Joint Attention Emerges through Bootstrap Learning

Yukie Nagai* 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
e-mail: yukie@er.ams.eng.osaka-u.ac.jp, {hosoda, asada}@ams.eng.osaka-u.ac.jp

***Abstract*— A human-like intelligent robot is expected to have the capability to develop its cognitive functions through experience without a priori knowledge or explicit teaching. In addition, the realization of this kind of robot can lead us to understand the developmental mechanisms of human beings. This paper proposes a bootstrap learning model by which a robot can acquire the ability of joint attention without a caregiver's evaluation or a controlled environment based on the robot's embedded mechanisms: visual attention and learning with self-evaluation. Through learning based on the proposed model, the robot finds a correlation in sensorimotor coordination when joint attention succeeds and consequently acquires the ability of joint attention by accumulating the appropriate correlation and losing the uncorrelated coordination as statistical outliers. The experimental results show the validity of the proposed model.**

I. INTRODUCTION

It is a challenging problem to develop a robot that can acquire a cognitive function through interactions with the environment without *a priori* knowledge or explicit teaching. The realization of this kind of robot reduces the designer's burden and could lead us to develop a human-like artificial agent and to understand the developmental mechanisms of human beings through it [1].

We have focused on joint attention, which is one of the social cognitive functions, and developed a learning model by which a robot acquires the ability of joint attention through interactions with a human caregiver [2]. For our purposes, joint attention is defined as the process by which an agent attends to an object that another agent attends to [3]. In human beings, joint attention is considered to be a cornerstone for social communication and enables an infant to interact with the caregiver and to receive various kinds of knowledge from the caregiver [4], [5].

In robotics studies [6]–[9], joint attention has been upheld as a significant function for a social robot to realize interactions with humans. Note that these studies have a common problem that a robot's ability of joint attention is considered to be innate. In cognitive science, it is suggested that a human infant acquires the ability of joint attention through interactions with its environment without explicit teaching [4]. Therefore, the above robotics studies could not explain the developmental mechanisms of the human infant. On the other hand, Fasel *et al.* [10] proposed an idea how an infant could develop the ability of joint attention. However, the validity of their idea has not been shown by implementing it in an artificial agent. Nagai *et al.* [2] proposed a learning model for joint attention and showed that the model enables a robot to acquire the ability by receiving task evaluation from a human caregiver. The caregiver plays an important role in the robot's development just as a caregiver would in an infant's development. However, it is very interesting to argue how the robot or the infant can acquire higher cognitive functions based on its embedded or pre-learned capabilities without the caregiver's intervention.

This paper presents a learning model which enables a robot to acquire the ability of joint attention based on its embedded capabilities without a caregiver's task evaluation or a controlled environment. In this paper, independent learning without teaching, external evaluation, or a controlled environment is called *bootstrap learning*. The proposed bootstrap learning model consists of two embedded mechanisms of the robot. One is visual attention to find and attend to a salient object in the robot's view, and the other is learning with self-evaluation to evaluate the success of visual attention and then to learn a sensorimotor coordination. Through trials and learning based on the above mechanisms, the robot acquires the correlation of the sensorimotor coordination when joint attention succeeds while it cannot find the correlation when joint attention fails. In the latter situation, the uncorrelated coordination is expected to be lost as statistical outliers since the position of the object that the robot attends to changes randomly every trial. As a result, only the appropriate correlation survives in the learning module and consequently allows the robot to acquire the ability of joint attention.

In the rest of this paper, the proposed bootstrap learning model is first explained. Next, some experiments which show that the robot can acquire the ability of joint attention based on the proposed model without a controlled environment or external task evaluation are described. Finally, conclusions and future work are given.

II. EMERGENCE OF JOINT ATTENTION THROUGH BOOTSTRAP LEARNING

A. Definition of Joint Attention

Fig. 1 shows an experimental setup for joint attention, in which a robot with two cameras, a human caregiver, and multiple salient objects are indicated. In each trial, the objects are placed at random positions, and the caregiver attends to a different object. In Fig. 1, the caregiver attends to the square object. The robot can receive a camera image $\boldsymbol{I}$ and camera angles $\boldsymbol{\theta} = [\theta_{pan},\ \theta_{tilt}]$ as inputs, and output a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ for the camera head to rotate. The joint attention task in this situation is defined as the process by which the robot outputs a motor command $\boldsymbol{\Delta\theta}$ based on the sensor inputs $\boldsymbol{I}$ and $\boldsymbol{\theta}$, and consequently attends to the same object that the caregiver attends to. Note that $\boldsymbol{\Delta\theta}$ is incrementally generated to control the camera head because of two kinds of nonlinearity: the rotational center of the camera head does not coincide with the optical center of each camera, and it is impossible to determine which point along the caregiver's gaze is the focus of the caregiver's attention.

