*Connection Science, Vol. 15, No. 4, December 2003, 211–229*

Taylor & Francis
Taylor & Francis Group

# A constructive model for the development of joint attention

Yukie Nagai*, Koh Hosoda*†, Akio Morita* and Minoru Asada*†
**Department of Adaptive Machine Systems,*

†*Handai Frontier Research Centre, Graduate School of Engineering, Osaka University, 2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan*
email: yukie@er.ams.eng.osaka-u.ac.jp
tel: +81-6-6879-7349
fax: +81-6-6879-7348

*Abstract.* This paper presents a constructive model by which a robot acquires the ability of joint attention with a human caregiver based on its embedded mechanisms of visual attention and learning with self-evaluation. The former is to look at a salient object in the robot's view, and the latter is to learn sensorimotor co-ordination when visual attention has succeeded. Since the success of visual attention does not always correspond to the success of joint attention, the robot has incorrect learning data for joint attention as well as correct data. However, the robot is expected statistically to lose incorrect data as outliers since such data do not have any correlation in the sensorimotor co-ordination while correct data have a correlation. The robot consequently acquires the ability of joint attention by finding the correlation in the sensorimotor co-ordination even if multiple objects are placed at random positions in an environment and a human caregiver does not provide any task evaluation to the robot. The experimental results show that the proposed model makes the robot reproduce the developmental process of infants' joint attention. Therefore, the proposed model could be one of the models to explain how infants develop the ability of joint attention.

*Keywords*: joint attention, development, visual attention, learning with self-evaluation, constructivist approach.

## 1. Introduction

Human infants acquire various and complicated cognitive functions through interactions with their environments during the first few years. However, the cognitive developmental processes of infants are not completely revealed. A number of researchers (Bremner 1994, Elman *et al.* 1996, Johnson 1997) in cognitive science and neuroscience have attempted to understand infants' development. Their behavioural approaches have explained the phenomena of infants' development. However, the developmental mechanisms are still not clear. In contrast, constructivist approaches have potential to reveal the cognitive developmental mechanisms of infants. It has been suggested in robotics that the building of human-like intelligent robots based on evidence about infants could lead researchers to an understanding of the mechanisms of infants' development (Brooks *et al.* 1998, Asada *et al.* 2001)

*Connection Science* ISSN 0954-0091 Print/ISSN 1360-0494 online 
http://www.tandf.co.uk/journals
DOI: 10.1080/09540090310001655101

Joint attention with a caregiver is one of the abilities that help infants to develop their social cognitive functions (Scaife and Burner 1975, Moore and Dunham 1995). The process of joint attention is defined as looking at the same object that someone else is looking at. The ability of joint attention enables infants to acquire various kinds of social capabilities, e.g. language communication (Morales *et al.* 1998), mind reading (Baron-Cohen 1995) and so on, through interactions with caregivers. A number of researchers in cognitive developmental science have explained the developmental process of infants' joint attention (Moore and Dunham 1995). Butterworth and Jarrett (1991) have investigated how infants develop the ability of joint attention through interactions with their caregivers. They suggested that infants acquire the ability through three developmental stages from 6 to 18 months old. The developmental phenomena of infants' joint attention have been explained through these studies. However, the developmental mechanisms have not yet been revealed.

Robotics researchers (Breazeal and Scassellati 2000, Kozima and Yano 2001, Imai *et al.* 2001, Scassellati 2002) have built the mechanisms of joint attention for their robots. They have investigated how the robots acquire social cognitive functions or realize social communication based on joint attention. However, the behaviours of joint attention of the robots were fully programmed by designers in advance; in other words, how the robots acquire such an ability through interactions with their environments was not discussed. In contrast, Fasel *et al.* (2002) presented a developmental model of joint attention based on a proper interaction of innate motivations and contingency learning. However, the validity of their model has not been verified through implementation in an artificial agent. Nagai *et al.* (2002) proposed a constructive model by which a robot learns joint attention through interactions with a human caregiver. They showed that a robot was able to acquire the ability of joint attention based on task evaluation from a caregiver, and the learning process became more efficient owing to the development of the robot's and the caregiver's internal mechanisms. However, human infants do not always seem to be provided task evaluation from caregivers, and the intention of their study was not to explain the staged developmental process of infants' joint attention.

This paper presents a constructive model that enables a robot to acquire the ability of joint attention without any task evaluation from a human caregiver and to reproduce the staged developmental process of infants' joint attention. The proposed model consists of the robot's embedded mechanisms: visual attention and learning with self-evaluation. The former is to find and look at a salient object in the robot's view, and the latter is to evaluate the success of visual attention and then to learn sensorimotor co-ordination. Since the success of visual attention does not always correspond to the success of joint attention, the robot has incorrect learning data for joint attention as well as correct data. However, the robot is expected statistically to lose incorrect data as outliers because such data do not have any correlation in the sensorimotor co-ordination, while correct data have a correlation. As a result, the robot acquires the ability of joint attention by finding the correlation in the sensorimotor co-ordination even if multiple objects are placed at random positions in an environment and the caregiver does not provide any task evaluation to the robot. In addition, it is expected that the robot reproduces the staged developmental process of infants' joint attention by shifting the attention mechanism from the embedded one, which is the mechanism of visual attention, to the acquired one, which is the sensorimotor co-ordination for joint attention. To the best of our knowledge, this study is the first to propose a computational model to explain the developmental mechanisms of infants' joint attention and to demonstrate that the ability of joint attention could develop without any external evaluation.

The rest of the paper is organized as follows. First, the developmental process of infants' joint attention suggested in cognitive developmental science is explained.

