is found. We call this learning process Learning Through Interaction with Person (LTIP). LTIP does not necessitate databases of objects beforehand, which does not require a large storage.

The rest of the paper is organized as follows. Section 2 describes related works. Section 3 describes LTIP. Section 4 describes experimental evaluation. Section 5 describes conclusion.

2. Related Works

There is a method of creating a model to classify images on the Web [1]. This method needs to collect many images from the web and to create a database to classify them. The drawback of this method is that it does not accommodate classification of unknown objects. In addition, it requires large storage.

Another related area is recognition of objects using SOM with Scale-Invariant Feature Transform (SIFT) feature [4]. In this method, SIFT is calculated from the input image and key points are created, until finally an object is recognized. Creating the key points will need a large volume of images for learning and this method also needs preparation beforehand.

Unlike these previous works, we do not need the database of classified images and preparation. Furthermore, real-time identification of a required image is made more possible.

3. LTIP: Learning Through Interaction with Person

The LTIP (Learning Through Interaction with Person) is a learning scheme working on a robot, which identifies the visual appearance of a lost object efficiently. We assumed a scenario of finding things to discuss with the LTIP about more practical situations. When the robot is asked to look for a thing it does not know, it gathers images as candidates of the thing from the Web using the Bing Search API. In many cases, because the gathered images are too many, they need to be categorized to facilitate the selection process. We innovate 12 of the features extraction method before the categorization and adopt Self-Organizing Maps (SOM) for non-supervised categorization. The categorized results promote effective use of the narrow down process of the target object. The user and the robot cooperatively repeat the categorization process and the narrow down process so that the robot grabs the target object. The details are described in each section.

3.1. A Scenario of Finding a Thing

We designed a human-robot interaction scenario in the context of finding a thing. We assumed the case where a user searches for an object in a laboratory and he/she wants a partner robot to look for it together. However, the robot does not know the visual appearance of the object. So, the first action the user has to do is to teach the robot what the visual appearance of the target object is.

If the robot is a slow learner, the teaching process becomes boring. To prevent the user from boredom, the robot utilizes Web knowledge to rapidly estimate and grab the visual appearance of the target object. Then the user and the robot engage the cooperative narrow down process, using the LTIP to teach and learn target object efficiency.

3.2. Bing Search API

The Bing Search API is provided by Microsoft to enable developers to embed search results in applications or websites using XML or JSON. We use Bing Search API for gathering candidate images of lost objects with the objects' names as a keyword.

3.3. Features of Images

It is necessary for the robot to identify the visual appearance of the object in order to distinguish images from one another. However, the robot cannot distinguish prototypic images. Therefore, the robot calculates features of images, quantifies the features, and distinguishes the images. At this time, if the robot uses prototypic getting images, the features have occurred variance because the size of the images is non-constant. Therefore, the size of the images' aspect ratio is maintained and the size of the images kept constant in order to avoid the variation.

We prepare 12 features, features of RGB, features of binalized image, Hog features, features of Hough conversation, and Conner futures. So, each image has a 12 dimensional vector: f_q ($q=1,2,..,12$). We talk in detail below for each feature and visualize the image features of image with a cup image (Figure 1).

3.3.1. Features of RGB: Features of RGB represent a dominating color of an image. Each pixel of the image has three dimensional color axes such as red, green, blue, with value from 0 to 255. We divided each color axis into 32 bins and count pixels, which belongs to each bin range, to make a color histogram. Then the bin that has maximum pixels was chosen to the value of the feature after normalization. Let p_j^C be number of pixels belonging to the j-th bin of color C ($j=1, 2, 3...32$; $C = R, G, B$). Then the elements of features of RGB are given as follows,

$$f_1 = \frac{1}{32}\arg\max_j \left(p_j^R\right), \quad f_2 = \frac{1}{32}\arg\max_j \left(p_j^G\right), \quad f_3 = \frac{1}{32}\arg\max_j \left(p_j^B\right). \tag{1}$$

