Advanced Robotics, Vol. 20, No. 10, pp. 1165–1181 (2006)

Also available online - www.brill.nl/ar

Full paper

Learning for joint attention helped by functional development

YUKIE NAGAI [1,*], MINORU ASADA [2] and KOH HOSODA [2]

[1] *National Institute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0289, Japan*
[2] *Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan*

Received 7 April 2006; accepted 31 May 2006

Abstract—Cognitive scientists and developmental psychologists have suggested that development in perceptual, motor and memory functions of human infants as well as adaptive evaluation by caregivers facilitate learning for cognitive tasks by infants. This article presents a robotic approach to understanding the mechanism of how learning for joint attention can be helped by such functional development. A robot learns visuomotor mapping needed to achieve joint attention based on evaluations from a caregiver. The caregiver adjusts the criterion for evaluating the robot's performance from easy to difficult as the performance improves. At the same time, the robot also gradually develops its visual function by sharpening input images. Experiments reveal that the adaptive evaluation by the caregiver accelerates the robot's learning and that the visual development in the robot improves the accuracy of joint attention tasks due to its well-structured visuomotor mapping. These results constructively explain what roles synchronized functional development in infants and caregivers play in task learning by infants.

Keywords: Joint attention; visual development; adaptive evaluation; visuomotor learning.

1. INTRODUCTION

Human infants are born with immature capabilities. Their vision is blurred, their movement is uncoordinated and their memory is limited [1, 2]. As they grow, they develop and improve their capabilities through experiences with seeing, moving and thinking. Such development in perceptual, motor and memory functions of infants may help them to learn cognitive tasks. Their immature visual function, for example, causes them to detect only the important information in a complicated

*To whom correspondence should be addressed. Present address: Applied Computer Science, Faculty of Technology, Bielefeld University, PO Box 100 131, 33501 Bielefeld, Germany. E-mail: yukie@techfak.uni-bielefeld.de

environment. They can extract principal features from input information by using their immature visual capabilities as a filter. Newport [3] asserted that maturational constraints in infants' perceptual and memory functions aid their language learning. Their limited capabilities enable them to extract the essence of complicated utterances by adults and thereby acquire important language structures. This 'less is more hypothesis' is also considered to hold for learning other cognitive tasks. Caregivers, at the same time, adapt how they interact with infants as the infants grow. When infants only have immature capabilities, caregivers use simple and readily comprehensible behaviors. They amplify their actions, and talk slowly and rhythmically. Moreover, they adjust how to teach cognitive tasks to infants and to evaluate infant behavior. The difficulties of cognitive tasks are controlled from easy to difficult according to improvements in how well infants achieve the tasks. We suggest that these adaptations by caregivers have the effect of highlighting the important information in a complicated environment and consequently aid in learning for cognitive tasks by infants as functional development in infants does.

Several studies in computational science and robotics have evaluated these theories from a constructivist viewpoint [21]. Elman [4] empirically showed that functional development in a learner helped language learning. He compared learning involved in two types of neural networks: a fully formed network and one with limited memory that gradually changed into a fully formed one. His experiments revealed that only the latter could be trained to process complex sentences. Dominguez and Jacobs [5, 6] demonstrated the effect of visual development on visual information processing tasks, e.g., recognition of binocular disparity and motion velocity. They showed that neural networks with a mechanism for increasing the number of input neurons achieved higher accuracy in visual recognition tasks than networks without such a mechanism. In one robotic approach, Metta *et al.* [7, 8] showed that a developmental mechanism could improve robot's learning. Their robot with a mechanism for visual development efficiently acquired the abilities to gaze at and to reach out to a visual target. Uchibe *et al.* [9] investigated whether development not only in a robot but also in the environment helped the robot to learn a soccer task. They used a mobile robot with a developmental mechanism that increased the dimension of its state vector and trained it to shoot a ball. The environment was also controlled to increase its complexity by speeding up a goalkeeper robot as learning proceeded. They found that development in both the robot and the environment facilitated learning. All these studies empirically verified the validity of the 'less is more hypothesis'. However, it is still an open question as to how developmental mechanisms affect the learning mechanisms involved in acquiring sensorimotor mapping. We speculate that they affect how well the mapping is structured.

We introduce joint attention tasks to investigate what effects functional developments in both infants and caregivers have on task learning. Joint attention is defined as looking at an object that someone else is looking at by following the direction of

his or her gaze [10, 11]. Infants are suggested to acquire this ability by 18 months of age [10], which means that they develop their perceptual, motor and memory functions as they learn to achieve joint attention. Caregivers are also considered to adapt how they interact with and evaluate infants. They may adjust the position of an object to be gazed at so that an infant can easily detect it. We therefore suggest that joint attention is an adequate task for investigating how functional developments in infants and caregivers relate to each other, and how these developments affect task learning.

Although many robotic models for performing joint attention have been built [12–15], they were fully structured ones, i.e., they did not incorporate any mechanisms for learning or functional development. They focused only on establishing human–robot interactions and were aimed at investigating the psychological effects of joint attention on communications. In contrast, Triesch *et al.* [16, 17] and Nagai *et al.* [18] developed learning models by which a computational agent or a robot acquired the ability to establish joint attention with a caregiver. They were motivated by the results of infant studies and designed their models so that learners acquired their abilities like infants. Their models, however, did not enable the learners and caregivers to develop their perceptual, motor and evaluation functions. They only discussed how mature functions enabled the learners to acquire the abilities.