Then, the proposed constructive model by which a robot acquires the ability of joint attention based on visual attention and learning with self-evaluation is described. The learning process through which a robot finds a correlation in its sensorimotor co-ordination for joint attention is explained using a simplified example. Next, some experiments that verify the validity of the proposed model are shown. Finally, there is a discussion.

## 2. The staged developmental process of infants' joint attention

Butterworth and Jarrett (1991) have investigated how human infants develop the ability of joint attention through interactions with their caregivers. They examined what kinds of cues infants utilized for selecting the target to be gazed at in an experimental environment where an infant and his/her caregiver were seated face-to-face and several objects were placed around them. Their observational experiments found that infants develop the ability of joint attention through three stages, as shown in figure 1.

(1) *Ecological stage at 6 to 9 months old:* In the first stage, an infant does not have the ability of joint attention. The infant shows a tendency to look at an interesting object in its field of view regardless of the direction of the caregiver's gaze (see figure 1(a)). Even if the caregiver is looking at the far object, the infant prefers to look at a salient object, such as a moving one, in the field of its view because salient objects attract the interest of the infant.

(2) *Geometric stage at 12 months old:* In the second stage, an infant comes to realize joint attention. The infant in this stage tracks the direction of the caregiver's gaze and looks at the same object that the caregiver is looking at (see figure 1(b)). However, even in this stage, the infant exhibits the gaze-following only when the object is observed within the field of the infant's view.

(3) *Representational stage at 18 months old:* In the final stage, an infant acquires the ability of joint attention. The infant is able to turn around along the direction of the caregiver's gaze and to identify the object that the caregiver is looking at even if the object is outside the field of the infant's first view (see figure 1(c)).

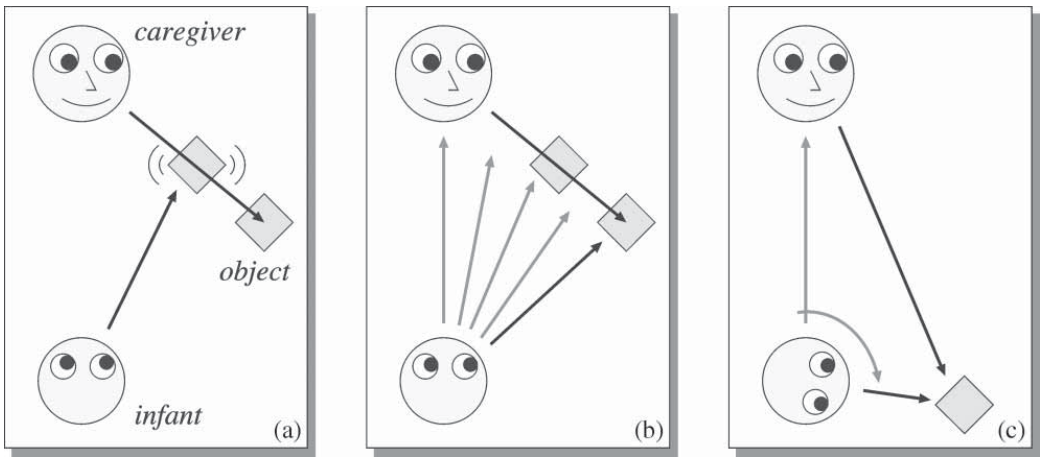



Figure 1. The staged developmental process of infants' joint attention: (a) in the ecological stage, an infant shows a tendency to look at an interesting object regardless of the direction of the caregiver's gaze; (b) in the geometric stage, the infant comes to realize joint attention only when the object that the caregiver is looking at is observed in the field of the infant's view; and (c) in the representational stage, the infant acquires the ability of joint attention and realizes it even if the object that the caregiver is looking at is outside the infant's view.

The developmental phenomena of infants' joint attention have been explained in this way; however, the developmental mechanisms have not been revealed yet. Furthermore, any computational model to explain the development of infants' joint attention cannot be found.

## 3. A constructive model for the development of joint attention

### 3.1. *The task definition of joint attention*

An environmental set-up for joint attention is shown in figure 2, in which a robot with two cameras, a human caregiver and multiple salient objects, which have a bright colour, an intricate pattern, or a motion, are indicated. The environment is not structured, in other words, the positions of objects change randomly every trial. The caregiver looks at one object (in figure 2, he/she is looking at a square object) and changes the object to be gazed at every trial. The robot obtains its camera image $\mathbf{I}$ and the angle of its camera head $\boldsymbol{\theta} = [\theta_{pan}, \theta_{tilt}]$ as inputs, and outputs a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan}, \Delta\theta_{tilt}]$ to rotate the camera head. The joint attention task in this situation is defined as a process by which the robot outputs a motor command $\boldsymbol{\Delta\theta}$ based on sensor inputs $\mathbf{I}$ and $\boldsymbol{\theta}$, and consequently looks at the same object that the caregiver is looking at. The robot is required to acquire the sensorimotor co-ordination to realize joint attention through learning.