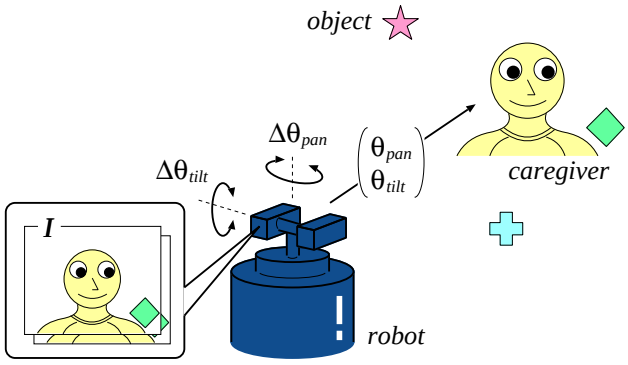


Fig. 1. An environmental setup for joint attention

B. A Basic Idea

The robot acquires the ability of joint attention through bootstrap learning based on the following embedded capabilities:

(a) *visual attention:* to find and attend to a salient object in the robot's view, and
(b) *learning with self-evaluation:* to evaluate the success of visual attention and then to learn a sensorimotor coordination.

First, the robot attends to the caregiver because it is the most interesting feature for the robot[1]. Then, if the robot finds a salient object in its view, the robot shifts its gaze direction from the caregiver's face to the object based on the visual attention mechanism. When visual attention succeeds, the robot evaluates it by itself and learns the sensorimotor coordination between the inputs $\boldsymbol{I}$, $\boldsymbol{\theta}$ and the output $\boldsymbol{\Delta\theta}$ based on the mechanism of learning with self-evaluation.

Note that visual attention is not always joint attention. The reason is that there are some salient objects in the environment, and the object that the robot attends to based on the mechanism of visual attention is just an interesting one for the robot, but does not always correspond to the object that the caregiver attends to. Therefore, the sensorimotor coordination that the robot learns in each trial can be either correct or incorrect for joint attention. The correct learning data are acquired when joint attention succeeds, and the incorrect are acquired when joint attention fails while visual attention succeeds. Through the learning process, however, the robot loses the incorrect data as outliers because it is supposed that the object position that the robot attends to changes randomly every trial and the sensorimotor coordination does not have any correlation. As a result, the robot acquires the appropriate sensorimotor correlation only when joint attention succeeds. In addition, the robot is expected to increase the success rate of joint attention over chance by utilizing the acquired sensorimotor coordination instead of the embedded mechanism of visual attention in subsequent trials. The robot consequently can find a better correlated coordination and acquire the ability of joint attention through bootstrap learning without a controlled environment or external task evaluation.

III. BOOTSTRAP LEARNING MODEL FOR JOINT ATTENTION

The proposed bootstrap learning model, which is based on visual attention and learning with self-evaluation, is shown in Fig. 2. As described above, the inputs to the model are $\boldsymbol{I}$ and $\boldsymbol{\theta}$, and the output is $\boldsymbol{\Delta\theta}$. The proposed model consists of the following modules, each of which corresponds to the embedded mechanisms: (a) visual attention and (b) learning with self-evaluation.

(a-1) The *salient feature detector* extracts distinguishing image areas from $\boldsymbol{I}$.
(a-2) The *visual feedback controller* receives the detected image features about objects and outputs ${}^{VF}\boldsymbol{\Delta\theta}$ to attend to an interesting object.
(b-1) The *internal evaluator* drives the learning mechanism in the learning module when the robot can attend to the interesting object.
(b-2) The *learning module* receives the image of the caregiver's face and $\boldsymbol{\theta}$ as inputs and outputs ${}^{LM}\boldsymbol{\Delta\theta}$. This module learns the sensorimotor coordination when it is triggered by the internal evaluator.

The salient feature detector and the visual feedback controller act as the visual attention mechanism, and the internal evaluator and the learning module carry out the

[1]The preference for looking at the caregiver is also innate for a human infant.

