Joint attention with strangers based on generalization through joint attention with caregivers

Akio Morita* Yuichiro Yoshikawa* Koh Hosoda*† Minoru Asada*†
*Dept. of Adaptive Machine Systems,
†Handai Frontier Research Center,
Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan
Email: {morita, yoshikawa}@er.ams.eng.osaka-u.ac.jp, {hosoda, asada}@ams.eng.osaka-u.ac.jp

***Abstract*— *Joint attention* is supposed to be a basis of the competence of communication with others. The authors have been attacking the issue how to learn joint attention only through interactions with caregivers, in other words, without external task evaluation from a viewpoint of a constructivist approach towards both establishing a design principle of communicative robots and understanding the developmental process of human communication. This paper presents a method for quick learning of joint attention with unfamiliar persons based on a hybrid architecture that consists of (1) a quick but person-dependent learning module and (2) a slow but person-independent learning module. Experimental results show the effectiveness of the proposed method.**

I. Introduction

A behavior of attending to an object to which another agent attends is defined as *joint attention* and is supposed to be a basis to estimate other's internal states such as intention and goal [1], and furthermore to acquire *the theory of mind* [2]. The conventional methods have fully preprogrammed the behavior of joint attention since they have focused on how higher social functions can be performed based on this behavior [3], [4], [5], [6]. However, the researchers have to reprogram the behavior when the environment changes too dramatically for the robot to acquire joint attention. Then it is a formidable issue to build a robot that learns joint attention through interaction with human beings since the competence of learning provides the robot with adaptability to the changes in the environment. On the other hand, although there are many studies on the mechanism of joint attention in human beings (e.g. [1], [7]), studies on understanding the acquisition process have just started. Therefore, building a robot that learns joint attention is an interesting issue from a viewpoint of a constructivist approach towards both establishing the design principle for an intelligent robot and understanding human developmental processes [8].

On the issue of acquiring joint attention, Nagai et al. [9] built a robot that learned joint attention based on the explicit evaluation from the caregiver. However, as a human infant does not seem to always need such an explicit evaluation from its caregiver to acquire joint attention. The robot should be able to learn it only through interactions with the caregivers, in other words, without supervision from them. Nagai et al. [10] and Fasel et al. [11] studied how the robot can acquire the competence of joint attention without explicit supervision. However, since their acquisition processes were tested only in the computer simulation and need enormous trials of joint attention for canceling out the influences of the failures, it was not clear whether their model worked in the real world. Furthermore, they did not explicitly consider how to extend the acquired behavior for jointing attention with unfamiliar persons (strangers) apart from the initial interaction partner (a caregiver).

In this paper, we extend the method in the previous work [10] for building a real robot that learns to perform joint attention with strangers based on generalization through joint attention with caregivers. The feedforward neural network (FNN) based architecture adopted to learn sensorimotor mapping for joint attention in the previous work could provide generalized mapping for persons. However, it needs too many trials to obtain the mapping through the real interactions since they involve many failures. Therefore, we propose a hybrid architecture that consists of a self-organized mapping (SOM) [12] based sensorimotor mapping and an FNN based one. In the SOM based sensorimotor mapping (SOM-SMM), the robot clusters the face image patterns of a caregiver and learns the mapping from the clusters onto the joint angle space to perform joint attention. Although the SOM-SMM highly depends on the caregiver since it clusters the face images based on the difference of the appearance, the learning can be done in real time since the mapping is much more simplified compared to the mapping from face image patterns onto the joint angle space. After learning some SOM-SMMs with caregivers, it generalizes these experiences through learning an FNN based sensorimotor mapping (FNN-SMM) by using only matched pairs of the face images and joint angles for joint attention, which are acquired in these SOM-SMMs.

In the rest of this paper, first we revisit the previous work [10] and point out how it is extended. Then, we present a hybrid architecture for learning joint attention with strangers based on generalization through the interac-

tions with caregivers. We show that the robot can quickly learn and generalize how to perform joint attention, and the effectiveness of the proposed method by the experiment.