We propose a developmental learning model by which a robot develops its visual function as it learns to achieve joint attention based on adaptive evaluation by a human caregiver. The robot improves its visual ability by gradually sharpening its input images as learning proceeds. This approach is based on evidence that human neonates have only a 1/30th of the visual acuity of adults and that their acuity improves as they grow [2]. The caregiver, on the other hand, adjusts the criterion used to evaluate the robot's performance of joint attention tasks. He or she changes the difficulty of the tasks from easy to difficult by reducing the tolerance against the robot's output error according to improvements in the robot's performance. This corresponds to a caregiver positioning objects so that an infant in the early stages of development can easily find them [11]. We investigate how these developmental mechanisms facilitate learning for joint attention and how they affect the structuring of visuomotor mapping in a robot.

The rest of the paper is organized as follows. First, we define human–robot joint attention. Then, we describe our developmental learning model for joint attention. The mechanisms responsible for visual development in a robot, adaptive evaluation by a caregiver and visuomotor learning by a robot are explained. Next, we show our experiments for evaluating the effects of functional development on learning to achieve joint attention. To clarify the effects of development in a robot and a caregiver, we compare our experimental results with those for three other models without development in either or both the robot and caregiver. Finally, we conclude with a discussion of the results and of future directions for additional research.

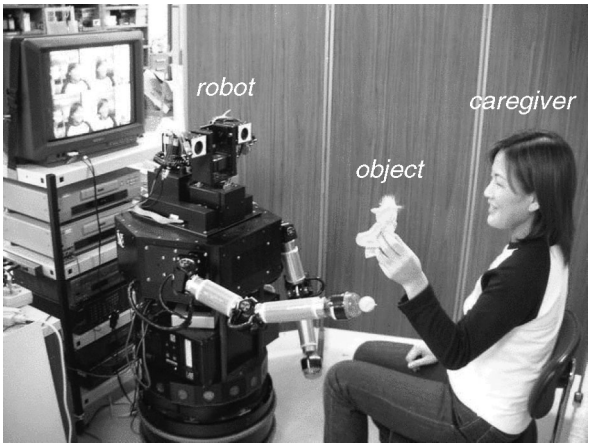


Figure 1. Experimental environment for joint attention, where the robot learned to gaze at an object that the caregiver was viewing by following the direction of her gaze.

2. HUMAN–ROBOT JOINT ATTENTION

We used joint attention as a task for evaluating the effects of functional development on task learning. Figure 1 shows the experimental environment, where the robot learned to look at the object that the human caregiver was viewing by following the direction of her gaze. In each trial, the caregiver replaced the object at different positions and gazed at it in front of her face. The robot observed the caregiver with head-mounted cameras, and visually tracked certain directions in the environment by panning and tilting its camera head. The robot acquired the ability to establish joint attention with the caregiver by learning mapping from the visual input, i.e., camera images, to the motor output, i.e., displacement angles of the camera head.

Note that the process of joint attention discussed here does not involve the robot's understanding of sharing attention with the caregiver, but is realized only based on its visuomotor learning. This corresponds to the first stage of the development of joint attention in infants. Infants are suggested to first engage in joint attention without understanding the nature of attention of others and come to comprehend the attention through these experiences [19]. We supported this idea and investigated how visuomotor learning for achieving joint attention could be helped by functional development.

3. DEVELOPMENTAL LEARNING MODEL FOR JOINT ATTENTION

Figure 2 presents a developmental learning model for joint attention, which consists of a neural network for the robot and a task evaluator for the caregiver. The neural network enables the robot to acquire the visuomotor mapping needed to achieve joint attention as it develops its visual function. The task evaluator enables the caregiver to provide appropriate feedback to the robot regarding its performance of joint attention tasks. The learning procedure is as follows:

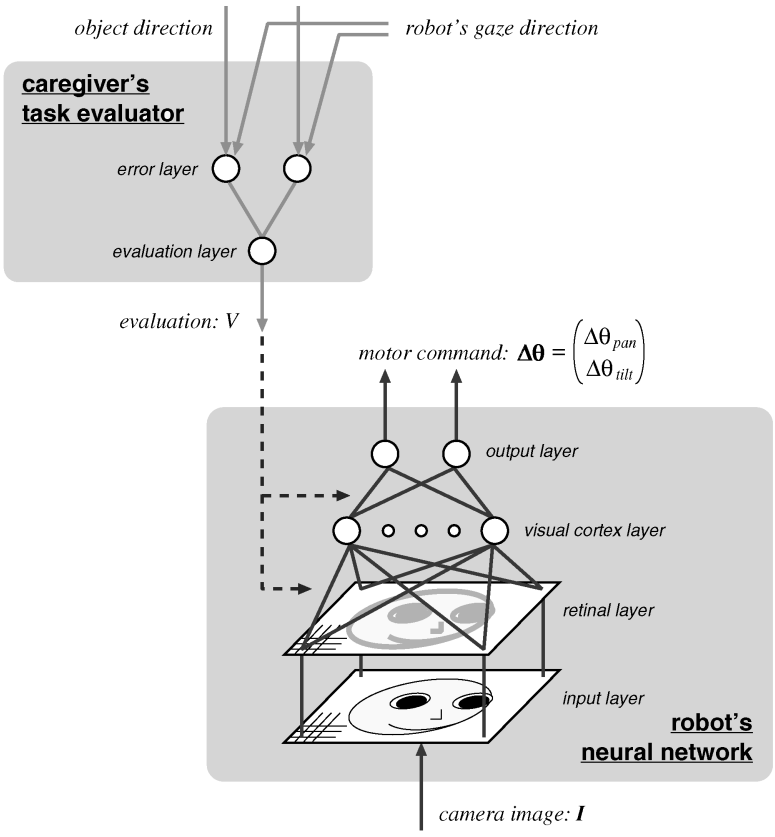


Figure 2. Developmental learning model for joint attention.