### 3.2. *The proposed model*

The proposed constructive model for the development of joint attention is shown in figure 3. As described earlier, a robot obtains its camera image $\mathbf{I}$ and the angle of its camera head $\boldsymbol{\theta}$ as inputs and outputs a motor command $\boldsymbol{\Delta\theta}$ to rotate the camera head. In the model, the robot has the following mechanisms:

(a) *visual attention*, which consists of a *salient feature detector* and *a visual feedback controller*, and has the capability to find and gaze at a salient object in the robot's current view;

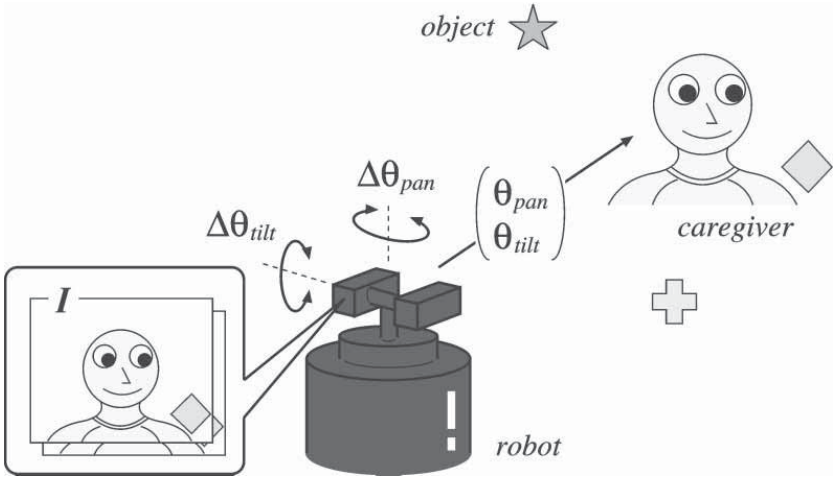



Figure 2. An environmental set-up for joint attention between a robot and a human caregiver. The environment includes multiple salient objects, and the robot and the caregiver, who is looking at one object, are seated face-to-face. The joint attention task in this situation is defined as a process by which the robot outputs a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan}, \Delta\theta_{tilt}]$ based on sensor inputs $\mathbf{I}$ and $\boldsymbol{\theta} = [\theta_{pan}, \theta_{tilt}]$, and consequently looks at the same object that the caregiver is looking at.

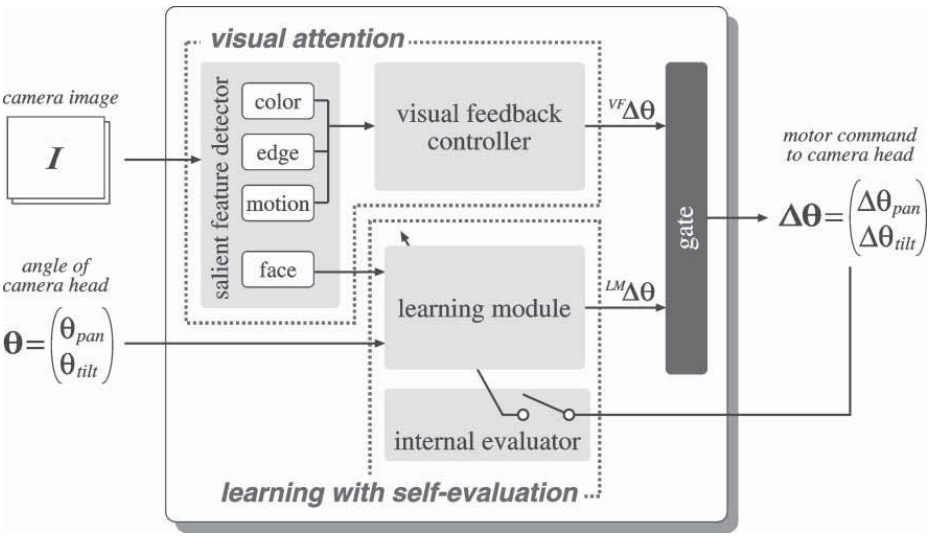



Figure 3. The proposed constructive model for the development of joint attention. A robot obtains its camera image **I** and the angle of the camera head **θ** as inputs, and outputs a motor command **Δθ** to rotate the camera head. The model includes the mechanisms of visual attention, which consists of a salient feature detector and a visual feedback controller, and learning with self-evaluation, which consists of a learning module and an internal evaluator. The former generates an output to look at a salient object in the robot's view, and the latter learns sensorimotor co-ordination when visual attention has succeeded. The gate module makes a choice between outputs from the visual feedback controller ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ and the learning module ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$.

(b) *learning with self-evaluation*, which consists of *a learning module* and *an internal evaluator* and has the capability to evaluate the success of visual attention and then to learn the sensorimotor co-ordination;
(c) *gate*, which makes a choice between an output from the visual feedback controller and an output from the learning module according to a selecting rate.

Based on the embedded mechanisms, the robot acquires the sensorimotor co-ordination for joint attention in the learning module through the following process.

(1) The robot first looks at the caregiver who is looking at an object and obtains the sensor inputs **I** and **θ**.
(2) If a salient object is observed in **I**, the robot detects the object by the salient feature detector and then generates a motor command ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ to look at the object by the visual feedback controller.
(3) At the same time, the robot generates a motor command ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$ by the learning module based on the inputs of the caregiver's face image, which has been detected by the salient feature detector, and the angle of the camera head **θ**.
(4) The gate makes a choice between ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ and ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$ according to the selecting rate that is designed to select mainly the former one at the beginning of learning and gradually to come to select the latter one as learning advances. The robot outputs the selected motor command as **Δθ** ( $=$ ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ or ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$).
(5) After the motor output, if an object is observed in the centre of the robot's camera image, the robot evaluates the success of visual attention by the internal evaluator and triggers learning in the learning module.

(6) The robot learns the sensorimotor co-ordination in the learning module by backpropagation using the output $\mathbf{\Delta\theta}$ when visual attention has succeeded as a reference.
(7) Repeat the above.

It should be noted here that the success of visual attention does not always correspond to the success of joint attention. Because the environment includes multiple salient objects as shown in figure 2, the robot mostly looks at a different object from that at which the caregiver is looking, based on the mechanism of visual attention and looks at the same object only at a chance level. Therefore, the robot has two kinds of learning data: correct learning data for joint attention and incorrect data.

