Proceedings of the 2nd International Symposium on
Adaptive Motion of Animals and Machines
SaA-II-3, March 2003

Emergence of Joint Attention based on Visual Attention and Self Learning

Yukie Nagai[1], Koh Hosoda[2], Akio Morita[3], and Minoru Asada[4]

Dept. of Adaptive Machine Systems, Graduate School of Engineering,

[2,4]HANDAI Frontier Research Center,

Osaka University, Suita, Osaka, 565-0871 Japan

e-mail: {[1]yukie, [3]morita}@er.ams.eng.osaka-u.ac.jp, {[2]hosoda, [4]asada}@ams.eng.osaka-u.ac.jp

Abstract

This paper presents a method which develops the robot's ability of joint attention with a human caregiver based on visual attention and self learning mechanisms. The visual attention is to find and attend to a salient object in the robot's view, and the self learning is to learn a sensorimotor coordination when the visual attention succeeds. Based on these mechanisms, the robot learns the sensorimotor coordination when the robot can watch the salient object by shifting its gaze direction from the caregiver's face to the object. Through the learning process, the robot can find the sensorimotor correlation only when joint attention with the caregiver is achieved. The correlation allows the robot to acquire the ability of joint attention. The experimental results show the validity of the proposed method in a simplified situation.

1. Introduction

The constructive approach to understand a process how a robot acquires cognitive functions through interactions with its environment attracts many researchers' interests. Asada *et al.* [1] proposed cognitive developmental robotics as a paradigm for the design of the humanoid robot that develops itself, and for the understanding of the cognitive developmental mechanisms of human beings. In the cognitive developmental robotics, not only the design of the robot but also that of the environment and the interactions between them should be argued.

The capabilities to interact with the environment include the ability of joint attention. Joint attention is defined as a process that an agent attends to an object which another attends to [2]. Researchers in the cognitive developmental science have interests in the ability of joint attention as the cornerstone of the development of cognitive functions. The reason is that it is known that a human infant acquires the ability of joint attention in the early of development and learns many kinds of social knowledges, e.g. language and mind reading [3], based on it [4]. On the basis of the insight, studies to build the mechanism of joint attention for a robot and to make the robot acquire the social knowledges have been attempted [5, 6, 7]. However, the mechanisms of joint attention are fully-developed by the designers, and it is not argued how the robot can acquire such an ability of joint attention through interactions with its environment. In contrast, Nagai *et al.* [8] proposed a learning system for joint attention with a human caregiver. They showed that their proposed system enables a robot to acquire the ability of joint attention through learning with the caregiver. In their system, however, the robot requires a task evaluation, that is success/failure information about joint attention, from the caregiver during the learning.

This paper presents a method which develops the robot's ability of joint attention based on visual attention and self learning mechanisms. The visual attention is to find and attend to a salient object in the robot's view, and the self learning is to learn a sensorimotor coordination when the visual attention succeeds. Based on these mechanisms, the robot learns the sensorimotor coordination when the robot can watch the salient object by shifting its gaze direction from the caregiver's face to the object. Through the learning process, the robot can find the sensorimotor correlation only when the robot achieves joint attention with the caregiver. The correlation allows the robot to acquire the ability of joint attention without the success/failure evaluation about joint attention. In addition, the proposed method could be useful to understand the cognitive developmental mechanisms of a human infant, because it is considered that the infant acquires the ability of joint attention not only based on the task evaluation but also based on the visual attention and the self learning mechanisms in the infant.

In the rest of this paper, the mechanism of the proposed method is described first. Next, the experimen-

tal results show that the robot can acquire the ability of joint attention based on the proposed method in a simplified situation. Finally, it is indicated that the learning process of the robot's joint attention based on the proposed method is similar to the developmental process of the infant's one.

2. Emergence of Joint Attention

An environmental setup for joint attention is shown in Figure 1. There are a robot with two cameras, a human caregiver, and multiple objects. The caregiver attends to one of the objects. The inputs to the robot are the camera images I_L, I_R and the angles of the camera head θ_{pan}, θ_{tilt}, and the outputs are the camera head displacement $\Delta\theta_{pan}$, $\Delta\theta_{tilt}$.

For the emergence of joint attention, the robot has the followings:

(a) *Visual attention*: to find and attend to a salient object in the robot's view.

(b) *Self learning*: to learn the sensorimotor coordination when the visual attention succeeds.