3.3.2. Features of Binalized Image: An image binarization is a conversion method to change an input image into an output image only with black and white pixels (Figure 2). The luminance of each pixel is used for this conversion. To evaluate the luminance of each pixel, the input image is converted into gray-scale image with a value from 0 to 255. Then each pixel is converted into black (=0) or white (255) depending on lower or higher to the threshold value: T_b. The feature of binalized image is given as the ratio of number of black pixels: n_T to all pixels: N in the image. We prepared three thresholds: $T_b(102)$, $T_b(153)$, $T_b(204)$, so three features are given as follows,

Figure 1. A Cup Image

Figure 2. Examples of Binalized Images (from Left *T=102*, *T=153*, *T=204*)

$$f_4 = \frac{n_{T_b(102)}}{N}, \quad f_5 = \frac{n_{T_b(153)}}{N}, \quad f_6 = \frac{n_{T_b(204)}}{N}. \tag{2}$$

3.3.3. HOG Feature: The HOG (Histogram of oriented Gradients) is often used for the purpose of object detection and recognition. The HOG is robust to geometric transformation and variations in lighting. Each pixel, except for edges in the picture, has eight pixels neighboring it (Figure 3). Let gradient direction and gradient strength denote *d (1,…,8)* and *s (-255 < s <255)*, respectively. The gradient of direction *d* on *i-th* ($i \in N$ - edges) pixels are s_i^d. Then, we defined the HOG feature in this paper as follows,

$$f_7' = \frac{1}{N - (edges)} \sum_{i=1}^{N-(edges)} \frac{1}{8} \sum_{d=1}^{8} s_i^d \quad . \tag{3}$$

3.3.4. Features Based on Hough Conversion: The Hough conversion is one of coordinate conversions, which converts an orthogonal coordinate to a polar coordinate, to detect straight lines and circles in an image. After Hough conversion, the image has several line segments and circular segments (Figure 4). Let L_n denote the number of line segments, L_d denote the total length of lines, C_n denote the number of circular segments, and C_d denote the total length of circular segments. We defined the features of Hough conversion as follows,

$$f_8' = L_d \ , \quad f_9' = L_n \ , \quad f_{10}' = C_d \ , \quad f_{11}' = C_n \ . \tag{4}$$

3.3.5. Corner Feature: A corner is defined as the intersection of two edges. It means that the corner pixel has two different dominant gradients among the neighbor pixels. Therefore, the calculation of corner feature is the same as HOG feature calculation until getting the parameter s_i^d. We defined T_c as the threshold representing the dominancy of the gradient, and that if two or more s_i^d exceed the T_c, then the pixel *i* is a corner (Figure 5). Finally we defined the corner feature as the number of corners in the picture,

$$f_{12}' = (number \quad of \quad corners) \ . \tag{5}$$

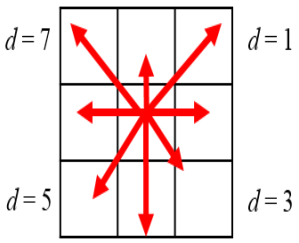


Figure 3. Gradients of Eight Directions Defined on a Pixel

Figure 4. Visualization of Features of Hough Conversion

Figure 5. Visualization of Corner Feature

Figure 6. Examples of Cup Image (left: I_1, right: I_2)

3.3.6. Normalization: The features f_1~f_6 have already been normalized. Meanwhile the features f_7'~f_{12}' have not been normalized yet. These latter features are normalized by using the relation among other images. Let *k (1,…,K)* denote index of image, while the features normalization are calculated as follows,

$$f^{k}_{q} = \frac{f_q^{k\prime} - \min\limits_{1\le k\le K}\left(f_q^{k\prime}\right)}{\max\limits_{1\le k\le K}\left(f_q^{k\prime}\right) - \min\limits_{1\le k\le K}\left(f_q^{k\prime}\right)}, \ (q=7\ldots 12). \tag{6}$$

3.3.7. Examples of Calculated Features: Figure 6 and Table 1 show examples of gathered images and 12 dimensions of their features. The 12 dimensional features are divided into two groups. One is a color based feature group, which consists of RGB features and binalized features represented as f_1 ~f_6. The other is shape based features group, which consists of HOG features, features based on Hough conversion, and corner features represented as f_7 ~f_{12}. We call the former group as CF and the latter as SF.