(i) The robot first gazes at the caregiver, who is looking at an object, and captures a camera image $\boldsymbol{I}$ of her face. The image $\boldsymbol{I}$ is input to the neural network.

(ii) The neural network produces a retinal image by blurring the input image with a smoothing filter and then generates motor output $\boldsymbol{\Delta\theta} = [\Delta\theta_{\text{pan}},\ \Delta\theta_{\text{tilt}}]$ based on the retinal image.

(iii) The robot pans and tilts its camera head based on $\boldsymbol{\Delta\theta}$ and looks in a certain direction in the environment.

(iv) The caregiver detects the output error between the direction of the robot's gaze and the direction of the target object, and provides evaluation V to the robot. Evaluation V has a value of 1 or 0, meaning joint attention has succeeded or failed.

(v) The robot modifies the connecting weights of its neural network based on V.

(vi) Return to (i).

As learning proceeds, the robot develops its visual function by adjusting the smoothing filter so that retinal images become less blurred. The caregiver, at the

same time, adapts the criterion for evaluating the robot's performance from easy to difficult.

3.1. Visual development in the robot

The robot develops its visual function by sharpening the smoothing filter between the input and retinal layers. A camera image $\boldsymbol{I}$ of the caregiver's face is first input to the neural network as grayscale information and is then reproduced as a retinal image through the smoothing filter. The filter W_k^{ir}, where k denotes the learning steps, is defined as a Gaussian function:

$$W_k^{\mathrm{ir}} = \exp\left(-\frac{(x - sx)^2 + (y - sy)^2}{2{\sigma_k}^2}\right), \tag{1}$$

where $(x,\ y)$, $(sx,\ sy)$ and σ_k are a position in the input image, the target position of the filter and the variance of the filter, respectively. This filter blurs the input image by being applied to all pixels in the image. The visual function develops by sharpening the filter as the robot improves its joint attention performance. The variance σ_k, which determines the sharpness of the filter, is updated by:

$$\sigma_k = \sigma_{\mathrm{init}}\left(\frac{\bar{e}_{k-1} - e_{\mathrm{fin}}}{\bar{e}_0 - e_{\mathrm{fin}}}\right), \tag{2}$$

where $\bar{e}_0$ and $\bar{e}_{k-1}$ are the means of the robot's output error at the beginning of learning and at learning step $k - 1$. This means that the filter becomes steeper as the error decreases. The parameters σ_{init} and e_{fin}, given by a designer, define the initial and end conditions for visual development. For example, a large σ_{init} value makes the robot start with a more immature visual function, i.e., the robot receives blurrier images at the beginning of learning. A small e_{fin} value makes it difficult for the robot to fully develop its visual function. Note that σ_k is updated only when:

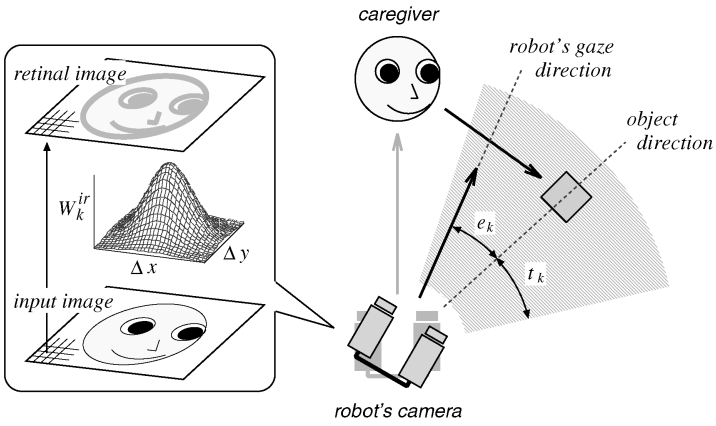
$$\bar{e}_{k-1} < \min \bar{e}_j \quad (0 \leqslant j < k - 1), \tag{3}$$

i.e., visual development is caused by improvements in the robot's performance of joint attention.

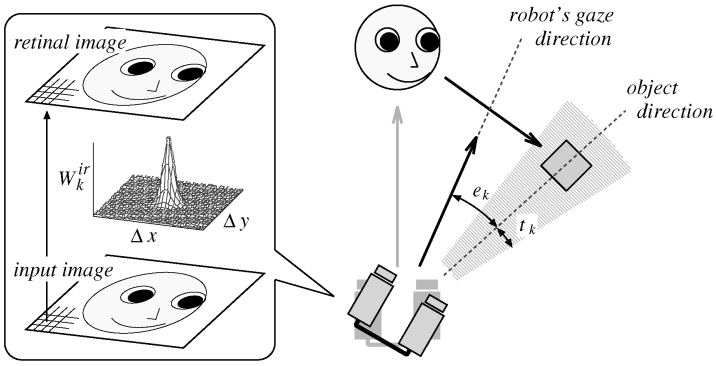
The mechanism responsible for visual development is illustrated on the left of Fig. 3, whereas the mechanism responsible for adaptive evaluation by the caregiver, which is explained in the next section, is on the right. The normal distribution surface between the input and retinal layers represents the smoothing filter W_k^{ir}, through which the retinal image is produced from the input image.

(i) In the early stages of learning, the filter has a large variance σ_k because the output error $\bar{e}_{k-1}$ nearly equals $\bar{e}_0$ in (2). The robot thus receives a blurred image on the retinal layer.

(ii) In the later stages of learning, the filter becomes steeper because $\bar{e}_{k-1}$ approaches e_{fin} and the robot receives a sharper image on the retinal layer.



(a) In early stages of learning, robot obtains blurred images on retinal layer because of large variance of smoothing filter. Caregiver uses loose criterion to evaluate robot's performance by setting large tolerance.



(b) In later stages of learning, robot obtains sharper images by reducing variance of filter. Caregiver tightens criterion for evaluation by reducing tolerance.