II. Joint attention learning without explicit task evaluation [10]

Joint attention is a behavior of attending to an object to which another agent attends. More practically, the process to perform joint attention is defined as follows: first the robot faces with the caregiver who sits in front of the robot and observes the caregiver's face. Then, according to the obtained face image, it changes its focus of attention to an object that is regarded to be attended by the caregiver. Nagai et al. [10] have proposed a method to learn a sensorimotor mapping from the face image space onto the joint angle space to perform joint attention without explicit task evaluation. Since it seems reasonable to follow the method from a view point of a constructivist approach, we revisit the previous work and point out its limitation.

In the learning process of the previous work [10], the robot iterates observing the caregiver's face which indicates that the caregiver attends to something in the environment and servoing its camera head to fixate one of the salient objects in the environment (see Fig. 1). The mechanism to perform servoing the camera head is called *visual attention* and implemented by a visual feedback controller. Note that the salient object fixated by the robot is not always the same as the one to which the caregiver attends since there are plural objects in the environment. In other words, the experiences involve the cases where it fails to perform joint attention. The sensorimotor mapping for joint attention is implemented by a feedforward neural network where the face image pattern $\boldsymbol{x}$ is input and the difference of the posture of the camera head $\Delta\boldsymbol{\theta}$ is output. Nagai et al. showed that, to learn the sensorimotor mapping, it can use these $\boldsymbol{x}$ and $\Delta\boldsymbol{\theta}$ obtained through interaction as training data of backpropagation where $\Delta\boldsymbol{\theta}$ is a difference between the postures of the camera head in observing the caregiver's face and in observing the salient object instead of utilizing explicit task evaluation given by the caregiver. The architecture is schematically shown in Fig. 2.

Since a feedforward neural network learned by the back-propagation method is expected to exhibit a generalization function, the robot may perform person-independent joint attention by the acquired sensorimotor mapping. However, the learning process takes a lot of time since it needs enormous trials to cancel out the influences of the failed trials of joint attention on learning. As a result, the method in the previous work can not be directly applied to a robot that learns joint attention without any explicit task evaluation in the real world.

III. The hybrid architecture

To learn joint attention with strangers, the robot should obtain the experiences with plural caregivers for generalization in real time. Although the learning mechanism proposed in [10] has generalizing ability, it needs hundreds of thousand of trials. The need of enormous training data

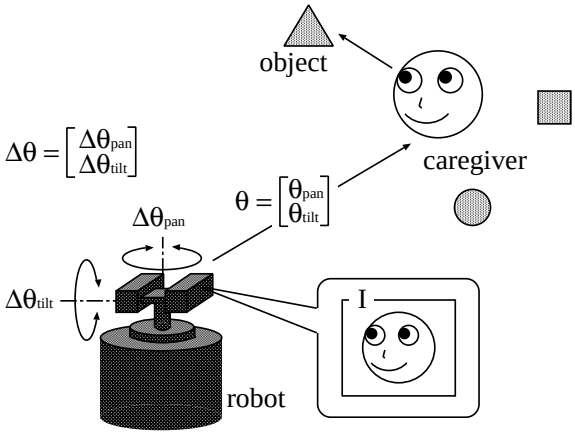


Fig. 1. Acquisition process of joint attention through the interaction with the caregiver

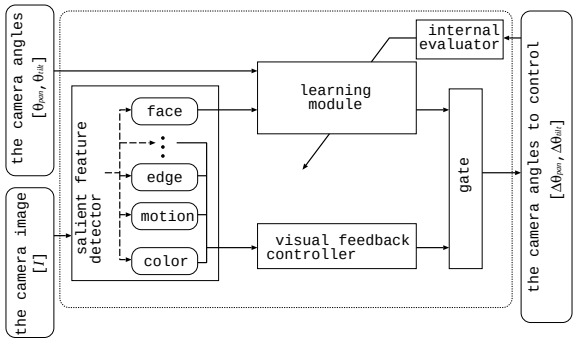


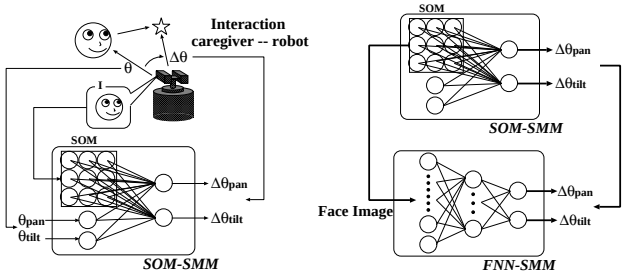
Fig. 2. A schematic architecture to learn joint attention with self-evaluation

is caused by the fact that the learning data contains the failed trials of which influence should be cancelled out. Therefore, we propose a hybrid architecture that consists of (1) quick but person dependent learning modules and (2) slow but person independent learning module.