Table 1. Example Feature Vectors of Cup Image I_1 and I_2

	Color based features (CF)						Shape based features (SF)					
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}
I_1	*0.688*	*0.688*	*0.281*	*0.442*	*0.554*	*0.748*	*0.784*	*0.590*	*0.349*	*0.677*	*0.894*	*0.938*
I_2	*0.938*	*0.156*	*0.188*	*0.565*	*0.640*	*0.712*	*0.126*	*0.194*	*0.133*	*0.137*	*0.156*	*0.325*

3.4. Clustering of Images by SOM

This section explains the SOM (Self-Organizing Maps) algorithm for non-supervised categorization. Let $\boldsymbol{x} = [x_1, x_2, \cdots, x_{12}]^T$ denote input vector, y_r $(1 \le r \le R)$ denote *r-th* output unit whose *r* represents a class name, and $\boldsymbol{w}_r = [w_{r1}, w_{r2}, \cdots, w_{r12}]^T$ denote weight vector, then they have a relation as follows,

$$y_r = \boldsymbol{w}_r^T \boldsymbol{x} \ . \tag{7}$$

The output units are competing with each other and only one winner *(r=c)* unit can output (winner-takes-all),

$$y_r = \begin{cases} \boldsymbol{w}_r^T \boldsymbol{x}, & \boldsymbol{w}_r^T \boldsymbol{x} = \max\limits_{1\le l\le R}\left(\boldsymbol{w}_l^T \boldsymbol{x}\right) \\ 0, & \boldsymbol{w}_r^T \boldsymbol{x} < \max\limits_{1\le l\le R}\left(\boldsymbol{w}_l^T \boldsymbol{x}\right) \end{cases} . \tag{8}$$

When the $\boldsymbol{w}_r$ and the $\boldsymbol{x}$ are normalized, the maximization of inner product in eq. (7) is equivalent to the minimization of Euclidean norm between $\boldsymbol{w}_r$ and the $\boldsymbol{x}$. So, the y_c is given as follows,

$$y_c = \arg\min_{1\le r\le R} \left\| \boldsymbol{x} - \boldsymbol{w}_r \right\| . \tag{9}$$

The learning of weight of the winner unit and its neighbor is done with apportioning of the difference between weight vector and input vector. The amount of apportionment is determined by topological neighborhood function: h_c^r using the distance: d_{cr}. Let $\boldsymbol{Y}_c$ and $\boldsymbol{Y}_r$

represent position vector of output units, then $d_{cr} = \|\mathbf{Y}_c - \mathbf{Y}_r\|$. So that the weight update process should have time convergence, the h_{cr} is defined as follows,

$$h_{cr}(t) = c \cdot \exp\left(- \frac{\|\mathbf{Y}_c(t) - \mathbf{Y}_r(t)\|^2}{\left(1 - \frac{t}{T}\right)^2} \right), \quad (10)$$

where t $(<T)$ represents the leaning times. Finally, the weight renewal rule is given as follows,

$$\mathbf{w}_r(t) \leftarrow \mathbf{w}_r(t) + h_{cr}(t)\big(\mathbf{x}(t) - \mathbf{w}_r(t)\big) \ . \quad (11)$$

When there are K number of input images, the SOM algorithm can be summarized in Table 2.

3.5. Narrow Down Process for Identifying the Object

The narrow down process is conducted using the result of SOM categorization. The robot shows the divided categories to the user and asks which category is closest to the target object. The disclosure of the divided categories is done with showing one representative image that is closest to the center of the category. This is a consideration for lightening the burden of the user. The user selects a category by touching a representative image that is closer to the target object.