Figure 3. Mechanisms for visual development in the robot (left) and adaptive evaluation by the caregiver (right).

As a result, the robot learns using only the principal features of input images in the early stages of learning whereas it learns using more features in the later stages. This should enable the robot to acquire well-structured visuomotor mapping needed to achieve joint attention.

3.2. Adaptive evaluation by the caregiver

The caregiver adjusts the criterion for evaluating the robot's performance of joint attention according to improvements in how well it establishes the tasks. After the

robot has turned its camera head based on the output from its neural network, the caregiver detects output error e_k between the direction of the robot's gaze and the direction of the target object. She then determines the value of evaluation V_k:

$$V_k = \begin{cases} 1, & \text{if } |e_k| \leqslant t_k \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

where t_k is the tolerance against the output error. Evaluation $V_k = 1$ means joint attention has been successful, while $V_k = 0$ means failure. In other words, the caregiver counts the robot's output as successful joint attention if it gazed at the object within a center circle with radius t_k in its camera image or failure otherwise. Adaptive evaluation is achieved by changing t_k according to the improvements in the robot's performance:

$$t_k = \bar{e}_{k-1} - \epsilon, \tag{5}$$

where $\bar{e}_{k-1}$ is the mean of the robot's output error at $k-1$ and ϵ is a small value. This means that the caregiver sets the difficulty of the joint attention task a little higher than the current level. Note that t_k is updated only when:

$$\bar{e}_{k-1} < \min \bar{e}_j \quad (0 \leqslant j < k-1), \tag{6}$$

i.e. the criterion for evaluating the robot's performance becomes more difficult as learning proceeds.

The mechanism responsible for adaptive evaluation by the caregiver is illustrated on the right of Fig. 3, where the sectored area represents the tolerance t_k against the robot's output error. If the direction of the robot's gaze is within the area, V_k is set to 1; otherwise it is set to 0.

(i) In the early stages of learning, the caregiver sets a large tolerance t_k because the robot has a large error $\bar{e}_{k-1}$ in (5). She, therefore, allows the robot to easily acquire a rough visuomotor map to achieve joint attention.

(ii) In the later stages of learning, the caregiver decreases t_k because the robot has reduced $\bar{e}_{k-1}$. She, thus, enables the robot to improve the accuracy of its visuomotor map.

This adaptive evaluation should accelerate the robot's learning.

3.3. Visuomotor learning based on task evaluation

The robot learns its visuomotor mapping based on evaluation V_k from the caregiver. It uses V_k to modify the connecting weights W_k^{rc} between the retinal and visual cortex layers and W_k^{co} between the visual cortex and output layers:

$$W_{k+1}^{\mathrm{rc,co}} = \begin{cases} W_k^{\mathrm{rc,co}}, & \text{when } V_k = 1 \\ W_k^{\mathrm{rc,co}} \pm \Delta W, & \text{when } V_k = 0, \end{cases} \tag{7}$$

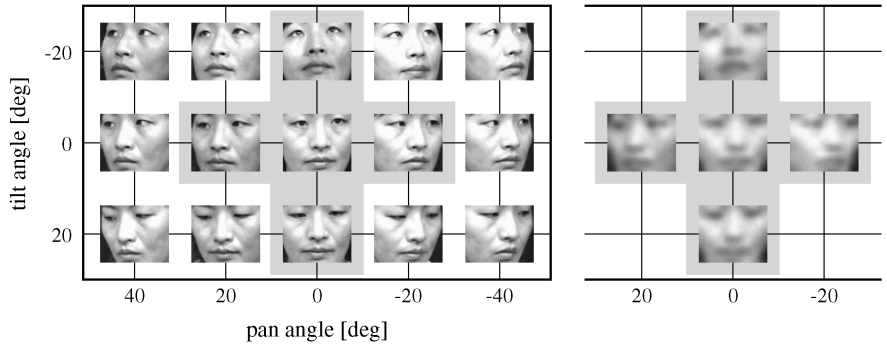
where ΔW denotes a small random value. This means that the neural network remains unchanged when the robot has received a good evaluation. Otherwise, it

is slightly modified by random changes to the connecting weights. The weights are changed randomly because the caregiver cannot teach how the robot should modify them. She can only inform it whether joint attention has succeeded or failed, but cannot teach it how to change its visuomotor mapping. In this way, the robot gradually improves the accuracy of its mapping.

4. EXPERIMENTS

4.1. Method

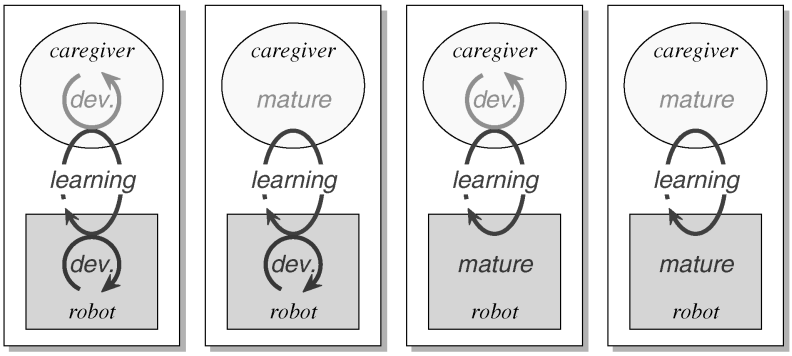
We experimentally evaluated how functional development affected learning. To conduct learning experiments off-line, we had the robot shown in Fig. 1 acquire input–output datasets beforehand. The input data were the camera images of the caregiver's face detected with 30 × 25 pixels. The corresponding output data were the displacement angles of the pan and tilt of the robot's head when it gazed correctly at the object that the caregiver was viewing. The examples of input images shown in Fig. 4a were captured when the caregiver was gazing at an object by panning from $-40°$ to $40°$ and tilting from $-20°$ to $20°$. The angles correspond to the motor output acquired when the robot gazed at the same object. Figure 4b shows the retinal images generated from the input images in Fig. 4a; only five images are presented as examples. The robot started learning with blurred images like these. Seventy-five datasets, five datasets at each position, were acquired in advance and used repeatedly throughout learning experiments. The neural network consisted of 750 input neurons, 750 retinal neurons, seven visual cortex neurons and two output neurons. The parameters were set to $\sigma_{\text{init}} = 3.0$, $e_{\text{fin}} = 0.05$, $\epsilon = 0.02$ and $\Delta W = 0.007$ by trial and error.



