Developmental Learning Model for Joint Attention

Yukie Nagai* Minoru Asada*† Koh Hosoda*†

*Dept. of Adaptive Machine Systems,
†HANDAI Frontier Research Center,
Graduate School of Engineering, Osaka University
e-mail: yukie@er.ams.eng.osaka-u.ac.jp, {asada, hosoda}@ams.eng.osaka-u.ac.jp

Abstract

This paper proposes a developmental learning model for joint attention between a robot and a human caregiver. The proposed model has abilities to accelerate the learning and improve the final task performance owing to two kinds of developments: a robot's development and a caregiver's one. The robot's development means that the sensing and the actuating capabilities of the robot change from immaturity to maturity. On the other hand, the caregiver's development is defined as that the caregiver changes the task from easy situation to difficult one. The proposed model causes these developments according to the learning progress of the robot. The experimental results showed what kinds of effects the developments bring to the learning.

1 Introduction

Robot learning can be accelerated and/or its final task performance can be improved by the developmental approaches. From a view point of cognitive developmental robotics [1], it is also meaningful that a learning system for a robot has developmental mechanisms toward making clear the emergence of intelligence and applying it to engineering.

Some researchers have attempted to apply the developmental approaches to robot leaning. Dominguez and Jacobs [4] showed that the learned performance of binocular disparity sensitivities by a developmental model was better than that by a non-developmental model. They constructed their developmental model by adding the spatial frequency information of an input image, that is divided into three levels: low, medium, and high frequencies, according to the learning time step. Metta *et al.* [10] built their developmental model to learn saccadic motion by changing the resolution of the input image from coarse to fine states. They also showed that their developmental model improved the final task performance. As above, it was realized that the developmental approaches improved the final task performance. However, the developmental mechanisms of their learning models were only built in the internal system of the robot. Uchibe *et al.* [15] proposed a method to control the environmental complexity and the cognitive capability of the learning robot. They showed that the robot learning of a soccer task was accelerated by their developmental model. However the final task performance was not improved by their model compared to a model with only the change of the environmental complexity.

It is well known in the developmental cognitive science that the developments of the attentional and the memorial abilities of children and caregiver's adaptive responses according to the children's level help the children to learn their first languages [5, 8, 13]. In other words, both development of internal mechanisms of a learner (robot's development) and that of caregiver's response mechanisms (caregiver's development) have potential to make the learning more efficient and improve the final task performance.

Joint attention is one of implicated tasks with development. Because the ability of joint attention is a fundamental one to communicate with others and acquire other social abilities, such as language understanding and mind reading [2]. Many researchers in cognitive developmental psychology have interests in it, because it is known that the ability of joint attention helps the children's development [12]. In the research area of engineering, some researchers have attempted to realize social communication between a human and a robot by constructing a joint attention mechanism inside the robot [7, 9, 14]. However, the joint attention mechanisms for their robots were built by the designers, and the psychological implication between the learning of joint attention and the development has not been involved.

This paper proposes a developmental learning model for joint attention that can accelerate the learning and improve the final task performance.

This model has two kinds of developments: a robot's development and a caregiver's one, and causes these developments by the learning progress of the robot. The robot's development is defined as that the sensing and the actuating capabilities of the robot change from immaturity to maturity. The caregiver's development means that the caregiver changes the task from easy situation to difficult one. First, the joint attention problem is defined, and it is argued how an infant acquires its ability. Next, the proposed developmental learning model is described and evaluated through some experiments. Finally, conclusions and future work are given.

2 Joint Attention

2.1 Task Definition

Joint visual attention is defined as that an agent attends to the same object which another attends to. **Figure 1** shows a process of joint attention between a robot and a caregiver. In this situation, the robot is the agent of the action. First, the robot observes the caregiver and estimates the direction of the caregiver's attention. Next, the robot turns the camera to the estimated direction and identifies the object which the caregiver attends to. These two steps are required for joint attention.