If the selected category still has number of images, the robot executes the SOM process and prompts the user to make selections again and again until the robot grabs the target image. This interaction process is not boring for the user, and also brings in a sense of togetherness for the user. Through this interactive process, the robot can approach the recognition of correct appearance of the target object, but also get a sense of understanding of the user's worth by analyzing this process afterward. Integrating the above-explained processes, the whole process of the LTIP can be written down as seen in Table 3.

Table 2. SOM Algorithm

(i)	Initialize the elements of y_r , set $t=0$ and $k=0$.
(ii)	$t \leftarrow t+1$
(iii)	$k \leftarrow k+1$
(iv)	Decide a winner unit y_c by eq. (16)
(v)	Update w_c and by eq. (18)
(vi)	If $k < K$,then go to (iii)
(vii)	If $t < T$, then go to (ii)
(viii)	End

Table 3. LTIP Interaction Protocol

(i)	A user tells the object name of what he/she is looking for to a robot
(ii)	The robot gathers images of the object from the Web using Bing Search API
(iii)	Resize the images to become same size each other
(iv)	Calculate features of all images
(v)	Categorize the images using SOM
(vi)	Prompt the user to make category selection with showing images representing the category
(vii)	The user select the closest image what he/she is looking for
(viii)	If the target object is not detected, then go to (v)

4. Experimental Evaluation

We conducted experiments to evaluate the LTIP. The experiments made use of pictures that were gathered before the experiment from Web pages using the Bing Search API with a prepared keyword. A user who joins the experiment voluntary is uninvolved the LTIP system development. The user selects a picture in his/her mind from the catalog of gathered pictures. Then the user teaches the selected picture to the system with iterative and interactive selection process. We count the iteration times in the whole process until the system gets the user-selected picture. We also analyze the within-class variance, inter-class variance and their ratio of groups in each step. Three types of experiment, the first one using only color features (CF), the second one using shape features (SF), and the third one using both features (SCF), are conducted.

4.1. Experimental Setup

We prepared 84 images (Figure 7) that were obtained by Bing Search API with the keyword "cup." Then, we defined three images as Target 1 (T1), Target 2 (T2), and Target 3 (T-3) (Figure 8). The test user interactively teaches the target image to the system using the LTIP manner. Figure 9 shows a scene that the system discloses using the divided categories. In this case, each category has more than 3 images; so only one representative image is shown in the category window. On the other hand, when either category has less than 4 images, that category shows all images to the window (Figure 10).

We set an upper limit to repetitive times of interactive process including SOM categorization and user selection. When the repetitive times reach 10, the trial is stopped and is regarded as failed. A number of ten users joined this experiment. Three trials per feature case, such as CF, SF, and CSF, were conducted per user. In total, 270 trials data were obtained.

Figure 7. The Gathered Images for Experiment

Figure 8. The Target Images in the Experiment (from Left, Target 1, Target 2 Target 3)

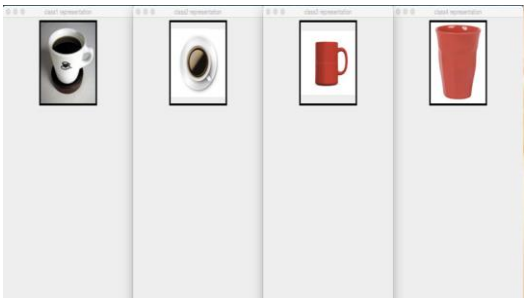


Figure 9. A Scene that the System Discloses the Categories Showing a Representative Image