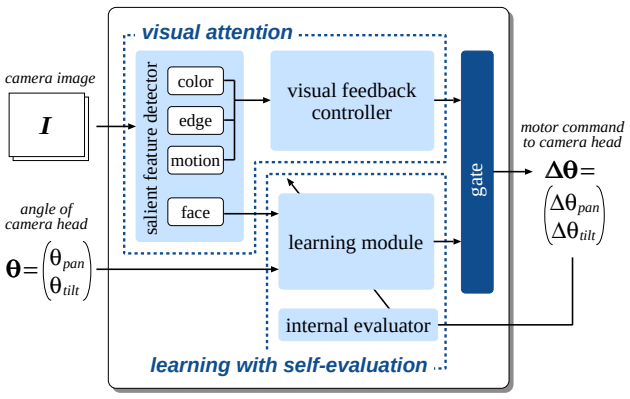


Fig. 2. Bootstrap learning model for joint attention based on visual attention and learning with self-evaluation

mechanism of learning with self-evaluation. In addition to these modules, the bootstrap learning model has another module to arbitrate the output of the robot.

(c) The *gate* makes a choice between ${}^{VF}\boldsymbol{\Delta\theta}$ and ${}^{LM}\boldsymbol{\Delta\theta}$, and outputs $\boldsymbol{\Delta\theta}$ as the robot's motor command.

The following sections explain these modules in detail.

A. Salient Feature Detector

The salient feature detector extracts distinguishing image areas in $\boldsymbol{I}$ by color, edge, motion, and face detectors. The color, edge, and motion detectors extract objects ($i = 1, \ldots, n$) which have bright colors, complicated textures, and motions, respectively. Then, the salient feature detector selects the most interesting object i_{trg} among the extracted objects by comparing the sum of the interests of all features.

$$i_{trg} = \arg\max_i (\alpha_1 f_i^{col} + \alpha_2 f_i^{edg} + \alpha_3 f_i^{mot}), \quad (1)$$

where f_i^{col}, f_i^{edg}, and f_i^{mot} indicate the size of the bright color area, the complexity of the texture, and the amount of the motion, respectively. The coefficients $(\alpha_1,\ \alpha_2,\ \alpha_3)$ denote how interesting each feature is and are determined according to the robot's characteristics and the context. At the same time, the face detector extracts the face-like stimuli of the caregiver. The detection of face-like stimuli is a fundamental ability for a social agent; therefore, it should be treated in the same manner as the detection of the primitive features. The detected primitive feature of the object i_{trg} and the face-like one of the caregiver are sent to the visual feedback controller and the learning module, respectively.

B. Visual Feedback Controller

The visual feedback controller receives the detected image feature of the object i_{trg} and outputs a motor command ${}^{VF}\boldsymbol{\Delta\theta}$ for the camera head to attend to i_{trg}. First, the controller calculates the object position $(x_i,\ y_i)$ of i_{trg} in the camera image. Then, the motor command ${}^{VF}\boldsymbol{\Delta\theta}$ is generated as

$${}^{VF}\boldsymbol{\Delta\theta} = \begin{pmatrix} {}^{VF}\Delta\theta_{pan} \\ {}^{VF}\Delta\theta_{tilt} \end{pmatrix} = g \begin{pmatrix} x_i - cx \\ y_i - cy \end{pmatrix}, \quad (2)$$

where g and $(cx,\ cy)$ denote a scalar gain and the center position of the image, respectively. The motor command ${}^{VF}\boldsymbol{\Delta\theta}$ is sent to the gate as the output of the visual feedback controller.

As described above, visual attention, which is one of the robot's embedded mechanisms, is performed by the salient feature detector and the visual feedback controller.

C. Internal Evaluator

The other embedded mechanism, that is learning with self-evaluation, is realized by the internal evaluator and the learning module.

The internal evaluator drives the learning mechanism in the learning module when

$$\sqrt{(x_i - cx)^2 + (y_i - cy)^2} < d_{th}, \quad (3)$$

where d_{th} is a threshold for evaluating whether the robot in looking at an object in the center of the camera image or not. Note that the internal evaluator does not know whether joint attention has succeeded but knows whether visual attention has succeeded.

D. Learning Module

The learning module consists of a three-layered neural network. In the forward processing, this module receives the image of the caregiver's face and the angle of the camera head $\boldsymbol{\theta}$ as inputs, and outputs ${}^{LM}\boldsymbol{\Delta\theta}$ as a motor command. The caregiver's face image is required to estimate the motor command ${}^{LM}\boldsymbol{\Delta\theta}$ to follow the caregiver's gaze direction, and the angle $\boldsymbol{\theta}$ is utilized to output ${}^{LM}\boldsymbol{\Delta\theta}$ incrementally and nonlinearly because the caregiver's attention cannot be narrowed down to a particular point along the line of the caregiver's gaze direction. The generated motor command ${}^{LM}\boldsymbol{\Delta\theta}$ is sent to the gate as the output of the learning module.