(a) Input images captured when caregiver panned her head from -40 to 40 [deg] and tilted from -20 to 20 [deg] to look at object.

(b) Retinal images generated from input images in (a) at beginning of learning.

Figure 4. Examples of input and retinal images.



(a) *RC-dev model* (proposed model) (b) *R-dev model* (c) *C-dev model* (d) *Mature model*

Figure 5. Conceptualizations of comparative learning models with and without a developmental mechanism in robot and/or caregiver.

We compared the performance of our learning model against that of three other models to evaluate how effectively functional development improved the robot's ability to achieve joint attention. Figure 5 shows conceptualizations of (a) the proposed model, called the RC-dev model, and three comparative models: (b) the R-dev model, (c) the C-dev model and (d) the Mature model. The RC-dev model has a developmental mechanism in both the robot and caregiver. The R-dev and C-dev models have a developmental mechanism in only the robot or caregiver, respectively, and the Mature model has no such mechanism. The caregiver in the R-dev and Mature models and the robot in the C-dev and Mature models are instead equipped with mature functions. That is, from the beginning of learning, the robot receives retinal images as clear as input images. The caregiver sets the criterion for evaluating the robot's performance to the most difficult level and never changes this over the learning period. We conducted learning experiments to evaluate (i) learning speed and (ii) accuracy in joint attention tasks employing these four models.

4.2. Results

4.2.1. Learning speed. We first compared the learning speed with the four models. We considered that functional development in the robot and/or caregiver would affect the learning speed for joint attention.

Figure 6 shows the changes in the output error over learning. The horizontal and vertical axes denote the learning steps k and the normalized output error $\bar{e}_k$ in the neural network, where $\bar{e}_k = 0.1$ means that the network has 9° of error between the direction of the robot's gaze and the direction of the target object. The four curves correspond to the four models in Fig. 5. Comparison of the results showed that adaptive evaluation by the caregiver accelerated learning for joint attention. The output error in the RC-dev model decreased faster than that in the R-dev model

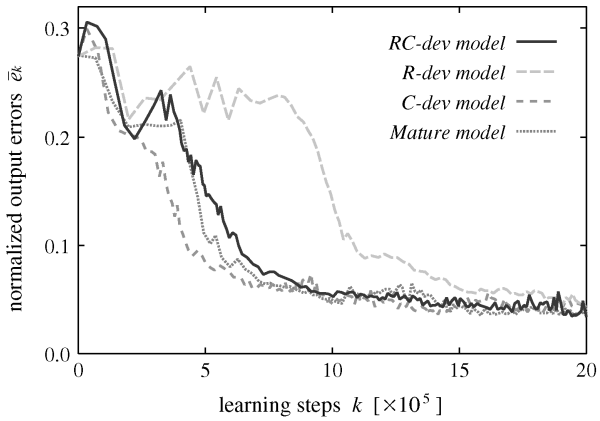
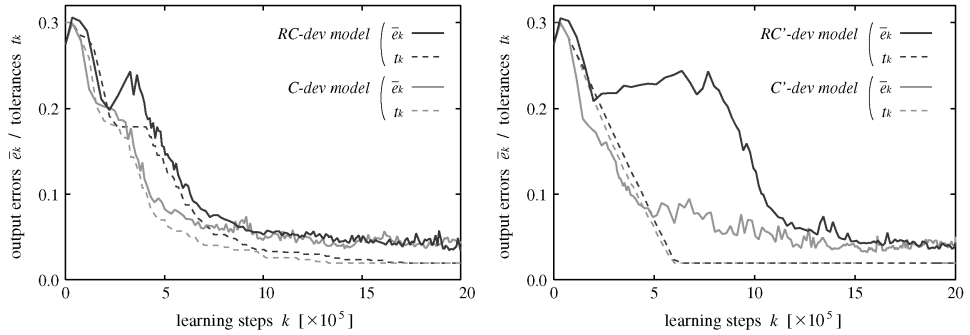


Figure 6. Changes in normalized output error $\bar{e}_k$ over learning.

and the output error in the C-dev model decreased faster than that in the Mature model. The learning speed was especially accelerated in the early stages of learning, although that in the later stages was almost the same. This suggests that adaptive evaluation enabled the robot to rapidly acquire a rough visuomotor map to achieve joint attention and to refine it as learning proceeded. In contrast to the acceleration caused by adaptive evaluation, the comparison of the results also showed that visual development in the robot decelerated learning. The learning speed with the RC-dev and R-dev models was lower than that with the C-dev and Mature models. The visual development decelerated learning because the blurred retinal images lacked the detailed information in the input images. As a result, the robot could not estimate the exact direction of the caregiver's gaze in the early stages of learning.

4.2.2. Relationship between learning speed and trigger for adaptation. How can the trigger for adaptation in evaluating the robot's performance affect the acceleration of learning? We assumed that appropriate timing for updating the tolerance t_k accelerated learning more. We thus compared the learning speed of the RC-dev and C-dev models, in which t_k was updated when the robot's output error $\bar{e}_k$ had decreased, with that of the RC′-dev and C′-dev models, in which t_k was updated based on a given clock.