Figure 10. A Scene Either Category has Three Images

4.2. Analysis Method

The trials are analyzed based on the number of times for detecting the target and the doneness of the categorization result on each SOM process. We introduce within-class variance, between-class variance and their ratio named class-variance ratio to measure the doneness of the categorization. Let N_i denote number of members in a class i, $\boldsymbol{x} \in X_i$ denote member vector in the class, $\boldsymbol{m}_i$ denote average vector in the class, and I denote number of classes, then the within-class variance is defined as follows

$$\sigma_w^2 = \frac{1}{I}\sum_{i=1}^{N}\frac{1}{N_i}\sum_{\boldsymbol{x}\in \boldsymbol{X}_i}(\boldsymbol{x}-\boldsymbol{m}_i)^T(\boldsymbol{x}-\boldsymbol{m}_i) \quad . \tag{12}$$

Let $\boldsymbol{m}$ denote average of all of vectors, the between-class variance is defined as follows,

$$\sigma_b^2 = \frac{1}{I}\sum_{i=1}^{N}(\boldsymbol{m}-\boldsymbol{m}_i)^T(\boldsymbol{m}-\boldsymbol{m}_i) \quad . \tag{13}$$

The class-variance ratio is defined as follows,

$$J_\sigma = \frac{\sigma_b^2}{\sigma_w^2} \quad . \tag{14}$$

The higher the class-variance ratio: J_σ is, the categorization doneness is better.

4.3. Experimental Results

The numbers of failures are shown in Table 4. The average of numbers of times for detecting the target except for failure cases are shown in Table 5. The average of class-variance ratio in each feature case is shown in Table 6.

In Table 4, the failure number of T-1 is 4 in the case of CF and 7 in the case of CSF. This is because the T-1 is more remarkable in color than in shape. The CSF includes the color feature, but in this case the shape feature wrongly affects and decreases the saliency of the T-1. On the other hand, the failure number of T-3 is 9 in the case of CF and 1 in the case of CSF. This is because the saliency of the T-3 becomes maximum in CSF vector space. These results lead to the fact that the number of features for little failure categorization depends on what the target is.

The same statement can be derived from Table 5. The number of repeatable interaction times depends on what the target is. It has also been observed that the number of times for detecting slightly correlate with number of failures.

The calculation results of class-variance show that the doneness of the categorization depends on how the gathered images population distributes in the vector spaces. Contrary to our expectation, the SF case is higher than the CSF case. This means that the suitable features set depends on the distribution of the population.

Consequently, it can be said that preliminarily determining the effective features set for categorization is difficult. One of the solutions is that the robotic system interactively tunes the feature vector space. Although our proposed LTIP has, so far, not implemented the online feature-tuning scheme, this is one of our future projects.

Table 4. Number of Failures in 30 Trials (%)

	CF	*SF*	*CSF*
T-1	*4 (13.3)*	*10 (33.3)*	*7 (23.3)*
T-2	*8 (26.7)*	*20 (66.7)*	*9 (30.3)*
T-3	*9 (30.0)*	*7 (23.3)*	*1 (3.3)*

Table 5. The Number of Times for Detecting the Target Except for Failure Cases

	CF (SD)	*SF (SD)*	*CSF (SD)*
T-1 (SD)	*2.8 (2.0)*	*3.6 (3.6)*	*3.3 (2.1)*
T-2 (SD)	*5.6 (1.0)*	*4.3 (1.6)*	*4.4 (0.7)*
T-3 (SD)	*3.3 (0.6)*	*1.7 (2.4)*	*1.4 (0.6)*

Table 6. The Average of Class-Variance Ratio in Respective Feature Cases

	CF (SD)	*SF (SD)*	*CSF (SD)*
J_σ *(SD)*	*26.2 (127.6)*	*30.7 (225.9)*	*25.0 (119.9)*

5. Conclusion

We proposed an LTIP (Learning Through Interaction with Person) scheme that enables a user and a robotic system to teach and learn effectively in a situation of cooperatively finding a lost object. When the user wants the robot to help him/her find a lost object, he/she has to teach the appearance of the object to the robot. However, since the object that he/she wants to teach is lost, the teaching process becomes difficult.