In the learning process, this module learns the sensorimotor coordination by back propagation when it is triggered by the internal evaluator. As described above, the internal evaluator drives the learning module according to the success of visual attention, not joint attention, this module has correct and incorrect learning data. The correct data mean joint attention has succeeded while the incorrect mean it has failed. In the case of correct data, the learning module can acquire the correlation between the inputs, the caregiver's face image and $\boldsymbol{\theta}$, and the output $\boldsymbol{\Delta\theta}$. On the other hand, in the case of incorrect data, this module cannot find the appropriate correlation; therefore,

such data is expected to be lost as outliers through the learning process. As a result, the acquired correlation of the sensorimotor coordination allows the robot to realize joint attention.

E. Gate

The gate arbitrates the motor command $\boldsymbol{\Delta\theta}$ between ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller and ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module. The gate sets a gating function to define the selecting rate of the outputs. At the beginning of the learning process, the selecting rate of ${}^{VF}\boldsymbol{\Delta\theta}$ is set to a high probability because the learning module has not acquired the appropriate sensorimotor coordination for joint attention yet. On the other hand, in the latter stage of the learning process, the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module, which has acquired the sensorimotor correlation, gradually comes to be selected at high probability. As a result, the robot can experience many learning situations which include both correct and incorrect data in the early stage of the learning process, and increase the correct ones according to the learning advance. It allows the robot to acquire more appropriate sensorimotor coordination for joint attention.

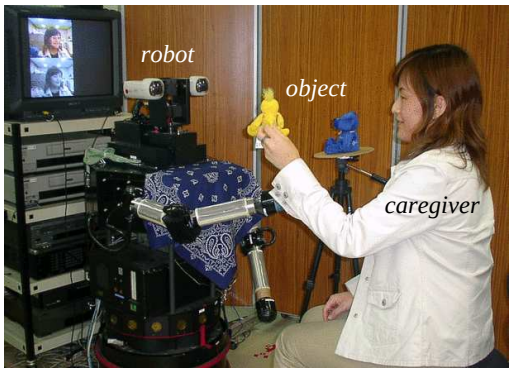
IV. EXPERIMENT

To show the validity of the proposed model, it was examined that an actual robot is able to acquire the ability of joint attention based on the proposed model in an uncontrolled environment in which multiple salient objects are placed.

A. Experimental Setup

An experimental environment is shown in Fig. 3 (a), and the left camera image of the robot is shown in (b). The caregiver sits in front of the robot and attends to the object in its hand. Other salient objects are set around the caregiver at random positions. The robot has two cameras and can turn them simultaneously to pan and tilt. The robot receives the camera image and detects the caregiver's face (left in Fig. 3 (b)) and the objects (right) by the salient feature detector. In the experiment, the degrees of the interests of the image features in Eq. (1) are set to $(\alpha_1,\ \alpha_2,\ \alpha_3) = (1,\ 0,\ 0)$, and the threshold of the success of visual attention in Eq. (3) is defined as $d_{th} = W_x/6$, where W_x is the width of the camera image.

To execute the learning in a simulated environment, the robot acquired 125 data sets, which included a camera image in which the caregiver's face was extracted as a window of 30 $\times$ 25 pixels and a camera angle when the robot attended to the caregiver, and a motor command for the camera head to shift its gaze direction from the caregiver to the object that the caregiver attended to in advance. Then, in each trial, we took one data set from the above and placed other salient objects at random positions in the simulated environment. The number of input, hidden, and output units of the learning module were set to 752 (30 $\times$ 25 for a camera image and 2 for the pan and the tilt angles of the camera head), 7, and 2, respectively. Under this condition, the robot repeated the trials and the learning alternately based on the proposed model.