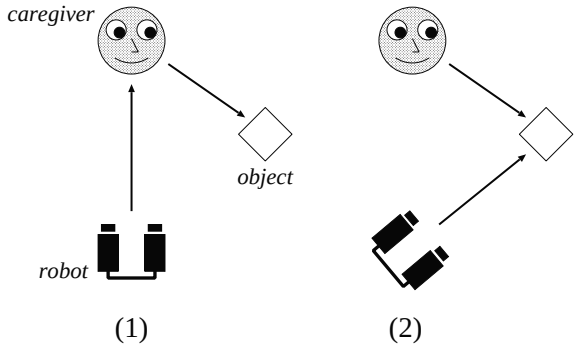


Figure 1: *Joint visual attention between a robot and a caregiver. The robot (1) observes the caregiver and (2) identifies the object which the caregiver attends to.*

2.2 Acquisition of The Capability of Joint Attention by A Human Infant

How does a human infant acquire the capability of joint attention? We refer to the following knowledge from developmental cognitive science to construct a developmental learning model for joint attention for a robot.

- An infant has an innate preference for a human face and observes it more closely than anything else [3, 6].
- A caregiver provides appropriate feedback to an infant, and it enables the infant to acquire the ability of joint attention [11].
- An infant can learn the ability of joint attention with a learning system based on reward, although he/she does not understand the meaning of the attention [11].

It is summarized that the learning of joint attention by an infant consists of two fundamental systems in the infant: a face detection system and a learning system based on reward, and one more system in the caregiver: an appropriate feedback system. Once the infant began joint attention, he/she can acquire the meaning of attention through the interactions among the three: the infant, the caregiver, and an object.

3 Developmental Learning Model for Joint Attention

The proposed developmental learning model for joint attention based on the previous arguments is shown in **Figure 2**. This model consists of two modules: one is a neural network for a robot and another is a task evaluator for a caregiver. The robot learns

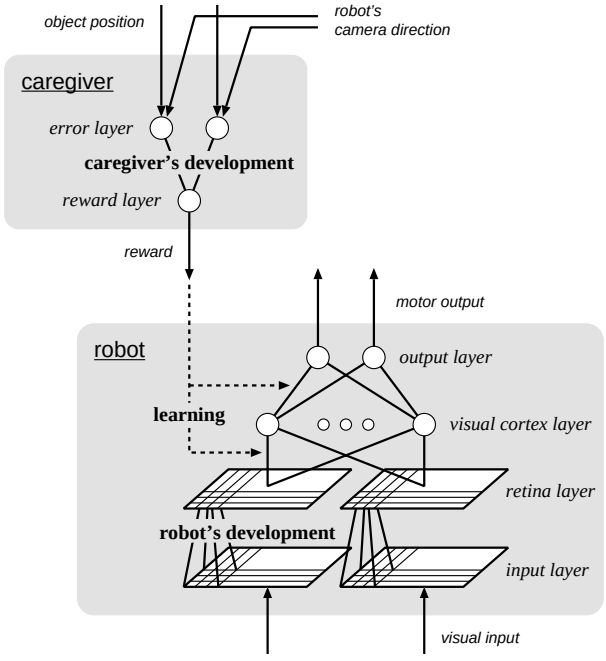


Figure 2: *A developmental learning model for joint attention*

the network based on a reward which is provided by the caregiver, and in parallel with the learning, the robot and the caregiver have each developmental mechanism.

3.1 Neural Network for A Robot

The neural network for a learning robot has four layers: an input layer, a retina one, a visual cortex one, and an output one. The inputs to the network are left and right camera images when the robot observes the caregiver (the situation shown in Figure 1 (1)), and the outputs are motor commands to attend to the same object which the caregiver attends to (the situation shown in Figure 1 (2)).