Based on these mechanisms, the robot can acquire the ability of joint attention along the following process. First, the robot attends to the caregiver and is supposed to find a salient object in its view. Then, the robot shifts its gaze direction to the salient object based on the visual attention mechanism. When the visual attention succeeds, the robot learns the sensorimotor coordination between the inputs (I_L, I_R, θ_{pan}, and θ_{tilt}) and the outputs ($\Delta\theta_{pan}$ and $\Delta\theta_{tilt}$) based on the self learning mechanism. Through the learning process, the robot can acquire the ability of joint attention by finding the sensorimotor correlation only when the robot achieves joint attention with the caregiver.

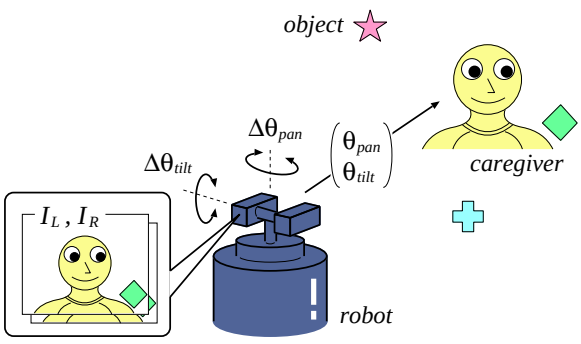


Figure 1: An environmental setup for joint attention

2.1. The proposed method with the visual attention and the self learning

The proposed method with the visual attention and the self learning is presented in Figure 2. The robot captures the camera images $I_{L,\,R}$ and the camera angles $\theta_{pan,\,tilt}$ as the inputs and outputs the displacement of the camera angles $\Delta\theta_{pan,\,tilt}$. The following modules corresponding to the visual attention and the self learning constitute the proposed method.

(a-1) *Salient feature detector* extracts distinguishing image areas from $I_{L,\,R}$.

(a-2) *Visual feedback controller* obtains the detected image features and outputs ${}^{VF}\Delta\theta$ to attend to the interesting object.

(b-1) *Learning module* receives the images of the caregiver's face and $\theta_{pan,\,tilt}$ as the inputs and outputs ${}^{LM}\Delta\theta$. This module learns the sensorimotor coordination when the internal evaluator triggers it.

(b-2) *Internal evaluator* drives the learning mechanism in the learning module when the robot can attend to the interesting object.

In addition to these modules, the proposed method has another one to adjust the output of the robot.

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

These modules are explained in detail in the following sections.

2.1.1. Salient feature detector

The salient feature detector extracts distinguishing image areas in $I_{L,\,R}$ by color, edge, motion, and face detectors. The color detector extracts objects with bright colors. The edge detector extracts image areas which have complicated textures. The motion detector finds objects with motions. These three mechanisms are used to detect the primitive features of the objects. On the other hand, the face detector extracts face-like stimuli of the caregiver. The detection of the face-like stimuli is a fundamental ability for a social agent and should be treated in the same manner with that of the primitive features. Then, the face detector keeps the caregiver's face image until it is detected again.

The detected primitive features of the objects and the face-like one of the caregiver are sent to the visual feedback controller and the learning module, respectively.

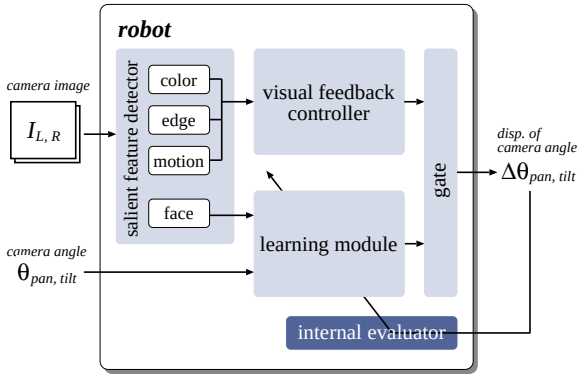


Figure 2: The proposed method for emergence of joint attention based on visual attention and self learning

2.1.2. *Visual feedback controller*

The visual feedback controller receives the detected image features by the salient feature detector and outputs the displacement ${}^{VF}\Delta\theta$ of the camera head to attend to the interesting object. The robot selects the most interesting object among the extracted image areas by summing the degrees of the interests of all features: color, edge, and motion. Then, the visual feedback controller calculates the image distance between the selected object's position and the center of the camera image and generates ${}^{VF}\Delta\theta$ by multiplying the distance with a gain. The robot can watch the most interesting object, if the gate selects ${}^{VF}\Delta\theta$ as the robot's motor command.