The hybrid architecture consists of a self-organized mapping (SOM) based sensorimotor mapping and an FNN based one where the latter is the same as the learning module used in the previous work. Fig. 3 illustrates a schematic explanation of the proposed architecture. Both of the components learn the sensorimotor mappings from the face image space onto the joint angle space to perform joint attention but the training data used in the learning process are different from each other.

A. Learning an SOM-SMM

The robot first learns with an SOM based sensorimotor mapping (SOM-SMM) through interaction with a caregiver. SOM [12] is one of the clustering methods that can self-organize the clusters to construct a vector quantizer of the input data with less mean square error to decode [13]. We suppose that the robot can detect the face image pattern of the caregiver using the ordinary template matching method. Before learning joint attention, the robot constructs an SOM of the face image patterns of the caregiver who sits



(a) learning SOM-SMM (b) learning FNN-SMM

Fig. 3. The hybrid architecture

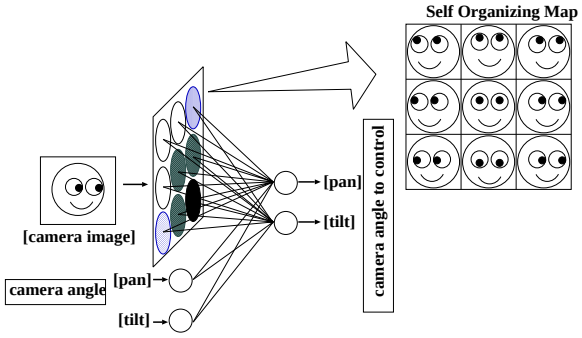


Fig. 4. The sensorimotor mapping based on two-layered perceptron with SOM input layer

in front of the robot and keeps changing his/her attention from an object to another.

Since the robot keeps observing the caregiver's face in this phase, it successively obtains the face image patterns of the caregiver $\boldsymbol{x} \in \Re^{D_I}$ in each of which the caregiver attends to something in the environment. According to the SOM algorithm, it clusters $\boldsymbol{x}$ in the following processes:

1) Codebook vectors $\boldsymbol{m}_i, i = 1, \cdots, N$ and the time stamp t are initialized.
2) When the robot perceives $\boldsymbol{x}$ which is the observed a face image pattern of the caregiver, the winner ID i^* of the codebook vectors is determined by the similarity to the input such as

$$i^* = \arg_i \min |\boldsymbol{x}(t) - \boldsymbol{m}_i(t)|, \tag{1}$$

where $|\cdot|$ indicates the Euclidean distance.
3) Codebook vectors are updated by

$$\begin{aligned} \boldsymbol{m}_i(t+1) &= \boldsymbol{m}_i(t) + \frac{1}{t}\Phi\left(|\boldsymbol{r}_i^* - \boldsymbol{r}_i|\right) \\ &\quad \cdot(\boldsymbol{x}(t) - \boldsymbol{m}_i(t)), \end{aligned} \tag{2}$$

where $\boldsymbol{r}_i$ is a position vector of the i-th codebook vector on the SOM. $\Phi(d)$ is a neighbourhood function such as

$$\Phi(d) = \exp\left(-\frac{d^2}{2\sigma^2(t)}\right), \tag{3}$$

$\sigma(t)$ is a time dependent decreasing parameter that determines to what extent neighbor codebook vector with the winner should be updated.
4) t is incremented and the process is returned to 2) until t is over the predefined iteration times.

The acquired SOM in these processes is expected to have sufficient variations of codebook vectors to represent the caregiver's face orientations since we suppose that the caregiver uniformly looks at objects in various position.