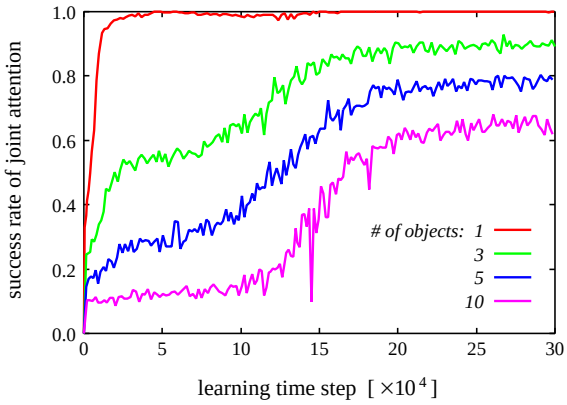
(a) an experimental environment

(b) the left camera image of the robot (left: the detected result of the caregiver's face by template matching, right: the detected result of the bright colors)

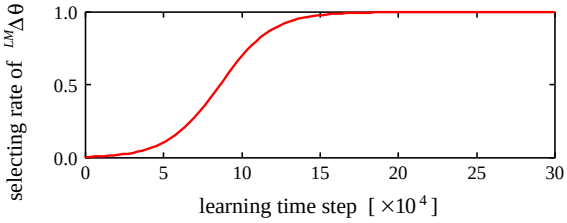
Fig. 3. An experimental setup for joint attention

B. Performance Change in Various Situations

It was verified that the proposed model enables the robot to acquire the ability of joint attention in an environment that includes multiple objects. The change of the success rate of joint attention in terms of the learning time is shown in Fig. 4 (a), where the number of the objects are set to 1, 3, 5, or 10. Fig. 4 (b) indicates the gating function (the selecting rate of ${}^{LM}\boldsymbol{\Delta\theta}$) as a sigmoid, which showed the best performance in some experiments. The number of objects 1 means that the robot has only correct learning situation every trial. By contrast, the number 10 means that the robot can experience the correct learning situation only at 1/10 probability at the beginning of the learning. However, it is expected to increase the correct one by utilizing the learning module, which has already acquired the correlated coordination until that time, according to



(a) success rate of joint attention



(b) gating function (sigmoid)

Fig. 4. The change of the success rate of joint attention and gating function (# of objects: 1, 3, 5, or 10, gate: sigmoid)

the advanced learning. From the result of Fig. 4 (a), we can see that the success rates of joint attention are at chance levels at the beginning of the learning process; however, they increase to high performance at the end even if many objects are set in the environment. Therefore, it is concluded that the robot can acquire the ability of joint attention based on the proposed bootstrap learning model without a controlled environment or external task evaluation.

Next, the effectiveness of the sigmoid gating function was verified. The result of Fig. 4 (a) (the number of the object: 5) was compared with the success rate of joint attention when the gating function was set to a constant value. The performance changes when the gating rate of ${}^{LM}\boldsymbol{\Delta\theta}$ is 0.7, 0.9, or 1.0 are shown in Fig. 5. The comparison of these results indicates that the gate designed as a sigmoid function can improve the task performance of joint attention. Especially, when the gating rate is 1.0, the success rate of join attention has not risen to the chance level, that is 0.2. The reason is that the learning module which had not acquired the appropriate correlation was utilized in the early stage of learning, and the learning data were biased to the initial experiences. These results

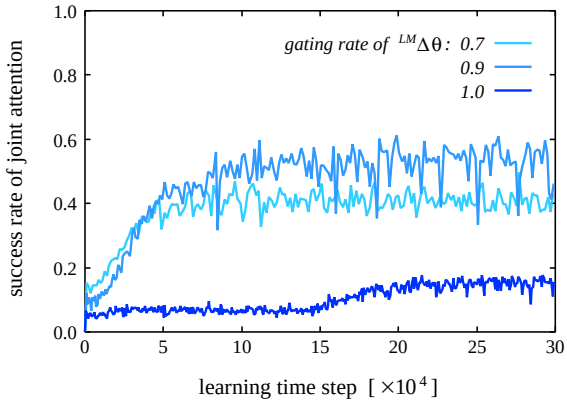


Fig. 5. The change of the success rate of joint attention (# of objects: 5, gate: constant)

show that the gating function should be designed so that the output from the visual feedback controller is selected with high probability in the beginning of learning and later that from the learning module comes to be adopted as learning advances, as described in Section III-E.

C. Final Performance of Joint Attention

After learning, we evaluated the final performance of the robot that used the sigmoid function for gating and learned in the environment with five objects. Fig. 6 shows the left camera images of the robot in which the input and the output of the learning module are indicated. In each of them, the caregiver's face image enclosed in a rectangle of 30×25 pixels is the input to the learning module, and the straight line shows the output from the learning module of which the width and the height indicate the pan and the tilt angles of the output, respectively. The robot is expected to find the object that the caregiver attends to by controlling the camera head along this line. From the results shown in Fig. 6, it is confirmed that the learning module can estimate the motor command to realize joint attention since the straight line corresponds to the gaze direction of the caregiver.