The developmental mechanism for the vision system of the robot is implemented in the connection weights between the input layer and the retina one. The information on the input and the retina layers are represented as images. The connection weight W_k^{ir} between the two layers at the learning time step k is given by a Gaussian spatial filter

$$W_k^{ir} = \exp\left(-\frac{(x-x_c)^2+(y-y_c)^2}{2{\sigma_k}^2}\right), \qquad (1)$$

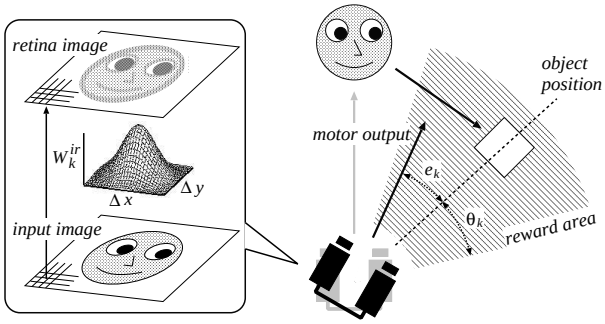
where (x, y) ,(x_c, y_c), and σ_k denote a position in the input image plane, a position of a target pixel of the spatial filter, and the sharpness of the spatial filter at k, respectively. The σ_k is determined by the task error[1] E

$$\sigma_k = \sigma_{init}\left(\frac{E_{k-1}-E_{fin}}{E_{init}-E_{fin}}\right), \qquad (2)$$

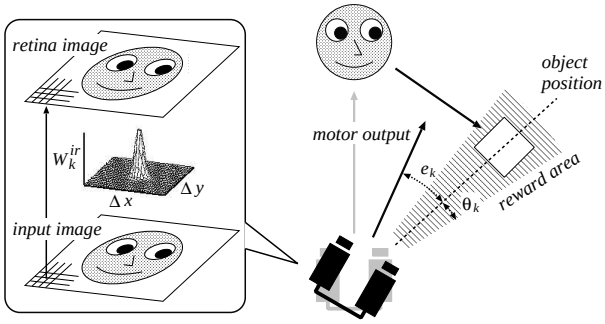
where σ_{init} is a initial value of σ_k, E_{k-1} is the task error at the learning time step $k-1$, E_{init} and E_{fin} denote the initial task error and the tolerance after the learning. This shift of σ_k means that the shape of the spatial filter changes from flat to steep, and this is regarded as visual development. The update of σ_k is triggered by the learning progress, that is when $E_{k-1} < \min E_j$ $(0 \le j < k-1)$. The left sides in **Figure 3** (a) and (b) show the appearances of the visual development inside the robot. In the early stage of the learning, (a) the retina image is blurred because the spatial filter is flat, and (b) the retina image becomes clear as well as the input one because the filter becomes steep in the later stage.

Other connection weights, W_k^{rc} between the retina layer and the visual cortex one and W_k^{co} between the visual cortex layer and the output one, are adjusted by the learning based on the reward. The robot executes the motor commands as the outputs

[1]Task error E_k is the average of the output errors e_k between the gaze direction of the robot's camera and the the object position in various situations, and it means the task performance at the learning time step k.



(a) the early stage of the learning



(b) the later stage of the learning

Figure 3: *The appearances of the robot's development (in the left sides) and the caregiver's one (in the right sides)*

of the neural network and receives a reward $R = 1$ or 0 as an evaluation of the action. The reward $R = 1$ means that the joint attention task is achieved, and $R = 0$ means that the task is not achieved. Then, the connection weights $W_k^{rc,co}$ are updated as follows

$$W_{k+1}^{rc,co} = \begin{cases} W_k^{rc,co}, & \text{when } R_k = 1 \\ W_k^{rc,co} \pm \Delta W, & \text{when } R_k = 0 \end{cases} \qquad (3)$$

where ΔW is a small random value. The reward R_k is determined by the caregiver according to the output error e_k between the motor output of the robot and the object position and the task error E_{k-1}.

3.2 Task Evaluator for A Caregiver

The task evaluator for a caregiver has two layers: an error layer and a reward one. The caregiver determines the reward R which is provided to the robot in accordance with the output error of the robot.

First, the caregiver measures the output error e_k between the gaze direction of the robot's camera and the object direction at the learning time step k.