- In the former case, i.e. when the robot has looked at the same object that the caregiver is looking at, the robot can find a correlation for joint attention in the sensorimotor co-ordination of the learning module. The reason is that the position of the object that the robot as well as the caregiver are looking at is uniquely determined by the image of the caregiver's face.
- In the latter case, i.e. when the robot has looked at a different object from that at which the caregiver is looking, the robot cannot find any correlation in the sensorimotor co-ordination of the learning module. The reason is that the position of the object that the robot has looked at does *not* uniquely correspond to the image of the caregiver's face.

As a result, the incorrect learning data in the latter case would be expected to be statistically lost as outliers through learning. In contrast, the correlation in the sensorimotor co-ordination that has been acquired when joint attention succeeded is relatively enhanced. The process of finding the correlation for joint attention in the sensorimotor co-ordination is illustrated in the Appendix using a simple example. This demonstrates that the proposed model enables the robot to acquire the ability of joint attention without any task evaluation from the caregiver. Furthermore, the robot is expected to improve the acquired sensorimotor co-ordination by increasing the selecting rate of an output from the learning module as learning advances. Applying an output from the learning module that has already acquired the sensorimotor co-ordination for joint attention enables the robot to have correct learning data at a higher probability. As a result, the robot is able to acquire more appropriate sensorimotor co-ordination for joint attention. This also changes the robot's behaviour. It is expected that the change of the selecting rate of outputs in the gate module makes the robot shift its behaviour from visual attention, which is embedded, to joint attention, which is acquired through learning.

The following sections explain the modules in the proposed model: the salient feature detector, the visual feedback controller, the internal evaluator, the learning module and the gate, in order.

3.2.1. *Salient feature detector.* The salient feature detector extracts distinguishing image areas from $\mathbf{I}$ by colour, edge, motion and face detectors. The colour, edge and motion detectors extract objects ($i = 1, \ldots, n$) that have a bright colour, an intricate pattern and a motion, respectively. Then, the salient feature detector selects the most interesting object $i_{\mathrm{trg}}$ among the extracted objects by comparing the sum of the interests of all features.

$$i_{\mathrm{trg}} = \arg\max_i (\alpha_{\mathrm{c}} f_i^{\mathrm{col}} + \alpha_{\mathrm{e}} f_i^{\mathrm{edg}} + \alpha_m f_i^{\mathrm{mot}}), \tag{1}$$

where $f_i^{\mathrm{col}}, f_i^{\mathrm{edg}}$ and $f_i^{\mathrm{mot}}$ indicate the size of the coloured area, the complexity of the pattern and the amount of the motion of the object $i$, respectively. The coefficients ($\alpha_{\mathrm{c}}$, $\alpha_{\mathrm{e}}$, $\alpha_{\mathrm{m}}$) denote the degrees of the interests in three features, which are determined based on the context or the characteristics of the robot. This mechanism makes the robot randomly change the object to be gazed at every trial. At the same time, the face detector extracts face-like stimuli of the caregiver by template matching. The detection of face-like stimuli is a fundamental ability for social agents, therefore it should be treated in the same manner as the detection of the primitive features. The detected primitive features of the object $i_{\mathrm{trg}}$ and the face-like one of the caregiver are sent to the visual feedback controller and the learning module, respectively.

3.2.2. *Visual feedback controller.* The visual feedback controller receives the detected image feature of the object $i_{\mathrm{trg}}$ and then generates a motor command ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ for the camera head to gaze at the object. First, this controller calculates the position $(x_i, y_i)$ of the object $i_{\mathrm{trg}}$ in the camera image and then generates a motor command ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ as

$$ {}^{\mathrm{VF}}\mathbf{\Delta\theta} = \begin{pmatrix} {}^{\mathrm{VF}}\Delta\theta_{\mathrm{pan}} \\ {}^{\mathrm{VF}}\Delta\theta_{\mathrm{tilt}} \end{pmatrix} = g\begin{pmatrix} x_i - cx \\ y_i - cy \end{pmatrix}, \tag{2} $$

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

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

3.2.3. *Internal evaluator.* The other embedded mechanism, which is learning with self-evaluation, is realized by the internal evaluator and the learning module.

The internal evaluator detects the success of visual attention when

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

where $d_{\mathrm{th}}$ denotes a threshold for evaluating whether the robot looks at an object in the centre of the camera image or not. If an object is observed in the centre of the camera image, the internal evaluator triggers learning processing in the learning module. Note that the internal evaluator does not know the success of joint attention but knows the success of visual attention.

3.2.4. *Learning module.* The learning module consists of a three-layered neural network, shown in figure 4. In the forward processing, this module receives the image of the caregiver's face and the angle of the camera head $\mathbf{\theta}$ as inputs, and outputs ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$ as a motor command. The caregiver's face image, which is input as the value of brightness of each pixel, is utilized to estimate the motor command ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$ to follow the caregiver's gaze direction. The angle of the camera head $\mathbf{\theta}$ is used to rotate the camera head incrementally because the caregiver's attention cannot be narrowed down to a particular point along the line of the caregiver's gaze. The generated motor command ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$ is sent to the gate module as an output from the learning module.