Fig. 7 shows the change of the robot's camera image when it shifts its gaze direction from the caregiver's face to the object based on the output from the learning module. The rectangle and the straight lines on the caregiver's face indicate the same meanings described above, and the circles and the cross lines show the gazing area of the robot and the object's position, respectively. The learning module incrementally generates a motor command at each step, and the robot consequently realizes the motion to follow the caregiver's gaze direction. During the camera motion, if the object is detected in the circle on the image, the robot stops its motion. This experimental result shows that the robot can realize joint attention based on the proposed model even if the object is far from the caregiver.

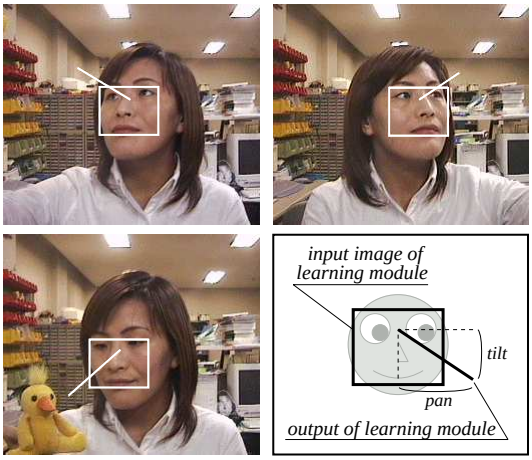


Fig. 6. The input and the output of the learning module when the caregiver is looking at various directions

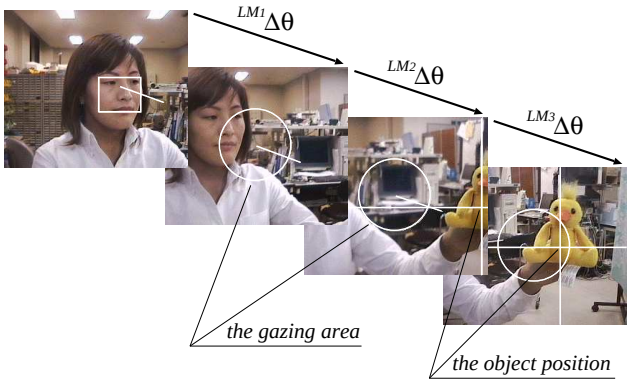


Fig. 7. The change of the camera image when the robot shifts its gaze direction from the caregiver's face to the object

V. CONCLUSION

The bootstrap learning model for joint attention has been presented in this paper. The model enables the robot to acquire the ability of joint attention by finding the sensorimotor correlation based on two embedded mechanisms: visual attention and learning with self-evaluation. Furthermore, the gate module in the proposed model makes learning more effective by utilizing the learning module which has already acquired the correlation. The experimental results showed that the robot can acquire the ability of joint attention based on the proposed model without a controlled environment or external task evaluation.

A more efficient learning mechanism should be developed so that the learning is executed not on the simulation but on the actual robot. In addition, the gating function should not be a deterministic one, like a sigmoid function, but designed by the performance of the robot. The realization of these changes will make the robot a truly developmental agent.

VI. ACKNOWLEDGMENTS

This study was funded by the Advanced and Innovational Research program in Life Sciences from the Ministry of Education, Culture, Sports, Science, and Technology of the Japanese Government.

VII. REFERENCES

[1] Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro, and Yasuo Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, Vol. 37, pp. 185–193, 2001.

[2] Yukie Nagai, Minoru Asada, and Koh Hosoda. Developmental learning model for joint attention. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 932–937, 2002.

[3] G. E. Butterworth and N. L. M. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, Vol. 9, pp. 55–72, 1991.

[4] Chris Moore and Philip J. Dunham, editors. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[5] Simon Baron-Cohen. *Mindblindness*. MIT Press, 1995.

[6] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, Vol. 8, No. 1, pp. 49–74, 2000.

[7] Brian Scassellati. Theory of mind for a humanoid robot. *Autonomous Robots*, Vol. 12, pp. 13–24, 2002.

[8] Hideki Kozima and Hiroyuki Yano. A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, 2001.

[9] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. In *Proceedings of 10th IEEE International Workshop on Robot and Human Communication*, 2001.

[10] Ian Fasel, Gedeon O. Deák, Jochen Triesch, and Javier 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*, pp. 21–27, 2002.