Then the caregiver determines the reward R_k

$$R_k = \begin{cases} 1, & \text{when } e_k \leq \theta_k \\ 0, & \text{when } e_k > \theta_k \end{cases} \quad (4)$$

where θ_k is the tolerance of the output error at k. The θ_k is defined as

$$\theta_k = E_{k-1} - \epsilon, \quad (5)$$

where ϵ is a small value. The update of θ_k is triggered by the learning progress as well as σ_k, that is when $E_{k-1} < \min E_j$ $(0 \leq j < k-1)$. This shift of θ_k means that the caregiver change the task evaluation from easy situation to difficult one, and this is regarded as the caregiver's development.

The appearances of the caregiver's development are shown in the right sides of Figure 3 (a) and (b) in contrast with the robot's development. In the early stage of the learning, (a) the caregiver sets the reward area defined by θ_k as wide because the task error E_k has large value, so that the robot can learn the task easily. By contrast, (b) the caregiver changes the reward area as narrow and makes the task learning more accurately in the later stage.

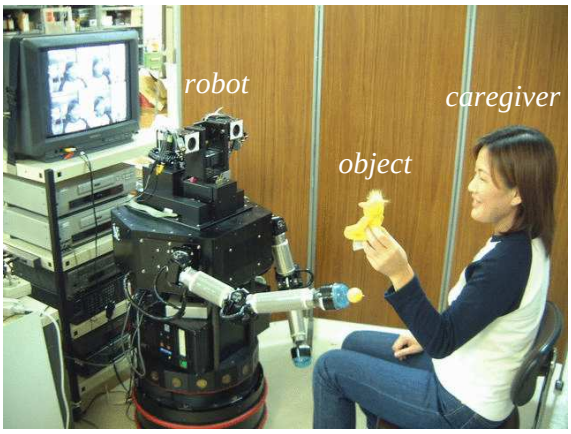
4 Experiment

4.1 Experimental Setup

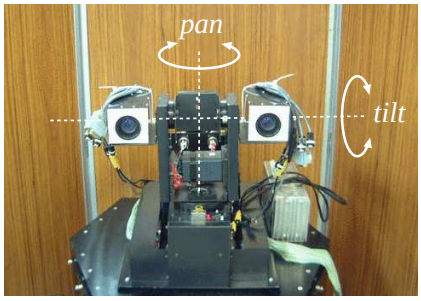
The experimental environment is shown in **Figure 4** (a), and the camera head of the robot is shown in Figure 4 (b). The robot is set in the face of the caregiver, and the each position of the robot and the caregiver is fixed. First, the caregiver holds an object in its hand and moves it to various positions within the range of the robot's vision. Then, the caregiver directs its attention to the object. At the same time, the robot observes this scene through its cameras and inputs the face images of the caregiver to the neural network. The neural network estimates the direction of the caregiver's attention and outputs two motor commands: pan and tilt angles common to left-and-right cameras, to direct the robot's attention to the same object. As a result of this process, the robot receives a reward from the caregiver and learns the neural network based on it.

4.2 Evaluation of The Learning Speed

To evaluate the learning speed of the proposed model, the error transition was compared with other three models. **Figure 5** shows the transitions of the average of the normalized output error through the learning process. Where the RC-dev. model is the proposed model which has the robot's development and the caregiver's one, and the R-dev. model, the C-dev. model, and the matured model denote a



(a) experimental environment



(b) the camera head of the robot

Figure 4: *An experimental setup for joint attention*

model which has only the robot's development, which has only the caregiver's one, and which does not have the developmental mechanism, respectively. What the model does not have the development means what the model has the matured mechanism. In other words, the connection weight W_k^{ir} of the matured mechanism (corresponding to Eq. (1)) is set as

$$W_k^{ir} = \begin{cases} 1 & x = x_c, y = y_c \\ 0 & x \neq x_c, y \neq y_c, \end{cases} \quad (6)$$

and the threshold θ_k of the matured mechanism (corresponding to Eq. (5)) is set as

$$\theta_k = \epsilon' \quad (7)$$

where ϵ' is a small value.