In the learning processing, when the learning module is triggered by the internal evaluator, the module learns sensorimotor co-ordination by backpropagation using the output $\mathbf{\Delta\theta}$ when visual attention has succeeded as a reference. As mentioned above,

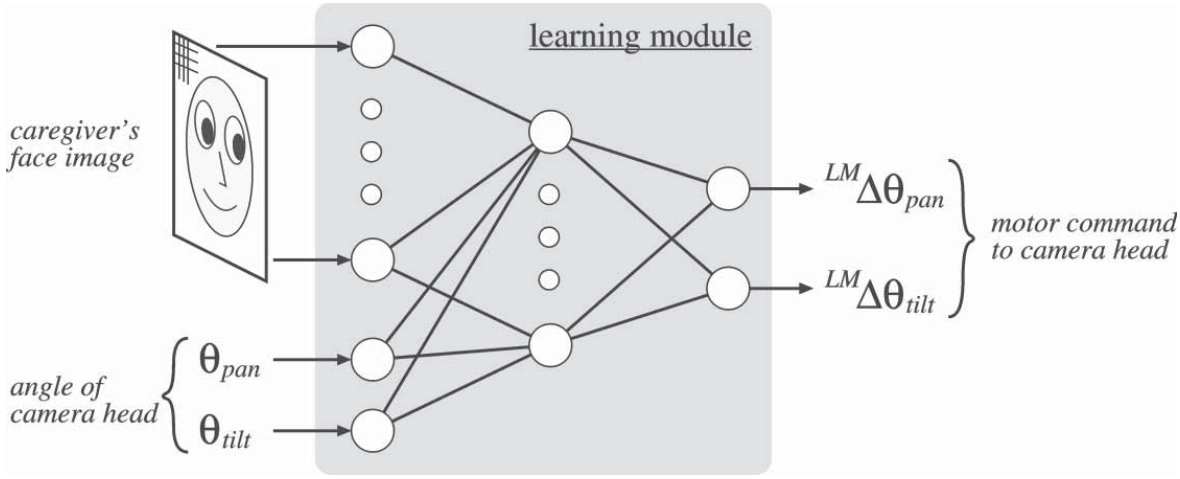



Figure 4. The learning module which consists of a three-layered neural network. The module receives a caregiver's face image and the angle of the camera head $\boldsymbol{\theta}$ as inputs and outputs a motor command ${}^{\mathrm{LM}}\boldsymbol{\Delta\theta}$. When the module is triggered by the internal evaluator, the module learns sensorimotor co-ordination by backpropagation using the output $\boldsymbol{\Delta\theta}$ when visual attention has succeeded as a reference.

the internal evaluator triggers learning according to the success of visual attention, not joint attention; therefore, this module has not only correct learning data for joint attention but also incorrect data. In the former case, the learning module is able to acquire a correlation in the sensorimotor correlation for joint attention. In contrast, in the latter case the module is not able to find any correlation because the inputs and the outputs do not have a unique correspondence but have a random correspondence. Thus, the learning module is expected statistically to lose incorrect data as outliers as described above and to enhance relatively the correlation in the sensorimotor co-ordination acquired when joint attention has succeeded. As a result, the robot acquires the ability of joint attention in the learning module without any task evaluation from the caregiver.

3.2.5. *Gate.* The gate module arbitrates a motor command $\boldsymbol{\Delta\theta}$ between ${}^{\mathrm{VF}}\boldsymbol{\Delta\theta}$ from the visual feedback controller and ${}^{\mathrm{LM}}\boldsymbol{\Delta\theta}$ from the learning module. The gate module sets a gating function to define the selecting rate of the outputs. At the beginning of learning, the selecting rate of ${}^{\mathrm{VF}}\boldsymbol{\Delta\theta}$ is set to a higher probability than that of ${}^{\mathrm{LM}}\boldsymbol{\Delta\theta}$ because the learning module has not yet acquired appropriate sensorimotor co-ordination for joint attention. On the other hand, in the later stage of learning, the output ${}^{\mathrm{LM}}\boldsymbol{\Delta\theta}$ from the learning module, which has acquired the sensorimotor co-ordination for joint attention, becomes more probable for selection. This gate module enables the robot to increase the proportion of correct learning data as learning advances and consequently to acquire more appropriate sensorimotor co-ordination for joint attention in the learning module. The experiments presented in this paper use a sigmoid function for the selecting rate, which is defined by a designer in advance.

### 3.3. *The staged learning process of joint attention*

It is expected that the proposed model makes the robot acquire the ability of joint attention through a staged learning process. Figure 5 represents the transition of the robot's behaviour through three stages. In each stage, the behaviour of the robot is represented as the change of its camera image when the robot shifts its gaze direction based on the