In the interaction phase with a caregiver, the robot learns a sensorimotor mapping for joint attention with a two-layered perceptron which outputs the joint angle for the input of the clustering result by the SOM (see Fig. 4). First, the robot observes the caregiver's face $\boldsymbol{x}$. The activations of the nodes of the SOM are calculated based on the similarity between the input and the codebook vector such as

$$a_i = \exp\left(-\frac{d_i^2}{{\sigma_s}^2}\right), \quad (i = 1, \cdots, N), \tag{4}$$

where σ_s is the parameters that determines how to evaluate the similarity. The vector $a = [a_1, \cdots, a_N]^T$ is input to the sensorimotor mapping. The weighted activations by the connection weights are summed up to calculate the output joint angle. This two-layered perceptron with an SOM input layer is learned through the backpropagation of the self-evaluation as in the previous work [10]. The proposed mechanism is expected to quick learn since the mapping is much more simplified compared to the direct mapping from face image patterns onto the joint angle space. However, note that the acquired sensorimotor mapping would be person dependent since the SOM input layer is specialized to one caregiver.

B. Learning an FNN-SMM

To perform joint attention with strangers, the robot must generalize its experiences of joint attention with caregivers since the acquired SOM-SMMs are person dependent. Although the previous work with an FNN needs enormous learning trials to cancel out the influence of the failed trials on learning the FNN, now the robot can utilize the acquired SOM-SMMs to obtain the training data instead of learning through interaction.

After learning an SOM-SMM, it is expected to output the correct joint angle to perform joint attention even when the robot obtains the codebook vectors of the SOM as an example of the caregiver's face image patterns. In other words, the robot can obtain N correct pairs of the face image patterns $\boldsymbol{m}_i$, that is a codebook vector of the SOM, and the joint angle to joint attention $\boldsymbol{\theta}_i^{som}(i = 1, \cdots, N)$, that is an output of the sensorimotor mapping for $\boldsymbol{m}_i$. Therefore, after the learning phase of SOM-SMMs with N_p caregivers, the robot obtains $N \times N_p$ data generated by N_p SOM-SMMs each of which is learned with a single caregiver. By using $N \times N_p$ data of the correct experience as the training data of learning an FNN, the robot can acquire the sensorimotor mapping with more generalized

competences. Note that although it takes numerous number of iterations to learn the sensorimotor mapping, the learning process can be performed in real time since it does not take a time to obtain the training data through interaction and the training data are expected to contain only the correct experiences to joint attention.

IV. Experiment

To show the effectiveness of the proposed hybrid architecture, we tested whether a real robot was able to learn joint attention with a stranger. First, we show the experiment of an SOM-SMM with one caregiver, and examine the performance of joint attention with a stranger. Then, the robot learns FNN-SMM with two SOM-SMMs and performs of joint attention with a stranger.

A. Experimental setup

The experimental setup is shown in Fig. 5. There are two yellow objects on a table between the robot and a person. Note that the positions of the objects change in every trial to joint attention. The person attends to one of these objects. On the other hand, the robot iteratively observes the person and then attends to one of the objects. When it observes the person's face, it obtains the face image pattern (30×30 pixels) that is found by the template matching method. To attend to the object, it extracts yellow color areas and controls its camera head to fixate a center of the regions by a visual feedback control.

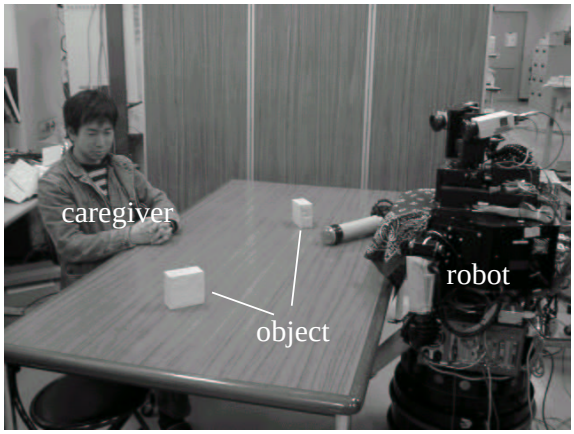


Fig. 5. The experimental setup

B. Learning an SOM-SMM

First of all, the robot kept observing the face of a caregiver who attended to one of the objects on the table. Since we supposed that the caregiver kept changing his/her gaze direction in this learning phase, the robot obtained the face image patterns in which the caregiver observed the object in various directions. Their face image patterns were clustered by an SOM algorithm. The input vectors consisted of the luminance values of the caregiver's face image (30×30 dimensions). The robot constructed an SOM by which the input vectors were mapped onto 8×8 clusters. Fig. 6 shows the face image patterns acquired as codebook vectors in the SOM. The SOM was self-organized through 2,000 times clustering of the observed caregiver's face image, which took five minutes.