2.1.3. *Learning module*

The learning module consists of a three-layered neural network. It receives the images of the caregiver's face and the camera angles $\theta_{pan,\,tilt}$ of the robot and outputs ${}^{LM}\Delta\theta$ as the motor command. The caregiver's face image is required to estimate the robot's camera head displacement ${}^{LM}\Delta\theta$ to follow the caregiver's gaze.

At the same time, this module learns the sensorimotor coordination by back propagation when it is triggered by the internal evaluator. Through the learning process, the robot can find the correlation between the inputs, the caregiver's face images and $\theta_{pan,\,tilt}$, and the output ${}^{LM}\Delta\theta$ only when joint attention with the caregiver is realized successfully. The correlation allows the robot to acquire the ability of joint attention without a success/failure evaluation about joint attention.

2.1.4. *Internal evaluator*

The internal evaluator drives the back propagation of the learning module when the robot can attend to the interesting object. It means that the internal evaluator does not take account of success/failure of joint attention.

2.1.5. *Gate*

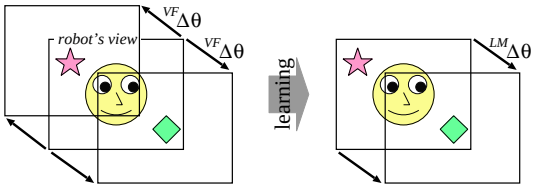
The gate decides the output $\Delta\theta_{pan,\,tilt}$ from ${}^{VF}\Delta\theta$ of the visual feedback controller or ${}^{LM}\Delta\theta$ of the learning module. The selecting method of the output is based on the success rate which the robot can attend to an interesting object by the output of the learning module. It means that ${}^{LM}\Delta\theta$ is selected at the rate which the robot successfully watches the interesting object by ${}^{LM}\Delta\theta$, and ${}^{VF}\Delta\theta$ is selected at the residual rate. By using this method, the robot can attend to the interesting object by ${}^{VF}\Delta\theta$ in the early stage of the learning. In the latter stage, the robot acquires the sensorimotor coordination to realize joint attention with the learning module and becomes to output ${}^{LM}\Delta\theta$ more frequently.

2.2. Incremental learning

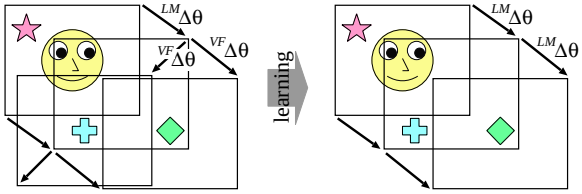
It is expected that the proposed method makes the robot generate an incremental learning process of joint attention.

stage I: In the beginning of the learning, the robot has a tendency to attend to the interesting object in its view by outputting ${}^{VF}\Delta\theta$, even though the caregiver attends to another object (left in Figure 3 (a)). Then, the robot starts to learn the sensorimotor coordination when the internal evaluator drives the learning module.

stage II: In the middle stage of the learning, the robot begins to utilize not only ${}^{VF}\Delta\theta$ but also ${}^{LM}\Delta\theta$. The robot can realize joint attention owing to the learning module learned in *stage I* only when the caregiver and the object which the caregiver attends to are observed in the same image (right in Figure 3 (a)). However, when the object is out of the robot's view, the robot can find it not at the center of the image but at the periphery by ${}^{LM}\Delta\theta$ and attend to the object by ${}^{VF}\Delta\theta$ (left in Figure 3 (b)). Then, the robot learns the sensorimotor coordination when it outputs ${}^{LM}\Delta\theta$ and ${}^{VF}\Delta\theta$, respectively.



(a) the learning from *stages I* to *II*



(b) the learning from *stages II* to *III*

Figure 3: Incremental learning of joint attention. The robot acquires the sensorimotor coordination which has the correlation between the caregiver's face image and ${}^{VF}\Delta\theta$.

stage III: In the last stage, the robot acquires the complete ability of joint attention owing to the learning in *stages I* and *II*. Even if the object which the caregiver attends to is far off the robot's first view, the robot can identify it by activating the learning module and generating ${}^{LM}\Delta\theta$ repeatedly until the object is detected in the robot's view (right in Figure 3 (b)).