From the graphs in Figure 5, we can see the effect of the developmental mechanisms in the learning speed. The caregiver's development has an ability to accelerate the learning, although the robot's development has an effect to delay it. The visual development of the robot means that the perceptual capability of it is limited in the early stage and the middle

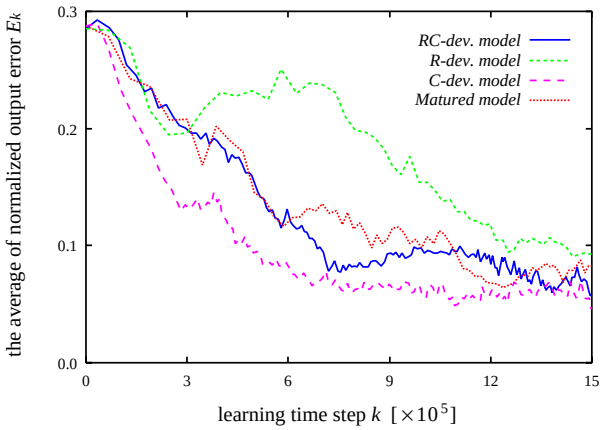


Figure 5: *The error transitions of four learning models. The RC-dev. model is the proposed model and the R-dev. model, the C-dev. model, and the matured model denote a model which has only robot's development, which has only caregiver's one, and which does not have any development, respectively.*

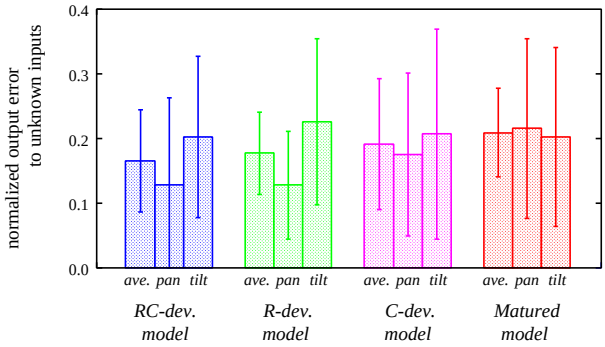


Figure 6: *Normalized task errors and standard deviations to unknown inputs*

one of the learning process. Because of its limitation, the learning with the robot's development is slowed down during these stages. On the other hand, the caregiver's development prompts the robot to learn the task, because it shifts the task level from easy to difficult one according to the learning progress by controlling the reward area.

4.3 Evaluation of The Final Task Performance

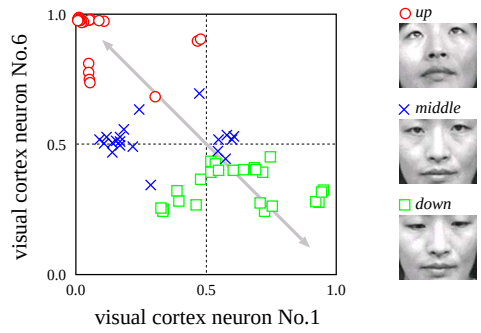
Experiments of joint attention were conducted with the neural network learned by the proposed developmental learning model. It was tested that the robot could identify the object which was set to a different point from the learned one by the caregiver. Through this experiment, it was confirmed that the robot was successful in joint attention task more than 90% of the trial.

Then, the final task performance of the proposed model was compared with the three models: the R-dev. model, the C-dev. model, and the matured model. The output errors of the neural networks learned by the each model to unknown inputs were measured. **Figure 6** shows the normalized output errors in the pan direction, the tilt one, and the average and their standard deviations of the four models. This result indicated that the RC-dev. model (the proposed model) has higher final task performance than other models because its output error is smaller than others. Both the robot's development and the caregiver's one have capabilities to improve the final task performance, and the capability of the robot's development is greater than that of the caregiver's one. It is supposed that a learning model with the robot's development, that is the visual development, could acquire generalized task performance, because the blurred retina image in the early stage of the learning enabled the robot to understand the abstract meaning of the joint attention task (input/output relationships).