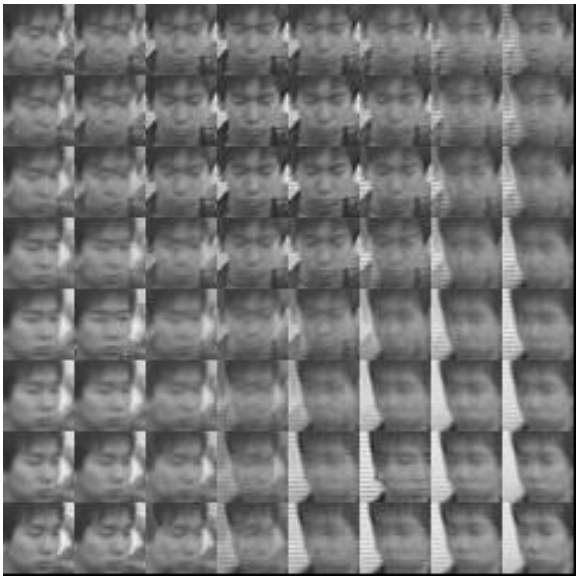


Fig. 6. Obtained codebook vectors of the self-organizing map

Next, to examine the learning process of an SOM-SMM through interaction with one caregiver, we conducted some computer simulation aided experiments. For each trial in the experiment, the robot obtained the activation pattern of the acquired SOM in observing the caregiver's face as an input, and the camera angle to fixate to one of two objects which are randomly located on the table as an output. We emulated this situation in the computer simulation, however we experimented with the real robot in order to make it more realistic. We obtained 300 data sets in advance, each of which consisted of the person's face image and camera angle to fixate the object to which the caregiver attended. The SOM-SMM learned a mapping from the activation pattern of the SOM to the motor command to fixate it through 800 trials to joint attention.

Fig. 7 shows the average and the standard deviation of the success rate of joint attention in 20 learning experiments each of which consists 800 trials. The success rate gradually increases from chance level at the beginning of the learning process to high performance at the end. We can see that the robot succeeds in quick learning of joint attention which is different from the previous work in which the robot needs ten thousands of trials. It seems to be caused by the fact that the sensorimotor mapping for joint attention is much more simplified by using SOM.

After the learning process, we investigated the performance of joint attention with the acquired SOM-SMM. We put only one object on each 8×4 points of the table and a person attended to it (see Fig. 8). In Fig. 8, x axis was a lateral position on the table, and y axis was a sagittal one for a person (caregiver or stranger). The robot observed the face image of a caregiver or a stranger and controlled its camera head to find the object by using acquired SOM.

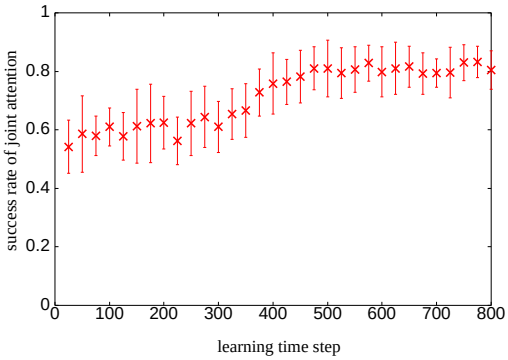


Fig. 7. The time course of the performance of joint attention in the learning process with the first caregiver

We counted the success number of joint attention out of five trials for each position.

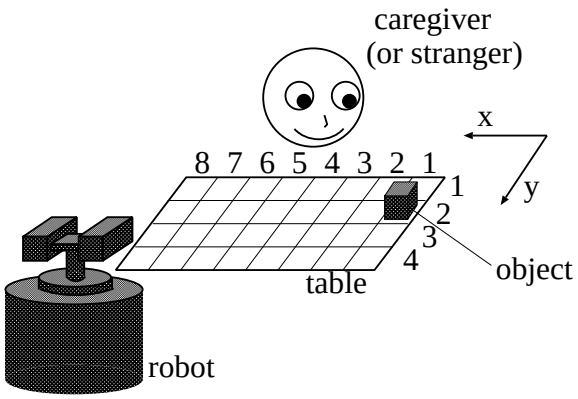


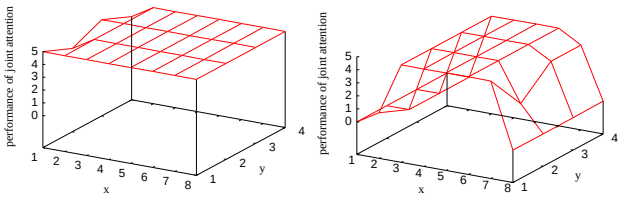
Fig. 8. The setup of evaluation of the task performance