It should be noted that the environment does not need to be controlled by the caregiver during the incremental learning process. The robot generates the incremental learning process by structuring the inputs from the environment and by learning the sensorimotor coordination step by step.

3. Experiment

3.1. Experimental setup

It was examined whether an actual robot can acquire the ability of joint attention based on the proposed method. An experimental environment is shown in Figure 4. The robot and the caregiver are set face-to-face, and the caregiver holds an object in its hand and attends to it. Then, the robot observes the caregiver with its two cameras and learns the sensorimotor coordination based on the proposed method.

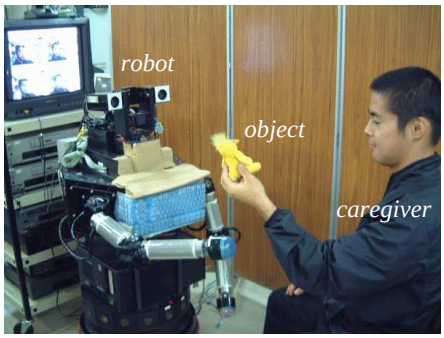


Figure 4: An experimental environment for joint attention

3.2. Incremental learning

It was verified that the robot was able to acquire the ability of joint attention when it was provided the incremental learning process explicitly. Each of the learning stages consists of a trial phase and a learning one. To simplify the problem, the environment is configured that single object with a color feature is placed and the caregiver is sure to attend to it. Under the condition, the robot learns the sensorimotor coordination along the following process.

stage I-trial: The object is set so that it can be observed in the same image that the robot observes the caregiver. Then, the robot tries to shift its gaze direction to the object by using the visual feedback controller.

stage I-learning: The robot learns the sensorimotor coordination of *stage I-trial* with the learning module.

stage II-trial: The object is set to be outside of the image that the robot observes the caregiver. Then, the robot tries to attend to the object based on the learning result in *stage I* and the visual feedback controller.

stage II-learning: The robot learns the sensorimotor coordination of *stages I-trial* and *II-trial* with the learning module which was already learned in *stage I*.

The learning curves through the all learning stages are shown in Figure 5. The curves plot the output error of the learning module through *stages I-learning* and *II-learning*. From this graph, we can see that the learning of joint attention progresses along the incremental learning process. The robot's final performance acquired through the all learning stages is presented in

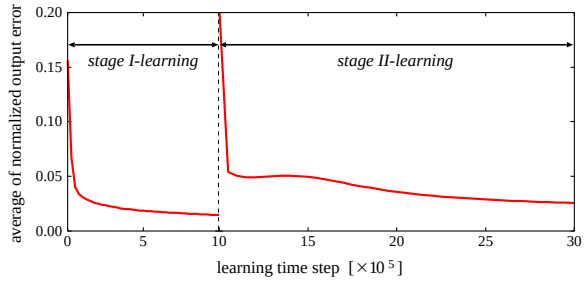


Figure 5: The error change through the incremental learning

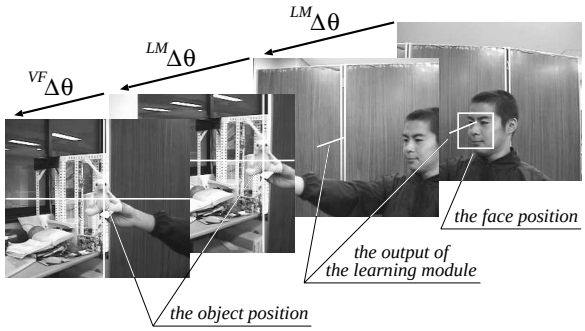


Figure 6: The final performance of the incremental learning

Figure 6. These images show the change of the left camera image of the robot by outputting ${}^{LM}\Delta\theta$ and ${}^{VF}\Delta\theta$. Here, the box, the lines, and the cross lines indicate the detected position of the caregiver's face, the output of the learning module, and the detected object's position, respectively. From this result, it can be concluded that the incremental learning based on the proposed method enables the robot to acquire the ability of joint attention in the situation with a single object.

3.3. Multiple objects situation

It was examined whether the robot was able to acquire the ability of joint attention based on the proposed method, even if multiple objects are set in the environment. Five objects with color features are placed to be observed in the same image that the robot attends to the caregiver. The positions of the objects are changed in each trial, and the caregiver also changes the object to attend to randomly. In the same manner as the previous experiment, the robot first obtains the learning data (the set of the sensory inputs and the motor outputs) through trials, and then learns the sensorimotor coordination of them. In the trials, the robot extracts the objects by the color detector and attends to one which has the largest size in the image by the visual feedback controller.