The results for the RC-dev and C-dev models are shown in Fig. 7a, and those for the RC′-dev and C′-dev models are shown in Fig. 7b. The solid and dashed curves denote changes in $\bar{e}_k$ and t_k, respectively. We can see in Fig. 7a that t_k decreased with the improvements in $\bar{e}_k$ while in Fig. 7b it decreased linearly. The clock trigger for t_k in Fig. 7b was designed through trial and error. These results showed that adaptive evaluation triggered by a given clock accelerated learning although its effectiveness strongly depended on the timing. The learning speed with the C′-dev model was higher than that with the C-dev model because the decrease in t_k was synchronized with the decrease in $\bar{e}_k$. However, the learning speed with the RC′-dev model was not as high as that with the RC-dev model because the decrease in t_k



(a) t_k was updated when $\bar{e}_k$ decreased. (b) t_k was updated based on given clock.

Figure 7. Relationship between changes in normalized output error $\bar{e}_k$ and tolerance t_k.

was too rapid. This means that adaptation in task evaluation that is not synchronized with the improvements in task performance may not accelerate learning. Therefore, we suggest that the timing for adaptation in evaluating tasks should be designed to match the improvements in task performance.

4.2.3. Task accuracy. We next compared the accuracy in joint attention tasks after learning. Although visual development in the robot had no advantages in evaluating learning speed, we expected it would produce good results for task accuracy.

Figure 8 shows the normalized output errors $\bar{e}$ in the acquired neural networks when unknown inputs were received after learning. The four bars correspond to the four models in Fig. 5. The unknown input data were 45 images of the same caregiver's face captured when she was looking in directions different from those in the learning experiments. By comparing the results in each graph, we can see that visual development in the robot improved the accuracy in joint attention tasks. The output error for the RC-dev model ($M = 0.128$, SD $= 0.081$) was less than that for the C-dev model ($M = 0.171$, SD $= 0.087$) and the error for the R-dev model ($M = 0.125$, SD $= 0.045$) was less than that for the Mature model ($M = 0.189$, SD $= 0.067$). The difference in normalized output error $\bar{e} \simeq 0.05$ equals 4.5° error in the displacement angles. Tukey's method showed that there were significant differences between the models with visual development and those without it ($P < 0.05$). We attributed the improvement in task accuracy to the immature visual function. It enabled the robot to gradually extract the principal features of the input images, to learn to achieve joint attention in stages and consequently to acquire a well-organized visuomotor map.

4.2.4. Relationship between task accuracy and structure of visuomotor mapping. How was the structure of visuomotor mapping affected by visual development? We

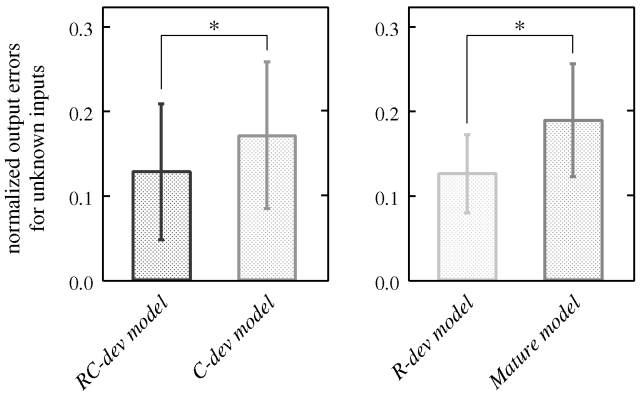


Figure 8. Normalized output error for unknown inputs after learning. $^*P < 0.05$, calculated using Tukey's method.

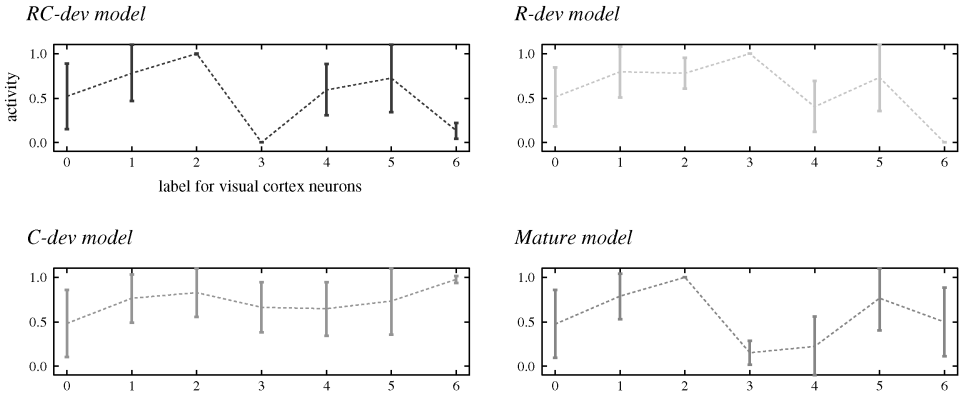


Figure 9. Activities of visual cortex neurons responding to unknown inputs.

postulated that the improvement in accuracy for joint attention tasks was due to well-structured mapping. To verify this, we compared the internal representations of the acquired neural networks.