Figs. 9 (a) and (b) show the distributions of the performance of joint attention with a caregiver and a stranger, respectively. The robot succeeded in joint attention with a caregiver independent of the location of the object. However, it failed with a stranger when the object was located on the edge of the table in this case $x = 1, 2, 7$, and 8 (see Fig. 9 (b)). Therefore, the robot seems to succeed in learning the sensorimotor mapping for joint attention with the caregiver while it failed in acquiring with a stranger. In other words, the acquired SOM-SMM was specialized for the caregiver.

C. Learning an FNN-SMM

To perform joint attention with a stranger, the robot must generalize the experiences of joint attention with caregivers. First, an FNN-SMM was learned by using an SOM-SMM with one caregiver, and we examined the performance of joint attention with a stranger. Then, the robot learned FNN-SMM by using the SOM-SMMs with two caregivers, and we evaluated whether the robot was able to perform joint attention.

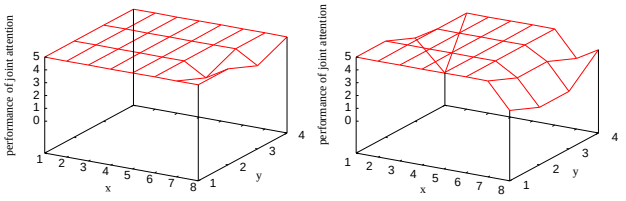
1) Learning with the first caregiver (64 data sets): We examined the performance of an FNN-SMM by using an SOM-SMM with one caregiver. The robot acquired the



(a) with the first caregiver (b) with a stranger

Fig. 9. The distribution of the performance that is the number of times (out of five) in which joint attention is successful, using the acquired SOM based sonsorimotor mapping with the first caregiver: (a) test with the first caregiver and (b) with a stranger

codebook vectors of the caregiver's face image pattern from the SOM as the input data of learning, and the motor command to attend to the object from an SOM-SMM as the learning output data. An FNN-SMM was trained by using 64 learning data sets which corresponded to the codebook vectors of the SOM and the output of the SOM-SMM. After the robot learned with these data sets for total 300,000 times, we examined its performance of joint attention with a caregiver and with a stranger. We counted the success number of joint attention in the same way of the previous experiment (see Fig. 8). Fig. 10 shows the distributions of the performance of joint attention with the first caregiver and a stranger, respectively. The robot succeeded in joint attention with a stranger even where the object was located on the edge of the table though the performance at those positions was not perfect in SOM-SMM. Therefore, the acquired FNN-SMM seemed to be more person independent than an SOM-SMM.

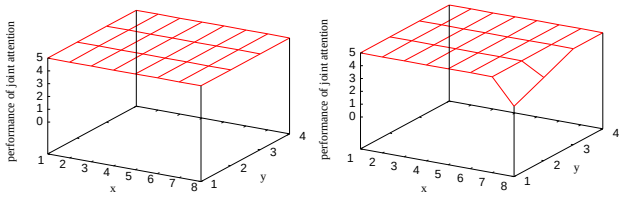


(a) with the first caregiver (b) with a stranger

Fig. 10. The distribution of the performance that is the number of times (out of five) in which joint attention is successful, using the acquired FNN based sonsorimotor mapping after learning with the first caregiver: (a) test with the first caregiver and (b) with a stranger

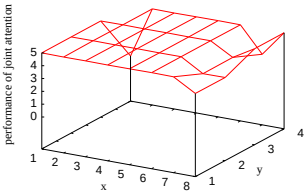
2) Learning with two caregivers (128 data sets): Then, we examined the performance of an FNN-SMM utilizing acquired two SOM-SMMs with two caregivers. The robot separately learned an SOM-SMM with each caregiver, and obtained 128 learning data sets. An FNN-SMM learned by using the 128 data sets. After the robot learned with these data sets for total 300,000 times, we evaluated the task performance of the FNN-SMM in joint attention with the first caregiver, with the second caregiver, and with a

stranger, respectively in the same way above (see Fig. 11). We can see that the robot succeeded in performing joint attention with the first and the second caregiver to the objects at any position on the table by using the FNN-SMM (see Figs. 11 (a) and (b)). Furthermore, it also succeeded in performing joint attention with a stranger even where the object was located on the edge of the table (see Fig. 11 (c)). We can see that the performance at the edge was slightly improved, compared to the one by the FNN-SMM with one caregiver (see Fig. 10). We may conclude that an FNN-SMM learned with plural caregivers is able to generalize persons' faces and performs person-independent joint attention.