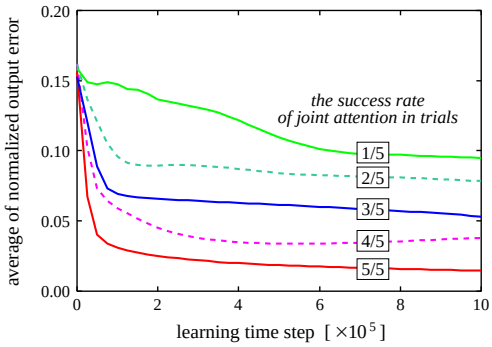


Figure 7: The error change in the multiple objects situation

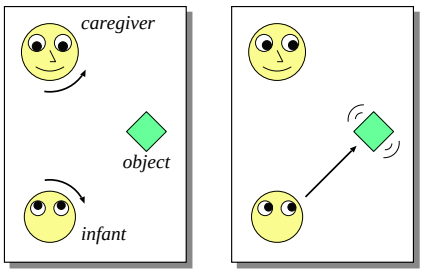
The change of the output error through the learning is shown in Figure 7. The learning curves indicate the error changes of the learning module in the cases that the robot was able to realize joint attention at the rates from $1/5$ to $5/5$ in the trials. Since the number of the objects is five, the rate $1/5$ means that the environment is not controlled so that the robot can realize joint attention in the trials. On the other hand, the rate $5/5$ means that the object which the caregiver attends to is controlled so that it can be detected as the largest size in the robot's view and the robot can realize joint attention in all trials. From this result, we can see that the final performance changes according to the success rate of joint attention in the trials, and the final performance at the small success rate is not sufficient to realize joint attention. About this problem, it is expected that the incremental learning can improve the final performance. The reason is that the robot can acquire the sensorimotor coordination which has stronger correlation by increasing the success rate of joint attention in the trials through the repetition of the trial and the learning.

4. Relationship to a Human Infant's Development

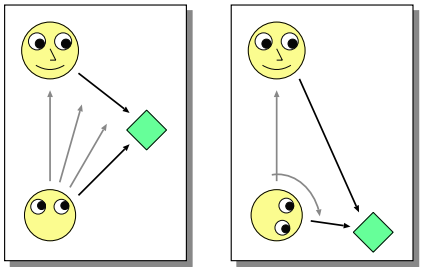
We can find a similarity between the incremental learning process of the robot's joint attention based on the proposed method and the development of an infant's joint attention. It is known in the developmental cognitive science that the infant has the following three stages of joint attention [2].

Ecological stage (6th months):

The infant follows the caregiver's gaze change or attends to an object which has interesting features (see Figure 8 (a)).



(a) 6th months



(b) 12th months (c) 18th months

Figure 8: Development of infant's joint attention

Geometric stage (12th months):
The infant tracks from the caregiver's eyes along the angle of its gaze and reliably attends to the object which the caregiver attends to. However, the infant performs the gaze following only when the object is within the field of the infant's view (see Figure 8 (b)).

Representational stage (18th months):
The infant turns around and attends to the object which the caregiver attends to by following the caregiver's gaze direction, even though the object is placed outside of the field of the infant's view (see Figure 8 (c)).

The developmental process of the infant's joint attention is closely similar to the incremental learning process of the robot's joint attention described in 2.2.. In addition, it is conjectured from the knowledge of developmental cognitive science that the visual attention and the self learning mechanisms are also seem to be prepared in the infant inherently [9]. Therefore, the proposed method could be useful to understand the developmental mechanisms of the infant's joint attention.

5. Conclusion

This paper has presented the method which develops the robot's ability of joint attention based on the visual attention and the self learning mechanisms. The experimental results showed that the ability of joint attention can be acquired based on the proposed method when the robot is provided the incremental learning process explicitly. Our future work is to examine whether the robot can acquire the ability of joint attention by generating the incremental learning process by itself. In addition, it is required to verify that the incremental learning can improve the final performance of the robot in the situation with multiple objects.

Acknowledgments

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

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] 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.

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

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

[5] 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.

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

[7] 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.

[8] 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.

[9] J. Gavin Bremner. *Infancy*. Blackwell, 1994.