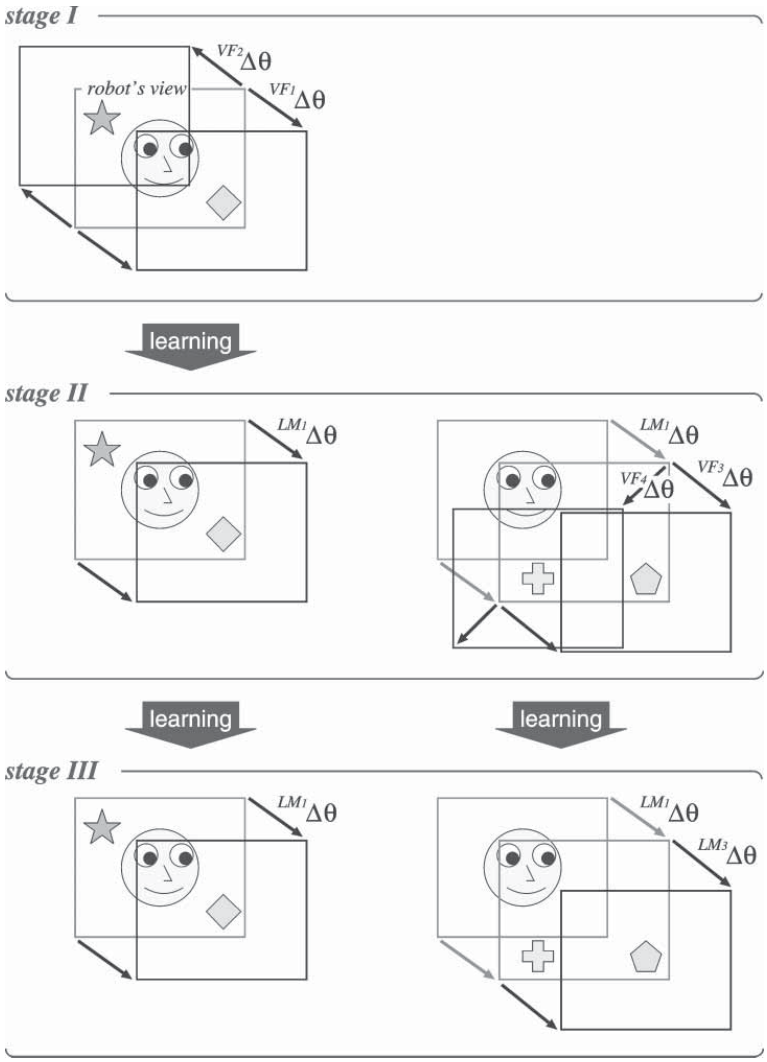



Figure 5. The staged learning process of the robot's joint attention. This figure shows the transition of the robot's behaviour, which is shown as the change of the robot's camera image, through the learning stages I, II and III. In stage I, the robot has a tendency to look at an interesting object in the field of the robot's view regardless of the direction of the caregiver's gaze since the gate module mainly selects the output $^{VF}\mathbf{\Delta\theta}$. In stage II, the robot realizes joint attention by generating the output $^{LM}\mathbf{\Delta\theta}$ only when the object that the caregiver is looking at is observed in the field of the robot's first view. In stage III, the robot realizes joint attention by generating the output $^{LM}\mathbf{\Delta\theta}$ incrementally even if the object is outside the robot's view. This staged learning process is considered to be equivalent to the developmental process of infants' joint attention shown in figure 1.

output $^{VF}\mathbf{\Delta\theta}$ or $^{LM}\mathbf{\Delta\theta}$. In the figure, a rectangle indicates a camera image of the robot, and arrows which connect the corners of two rectangles show a motor output of the robot.

- *Stage I:* In the first stage of learning, the robot has a tendency to look at an interesting object in the field of the robot's view based on the embedded mechanism of visual attention since the gate module mainly selects $^{VF}\mathbf{\Delta\theta}$ as a robot's motor command. At the top of figure 5, the robot outputs $^{VF_1}\mathbf{\Delta\theta}$ or $^{VF_2}\mathbf{\Delta\theta}$ case by case and looks at one object in the centre of the camera image regardless of the direction of the caregiver's gaze. At the same time, the robot starts to learn the sensorimotor co-ordination in each case.

- *Stage II:* In the middle stage of learning, the robot is able to realize joint attention only when the object that the caregiver is looking at is observed in the field of the robot's first view. At the middle left of figure 5, the robot looks at the same object that the caregiver is looking at based on the output ${}^{\mathrm{LM}_1}\mathbf{\Delta\theta}$ from the learning module that has acquired the sensorimotor co-ordination for joint attention in stage I. At the middle right of figure 5, if the object that the caregiver is looking at is outside the field of the robot's first view, the robot can find the object not at the centre of the camera image but at the periphery by generating ${}^{\mathrm{LM}_1}\mathbf{\Delta\theta}$. Then, if several objects are observed in the camera image, the robot outputs ${}^{\mathrm{VF}_3}\mathbf{\Delta\theta}$ or ${}^{\mathrm{VF}_4}\mathbf{\Delta\theta}$ to look at the most interesting object case by case. When visual attention has succeeded, the robot learns the sensorimotor co-ordination in each case as well as stage I.
- *Stage III:* In the final stage, the robot has acquired the complete ability of joint attention owing to learning in stages I and II. At the bottom of figure 5, the robot can identify the object that the caregiver is looking at by generating ${}^{\mathrm{LM}_1}\mathbf{\Delta\theta}$ and ${}^{\mathrm{LM}_3}\mathbf{\Delta\theta}$ incrementally even if the object is not observed in the field of the robot's first view. The sensorimotor co-ordinations of ${}^{\mathrm{LM}_1}\mathbf{\Delta\theta}$ and ${}^{\mathrm{LM}_3}\mathbf{\Delta\theta}$ have been acquired through learning in stages I and II because they had a correlation in the sensorimotor co-ordination.

The staged learning process of the robot's joint attention can be regarded as equivalent to the staged developmental process of infants' joint attention shown in figure 1. Stages I, II and III of the robot correspond to infants at 6–9, 12 and 18 months old, respectively. The next section verifies the learning process of the robot's joint attention through experiments.

## 4. Experiment

### 4.1. *An experimental set-up*

It was examined whether an actual robot can acquire the ability of joint attention based on the proposed model without any task evaluation from a human caregiver in an unstructured environment, including multiple objects. An experimental environment is shown in figure 6(a), and the robot's camera image in this situation is shown in figure 6(b). In the environment, several objects with a bright colour are placed randomly around the robot and the caregiver. The caregiver looks at one object that is randomly selected every trial (in figure 6, she is looking at the object in her hand). The robot has two cameras, which rotate on the pan and the tilt axes, and detect the caregiver's face and the objects by the salient feature detector from the camera image as shown in figure 6(b). The rectangle in the left image in figure 6(b) shows the position of the caregiver's face detected by template matching, and the highlighted areas in the right image show the objects with a bright colour extracted by using thresholds in colour space. The detected image of the caregiver's face is input to the learning module as the value of brightness of each pixel, and the robot learns its sensorimotor co-ordination for joint attention in the learning module.