(a) with the first caregiver

(b) with the second caregiver



(c) with a stranger

Fig. 11. The distribution of the performance that is the number of times (out of five) in which joint attention is successful, using the acquired FNN based sonsorimotor mapping after learning with the first caregiver: (a) test with the first caregiver, (b) with the second one, and (c) with a stranger

V. Conclusion

We proposed a hybrid architecture to quickly learn to perform joint attention with unfamiliar person based on generalization through joint attention with a familiar person in the real world. It consists of an SOM based sensorimotor mapping (SOM-SMM) and an FNN based one (FNN-SMM). The FNN-SMM learns generalized competence of joint attention by utilizing the examples acquired in the SOM-SMM by the real time interaction with caregivers. In computer simulation aided experiments, we confirm that the robot can fast learn joint attention in the real world by the proposed architecture and the possibility to acquire generalized competence of joint attention.

Although the robot independently learns the SOM-SMM with caregivers in the proposed method, the robot do not have to wait learning the FNN-SMM until it finished all learning process of the SOM-SMM with caregivers. Since the FNN-SMM learned by utilizing the already acquired SOM-SMMs could work so that the robot succeeds in joint attention in the learning phase of the SOM-SMM with new caregiver, how to balance the use of developing the FNN-SMM and one of visual attention in the trial of joint attention is one of our future work. Furthermore, the robot should be able to robustly perform joint attention in various environment both with the caregivers and strangers. On this purpose, although the SOM is fixed after once acquired in this paper, how to adapt the SOMs for the environmental changes involving the caregivers and how to modify the FNN-SMM according to the adaptation of the SOMs should be also addressed in the future.

Acknowledgment

The Advanced and Innovational Research program in Life Sciences of the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government and a Research Fellowship for Young Scientists from Japan Society for the Promotion of Science supported this research.

References

[1] C. Moore and P. J. Dunham, "Joint attention: It's origins and role in development," *Lawrence Erlbaum Associates*, 1995.
[2] Baron-Cohen, *Mindblindness*. MIT Press, 1995.
[3] B. Scassellati, "Theory of mind for a humanoid robot." *Autonomous Robots*, vol. 8, no. 1, pp. 13–24, 2002.
[4] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999, pp. 1146–1151.
[5] H. Kozima and H. Yano, "A robot that learns to communicate with human caregivers," in *Proceedings of the first International Workshop on Epigenetic Robotics*, 2001.
[6] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction." in *Proceedings of 10th IEEE International Workshop on Robot and Human Communication*, 2001, pp. 498–503.
[7] G. Butterworth, "The ontogency and phylogency of joint visual attention," *Blackwell*, 1991.
[8] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots." *Robotics and Autonomous System*, vol. 37, pp. 185–193, 2001.
[9] Y. Nagai, M. Asada, and K. Hosoda, "Developmental learning model for joint attention," in *Proceedings of the IEEE/RSJ International Conference on Interlligent Robots and Systems*, 2002, pp. 932–937.
[10] Y. Nagai, K. Hosoda, and M. Asada, "Joint attention emerges through bootstrap learning," in *Proceedings of the IEEE/RSJ International Conference on Interlligent Robots and Systems*, 2003, pp. 168–173.
[11] I. Fasel, G. O. Deák, J. Triesch, and J. Movellan, "Combining embodied models and empirical research for understanding the development of shared attention," in *Proceedings of the 2nd International Conference on Development and Learning*, 2002, pp. 21–27.
[12] T. Kohonen, *Self-organizing maps*. Heidelberg: Springer, 1995.
[13] S. Luttrell, "Derivation of a class of training algorithm," *IEEE Trans. Neural Networks*, vol. 1, pp. 229–232, 1990.