Using the LTIP scheme, the robot searches the World Wide Web to obtain candidate images with the keyword given by the user. Then the robot and the user cooperatively narrow down the images until they reach the most similar target object image that the user is looking for. In this process, the robot's role is categorizing images into a few categories. This categorization process is implemented based on the SOM algorithm. The user's role

is selecting a category that tends to include the target object. Since the number of categories shown at a time is not so large and only one representative image is disclosed in a category window, the user rarely feels burdened in the selection process. Although it is rare to obtain the target object in one time, this narrow down process can be repeated and the user and the robot can finally find a target object in high success rate. We experimentally confirmed that the average success rate is over 90%. We also investigated the relation among the various features set, success rate, number of times for detecting, and the class-variance ratio. We found that the statistical superiority on categorization has little association with the performance of narrow down process. This means that prior determinations of effective features set for categorization is difficult.

One of the solutions to this problem is that the robotic system interactively tunes the feature vector space during interaction with a person. So, one of our future projects for improving the LTIP is implementation of the online feature-tuning scheme the LTIP.

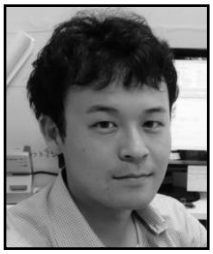
Acknowledgment

This research is partially supported by Grant-in-Aid for Scientific Research (C): 25330113 and Grant-in-Aid for Young Scientists (B): 22700202

References

[1] J. J. Lim, "Sketch Tokens: A Learned Mid-level Representation for Contour and Object Detection", Computer Vision and Pattern Recongnition, **(2013)**.
[2] T. Kawanishi, "Quick Object Search Using Color Histogrames and Pan-Tilz-Zoom Camera", The Institute of Electronics, Information and Communication Engineers, **(2001)**.
[3] H. Masuta, "Retinal Model based Plane Detection Method using Range Images for Unknown Object Recognition", The Robotics Society of Japan, **(2014)**.
[4] B. Akgun, "Trajectories and Keyframes for Kinesthetic Teaching: A Human-Robot Interaction Perspective", Session: Animating Robot Behavior, **(2012)**.
[5] H. Osawa and M. Imai, "Immersive Discovery Method for Exploring Interaction Strategies of an Agent" The Japanese Society for Artificial Intelligence, **(2013)**.
[6] K. Yanai, "Web Image Mining with Probabilistic Topic Models", The 22nd Annual Conference of the Japanese Society for Artificial Intelligence, **(2008)**.
[7] Y. Okada,"Genericimage clasithcation using SIFT and Higher rank SOM", IEICE, **(2009)**.

Authors



Junji Takahashi received his B.E., M.E., and Ph.D. degree in Engineering from Nagoya University, Nagoya, Japan, in 2005, 2007, and 2010 respectively. He was a Research Fellow on GCOE (Cybernics) at the Graduate School of Systems and Information Engineering in University of Tsukuba from April 2010 to March 2012, and a postdoctoral Research Fellow of Nagoya University from April 2012 to March 2013. He is currently an Assistant Professor of Aoyama Gakuin University. He is a member of IEEE, RSJ, SICE and JSME. He received a Best Paper Award at the 9th International Symposium on Distributed Autonomous Robotic Systems (DARS2008), a prize of encouragement from the Chubu Branch of The Society of Instrument and Control Engineers (SICE), 2012. His research interests include mechatronics, distributed autonomous robotic systems, SLAM, robotics for disaster response, factory automation, human robot interaction, and bio-signal processing.

Ryuji Suzuki received his B.E from Aoyama Gakuin University, Kanagawa, Japan, in 2016. Currently, he is a Master Degree Student at Graduate School of Science and Engineering, Aoyama Gakuin University. His main research interests include robotics and sensing.

Yoshito Tobe received his Ph.D. in Media and Governance from Keio University in 2000. He joined Aoyama Gakuin University as a professor in 2000. He joined Aoyama Gakuin University as a professor in 2012. His research includes wireless sensor networks, distributed systems, and participatory sensing systems.