The experiment presented in this section applied the following parameters. The degrees of the robot's interests in image features in equation (1) were defined as $(\alpha_c, \alpha_e, \alpha_m) = (1, 0, 0)$; in other words, the robot was designed to prefer to look at an object that had a bright colour and a larger size in the camera image. Note that the caregiver did not know the preference of the robot. The threshold $d_{\mathrm{th}}$ in equation (3) for the determination of the success of visual attention was defined as $d_{\mathrm{th}} = W_x/6$, where $W_x$ denotes the width of the robot's camera image. Under these conditions, learning data sets were acquired in the real environment in advance, and then offline learning was conducted. Each data set included:

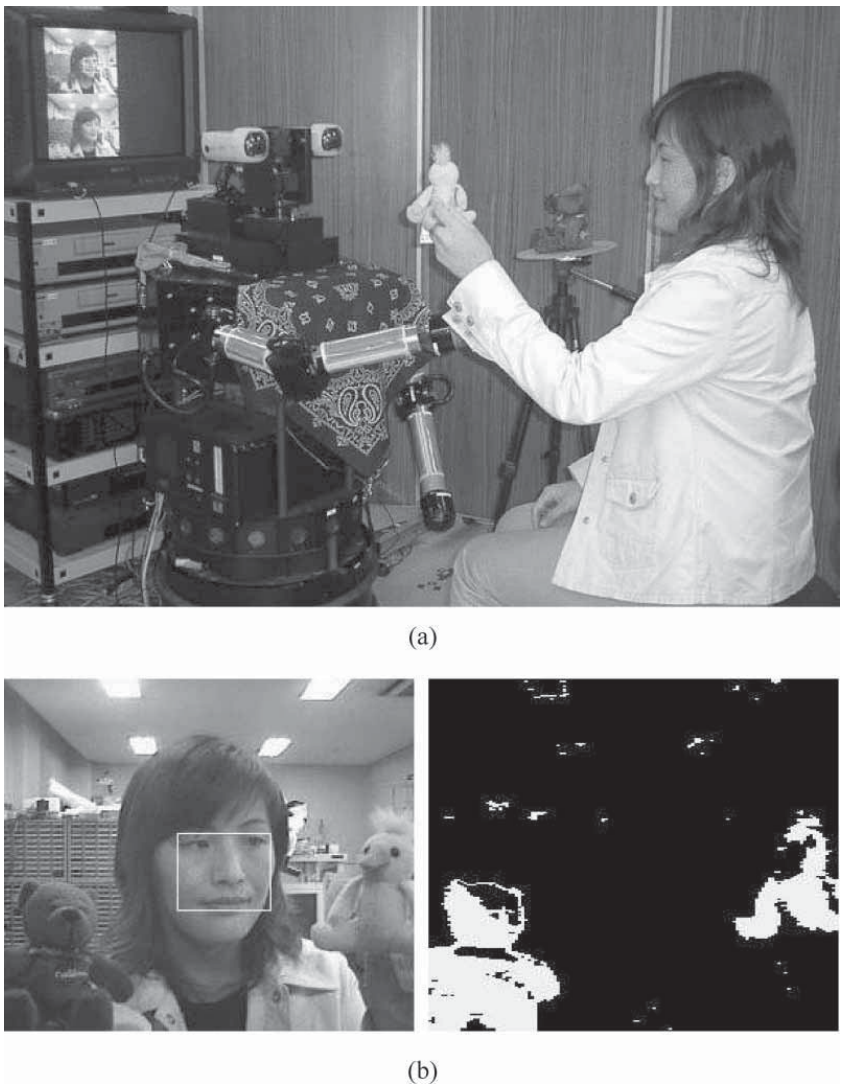


Figure 6. An experimental set-up for joint attention. (a) An experimental environment for joint attention in which a robot with two cameras, a human caregiver and multiple salient objects are shown. The objects are randomly placed every trial, and the caregiver looks at one object that has been selected at random. The robot first looks at the caregiver and captures its camera image as shown in (b). (b) The robot's camera image acquired in situation (a). The rectangle in the left image shows the position of the caregiver's face detected by template matching, and the highlighted areas on the right show the objects with a bright colour extracted by using thresholds in colour space. This processing is conducted by the salient feature detector in the proposed model.

- *input data:* a left camera image **I**, in which the caregiver's face was extracted as a window of size 30 × 25 [pixels], and the angles of the camera head $\boldsymbol{\theta} = [\theta_{pan}, \theta_{tilt}]$ when the robot was looking at the caregiver's face;
- *output data when joint attention succeeded:* a motor command $\boldsymbol{\Delta\theta}$ for the camera head to shift the robot's gaze direction from the caregiver's face to the object that the caregiver was looking at;
- *output data when joint attention failed while visual attention succeeded:* motor commands $\boldsymbol{\Delta\theta}$ to look at different objects from those which the caregiver was looking at, only these data were obtained in a simulation.

One hundred and twenty-five data sets were randomly utilized in offline learning. The robot learned its sensorimotor co-ordination in the learning module by backpropagation using the above input data and either of two kinds of output data as a reference. The number of units in the learning module was set as 752 $(30 \times 25 + 2)$ for the input units, seven for the hidden units and two for the output units. The number of hidden units was determined based on preliminary experiments.