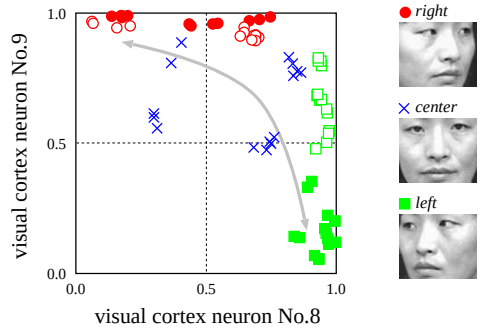
Next, the information extractions on the visual cortex layer in the neural network learned by the proposed model were investigated to verify the generalized task performance described above. **Figure 7** shows the responses of the visual cortex neurons to various input images. This distribution indicates that the visual cortex neurons of No.1 and No.6 estimate the vertical direction of the caregiver's attention (shown in Figure 7 (a)), and No.8 and No.9 estimate the horizontal one (shown in (b)). From these examination results, it is confirmed that the neural network learned by the proposed developmental learning model has acquired the response selectivity to the change of the direction of the caregiver's attention.

5 Conclusion

This paper presented the developmental learning model for joint attention between the robot and the caregiver. The proposed model has two developmental mechanisms: the robot's development and the caregiver's one. The robot's development means that the sensing (the visual) capability of the robot change from immaturity to maturity, and the caregiver's development means that the caregiver changes the task evaluation from easy situation to difficult one by controlling the tolerance of the output error. These two developments are caused by the learning progress of the robot. The experimental results showed that the caregiver's development accelerated the learning and the robot's development improved the final task performance. Therefore, we can



(a) No.1 and No.6 neurons have the response selectivity to the change of the vertical direction of the caregiver's attention.



(b) No.8 and No.9 neurons have the response selectivity to the change of the horizontal direction of the caregiver's attention.

Figure 7: *The responses of the visual cortex neurons to various input images*

conclude that the proposed model which has both the robot's development and the caregiver's one has the best capabilities.

Above effects which the robot's development and the caregiver's one brought the learning should be analyzed in detail, and it should be made clear how the developmental mechanisms made these effects. In addition, the acquired ability of joint attention by the proposed model is that the robot does not act with understanding the meaning of the attention but reacts to the visual inputs of the caregiver's face reflexively. The learning model for the robot should be added a mechanism to acquire the meaning of the attention through the experiences of joint attention and understand triadic representation among the robot, the caregiver, and an object.

Acknowledgments

This research was supported by the Japan Science and Technology Corporation, in Research for the Core Research for the Evolutional Science and Technology Program (CREST) titled Robot Brain Project in the research area "Creating a brain."

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 System*, Vol. 37, pp. 185-193, 2001.

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

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

[4] Melissa Dominguez and Robert A. Jacobs. "Visual Development and the Acquisition of Binocular Disparity Sensitivities," In *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[5] Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A connectionist perspective on development*, MIT Press, 1996.

[6] R. L. Fantz. *The Origin of Form Perception*, Scientific American, 1961.

[7] 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.

[8] Mark H. Johnson. *Developmental Cognitive Neuroscience*, Blackwell, 1997.

[9] 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.

[10] Giorgio Metta, Giulio Sandini, Lorenzo Natale, and Francesco Panerai. "Development and Robotics," In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pp. 33-42, 2001.

[11] Chris Moore and Valerie Corkum. *Social Understanding at The End of The First Year of Life*, Developmental Review, 1994.

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

[13] Elissa L. Newport. "Maturational Constraints on Language Learning," *Cognitive Science*, Vol. 14, pp. 11-28, 1990.

[14] Brian Scassellati. "Theory of Mind for a Humanoid Robot," In *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots*, 2000.

[15] Eiji Uchibe, Minoru Asada, and Koh Hosoda. "Environmental Complexity Control for Vision-Based Learning Mobile Robot," In *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 1865-1870, 1998.