Figure 9 plots the mean activities and standard deviations of the visual cortex neurons when the neural networks received unknown inputs after learning. The horizontal axis denotes the labels for the neurons, and the vertical axis denotes their activities. The unknown data were the same as in the previous experiment. We can see from the results that the number of neurons for which the standard deviation equaled zero was increased by visual development. There were two neurons with zero standard deviation each in the RC-dev model (nos 2 and 3) and R-dev model (nos 3 and 6), whereas there was only one in the Mature model (no. 2). The C-dev model did not have any such neurons. Zero standard deviation means that the neurons had not been used for joint attention tasks. In other words, only neurons with large variances had been used to recognize input images. The means for the number of the unutilized neurons were 1.2 in the RC-dev, 1.2 in the

R-dev, 0.7 in the C-dev and 0.6 in the Mature models. We thus confirmed that the internal representations of visuomotor mapping were more downsized by visual development.

The mechanism for downsizing visuomotor mapping is considered to be as follows. First, the maturational constraint in the visual function enabled the robot to extract only the principal features from the input images. As we can see from Fig. 4b, the blurred retinal images retained variances mainly in the horizontally distributed images. This enabled the robot to learn by first mainly focusing on the horizontal differences. As learning proceeded, the robot gradually came to detect the vertical differences as well by receiving sharper retinal images and to use both differences in learning. As a result, it learned to achieve joint attention in stages, i.e., first horizontally and then vertically. This is why the robot with visual development acquired downsized and well-organized visuomotor mapping, and consequently improved its accuracy in joint attention tasks.

4.3. Joint attention experiments

Finally, we conducted joint attention experiments in an actual environment shown in Fig. 1 to evaluate the effectiveness of the acquired neural network. The robot was embedded with a neural network learned with the RC-dev model. The caregiver, the same person as in the learning experiments, sat in front of the robot, placed an object at random positions and gazed at it. The timing at which the robot's camera captured an image of her face and turned its head based on the output from the neural network was controlled by the caregiver.

Figure 10 shows the examples of camera images for when the robot tried to establish joint attention. The rectangle in each denotes the input image for which grayscale information was input to the neural network. The line indicates the motor

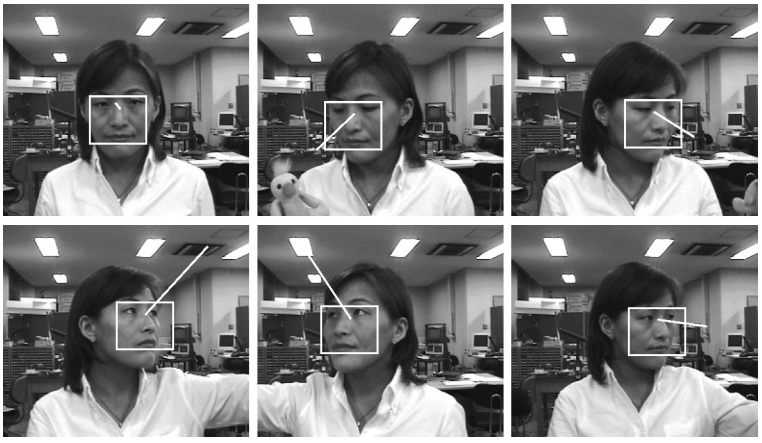


Figure 10. Examples of camera images for when the robot tried to achieve joint attention. Rectangles denote input image and lines denote motor output, where horizontal and vertical components correspond to displacement angles for panning and tilting.

output by which the robot turned its head. The horizontal and vertical components of the line correspond to the displacement angles for panning and tilting. Note that the line does not show the direction of the caregiver's gaze, but the motor output from the neural network. These results showed that the neural network could generate appropriate motor output to follow the direction of the caregiver's gaze. The success rate for joint attention was 95% (=19/20 trials), where a trial was counted as successful if the robot gazed at the target object within a centered circle of the camera image. We concluded that the proposed model enabled the robot to acquire appropriate visuomotor mapping to achieve joint attention.

5. DISCUSSION AND FUTURE WORK

This paper presented a developmental learning model for joint attention based on the theory that development helps learning. The model enabled a robot to develop its visual function as it improved its performance of joint attention. A caregiver provided appropriate feedback to the robot according to the improvements in the robot's performance. Employing the model, we examined how functional development in the robot and caregiver facilitated robot learning. Our experimental results revealed that:

- Adaptive evaluation by a caregiver accelerated the speed of learning, especially when the criterion for evaluation was tightened as the robot's performance improved.
- Visual development in the robot improved its task accuracy by enabling it to acquire downsized and well-organized visuomotor mapping.

Several researchers in the fields of cognitive science and developmental psychology have pointed out the importance of development in task learning by infants [3, 11]. They suggested that development in perceptual, motor and memory functions of infants as well as adaptive evaluation by caregivers may help infants to learn cognitive tasks. However, the mechanisms for how development affects learning have not been completely uncovered. Our experimental results empirically demonstrated the theory that development helps learning and provided explanations for the mechanisms. Caregivers assist infants to learn cognitive tasks by controlling the difficulty of the tasks so that infants learn in incremental steps. This enables infants to rapidly acquire rough coordination needed to perform the tasks. The developmental capabilities of infants themselves also play a role. Their immature perceptual, motor and memory functions in early infancy enable them to deal only with more important information and to increase the complexity of information they deal with as they develop. An interesting finding from our experiments is that visual development helped the robot to learn to establish joint attention first horizontally and then vertically, as in infants. Infants are also suggested to first come to follow the direction of another person's gaze when he/she has turned his/her head horizontally and then vertically [20]. This correspondence in learning between the robot

and infants should take us somewhat closer to revealing the learning mechanisms of infants [21].

We intend to modify the learning model so that a robot can use various image features, such as edges and motion, as input information. As the model here used only grayscale information from camera images, the robot was overly sensitive to lighting conditions and differences in the facial features of the caregiver. We expect that using various image features will enable the robot to acquire more generalized and robust capabilities. Furthermore, it should enable us to understand the roles that image features play in learning for joint attention. We also intend to investigate the extent to which motor and memory functions develop. Whereas our robot only developed perceptually, human infants develop their motor and memory functions as well. Development of these functions is considered to be intricately interrelated, and they all facilitate learning. Therefore, we intend to investigate how these developments assist task learning.

REFERENCES

1. D. Maurer and C. Maurer, *The World of the Newborn*. Basic Books, New York (1988).
2. J. G. Bremner, *Infancy*. Blackwell, Oxford (1994).
3. E. L. Newport, Maturational constraints on language learning, *Cognitive Sci.* **14**, 11–28 (1990).
4. J. L. Elman, Learning and development in neural networks: the importance of starting small, *Cognition* **48**, 71–99 (1993).
5. M. Dominguez and R. A. Jacobs, Developmental constaints aid the acquisition of binocular disparity sensitivities, *Neural Comput.* **15**, 161–182 (2003).
6. R. A. Jacobs and M. Dominguez, Visual development and the acquisition of motion velocity sensitivities, *Neural Comput.* **15**, 761–781 (2003).
7. G. Metta, G. Sandini and J. Konczak, A developmental approach to visually-guided reaching in artificial systems, *Neural Networks* **12**, 1413–1427 (1999).
8. G. Metta, G. Sandini, L. Natale and F. Panerai, Development and robotics, in: *Proc. IEEE–RAS Int. Conf. on Humanoid Robots*, Tokyo, pp. 33–42 (2001).
9. E. Uchibe, M. Asada and K. Hosoda, Environmental complexity control for vision-based learning mobile robot, in: *Proc. IEEE Int. Conf. on Robotics and Automation*, Leuven, pp. 1865–1870 (1998).
10. G. Butterworth and N. Jarrett, What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy, *Br. J. Dev. Psychol.* **9**, 55–72 (1991).
11. C. Moore and P. J. Dunham (Eds), *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum, Englewood Cliffs, NJ (1995).
12. C. Breazeal and B. Scassellati, Infant-like social interactions between a robot and a human caregiver, *Adaptive Behavior* **8**, 49–74 (2000).
13. B. Scassellati, Theory of mind for a humanoid robot, *Autonomous Robots* **12**, 13–24 (2002).
14. H. Kozima and H. Yano, A robot that learns to communicate with human caregivers, in: *Proc. 1st Int. Workshop on Epigenetic Robotics*, Lund, pp. 47–52 (2001).
15. M. Imai, T. Ono and H. Ishiguro, Physical relation and expression: Joint attention for human-robot interaction, in: *Proc. 10th IEEE Int. Workshop on Robot and Human Interactive Communication*, Bordeaux, pp. 512–517 (2001).
16. I. Fasel, G. O. Deák, J. Triesch and J. Movellan, Combining embodied models and empirical research for understanding the development of shared attention, in: *Proc. 2nd Int. Conf. on Development and Learning*, Cambridge, MA, pp. 21–27 (2002).

17. E. Carlson and J. Triesch, A computational model of the emergence of gaze following, in: *Proc. 8th Neural Computation and Psychology Workshop*, Canterbury (2003).
18. Y. Nagai, K. Hosoda, A. Morita and M. Asada, A constructive model for the development of joint attention, *Connection Sci.* **15**, 211–229 (2003).
19. V. Corkum and C. Moore, Development of joint visual attention in infants, in: *Joint Attention: Its Origins and Role in Development*, C. Moore and P. J. Dunham (Eds), pp. 61–83. Lawrence Erlbaum, Englewood Cliffs, NI (1995).
20. C. Moore, M. Angelopoulos and P. Bennett, The role of movement in the development of joint visual attention, *Infant Behav. Dev.* **20**, 83–92 (1997).
21. M. Asada, K. F. MacDorman, H. Ishiguro and Y. Kuniyoshi, Cognitive developmental robotics as a new paradigm for the design of humanoid robots, *Robotics Autonomous Syst.* **37**, 185–193 (2001).

ABOUT THE AUTHORS

Yukie Nagai received her BE and ME degrees in Engineering from Aoyama Gakuin University in 1997 and 1999, respectively, and her PhD degree in Engineering from Osaka University in 2004. From 2002 to 2004, she was a Research Associate of the Graduate School of Engineering, Osaka University. From 2004 to 2006, she was a Researcher of the National Institute of Information and Communications Technology. Since 2006, she has been a researcher in the Faculty of Technology, Bielefeld University. Her research interests are cognitive development robotics and human-robot interactions.

Minoru Asada received this PhD degree in Control Engineering from Osaka University in 1982. From 1982 to 1988, he was a Research Associate in Control Engineering, Osaka University. He became an Associate Professor in Mechanical Engineering for Computer-Controlled Machinery, Osaka University in 1989 and a Professor at the some Department in 1995. Since 1997, he has been a Professor of the Department of Adaptive Machine Systems, Osaka University. From 1986 to 1987, he was a Visiting Researcher at Center for Automation Research, University of Maryland. He received the 1992 best paper award of the IEEE/RSJ International Conference on Intelligent Robots and Systems and the 1996 best paper award of the Robotics Society of Japan. In 2001, he received the Commendation by the Minister of Education, Culture, Sports, Science and Technology, Japanese Government as a person of distinguished services to enlightening people on science and technology. He has been the President of the International RoboCup Federation since 2002 and a IEEE Fellow since 2005.

Koh Hosoda received his PhD degree in Engineering from Kyoto University in 1993. From 1993 to 1997, he was a Research Associate of the Department of Mechanical Engineering for Computer-Controlled Machinery, Osaka University. Since 1997, he has been an Associate Professor of the Department of Adaptive Machine Systems, Osaka University. From 1998 to 1999, he was a Guest Professor in the AI Laboratory, Department of Computer Science, University of Zurich.