### 4.2. *The change of the task performance*

It was verified how the task performance of joint attention changed depending on the number of objects over learning. The gating function to select a motor output in the gate module was defined as a sigmoid one, shown in figure 7(a). The horizontal axis denotes the learning time step, and the vertical one denotes the selecting rate of the output ${}^{\mathrm{LM}}\mathbf{\Delta\theta}$ from the learning module. The output ${}^{\mathrm{VF}}\mathbf{\Delta\theta}$ from the visual feedback controller was selected at the residual rate. This gating function was designed based on preliminary experiments.

Figure 7(b) shows the changes of the success rates of joint attention over learning, in which the number of objects is set to one, three, five, or ten. The case whose number of object equals one means that the robot always learns correct sensorimotor co-ordination for joint attention. In contrast, the case of ten means that the robot has correct learning data only at 1/10 proportion at the beginning of learning. However, the robot is expected to increase the proportion of correct data by adapting outputs from the learning module that has already acquired sensorimotor co-ordination for joint attention. From this result, it can be found that the success rates of joint attention are at chance levels at the beginning of learning; however, they increase to high performance at the end although the environment includes multiple objects. In the case that the number of objects is set to five, the success rate of joint attention improves from 20%, which is just a chance level, to 85%.

### 4.3. *The staged learning process of joint attention*

It was investigated how the robot changed its behaviour through the learning process. We focused on the result when the number of objects was set to five in figure 7(b) and examined the robot's behaviour in the three stages I, II and III, whose learning periods were 2–5, 20–23 and 45–48 [ $\times 10^4$], respectively.

Figure 8 shows the pan angle of the robot's camera head when it realized visual attention, in which '○' and '×' indicate the success of joint attention and the failure, respectively. In other words, the former means that the robot has looked at the *same* object as the caregiver, while the latter means that the robot has looked at a *different* object. Note that objects also exist at the positions of those areas that do not include any mark. The pan angle of the robot's camera head is $0^\circ$ when the robot is looking at the caregiver, and the view range of the robot is $\pm 18^\circ$. In other words, the objects found within $\pm 18^\circ$ are observed in the field of the robot's view when the robot is looking at the caregiver. From this result, we can see that the number of successes of joint attention increases as learning advances, and at the same time the range of the camera angle when the robot has realized joint attention gradually exceeds the range of $\pm 18^\circ$. The reason the robot seldom or never achieves visual attention or joint attention over $\pm 18^\circ$ in stage I is that the robot in this stage mainly selects the output from the visual feedback controller based on the gating function shown in figure 7(a) and looks at an object in the field of the robot's first view. It is confirmed that this experimental result demonstrates the stages of the learning process of joint attention shown in figure 5. In other words, the learning process of the robot's joint attention based on the proposed model can be regarded as

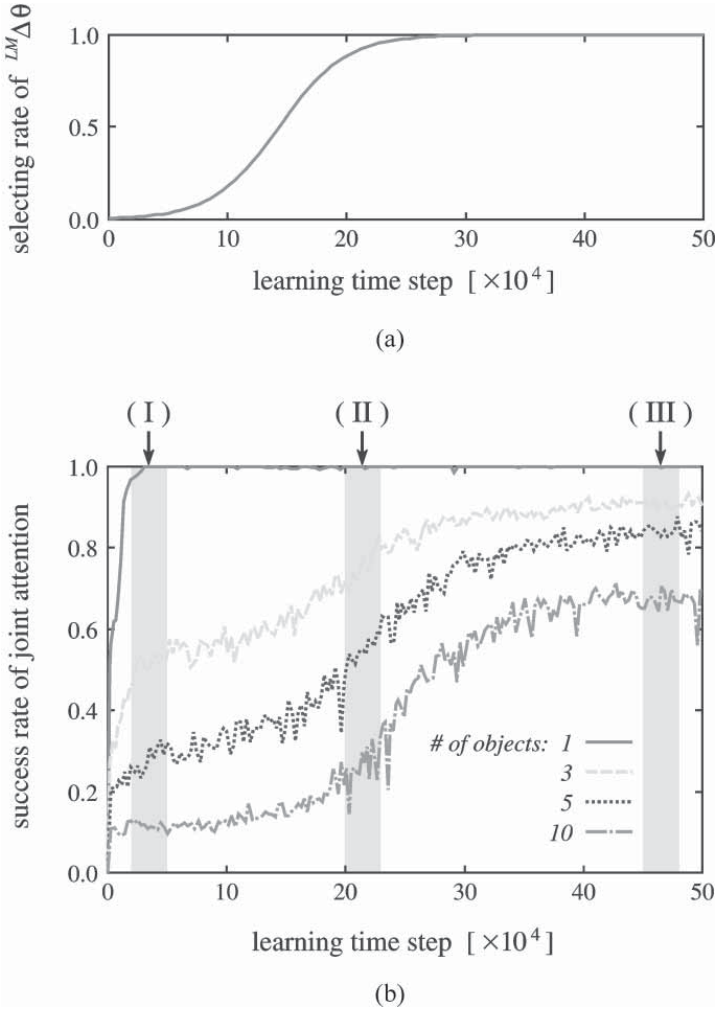



Figure 7. The gating function for selecting a motor output in the gate module and the changes of the success rates of joint attention over learning. (a) The gating function utilized in the experiments. The output $^{LM}\mathbf{\Delta\theta}$ from the learning module was selected at the rate of the value of this sigmoid function, and the output $^{VF}\mathbf{\Delta\theta}$ from the visual feedback controller was selected at the residual rate. (b) The changes of the success rates of joint attention over learning. Each curve shows the result when the number of objects was set to one, three, five, or 10. The success rates of all cases are at chance levels at the beginning of learning; however, they increase to high levels at the end. This result indicates that the proposed model enables a robot to acquire the ability of joint attention without any evaluation from the caregiver even if the environment includes several objects. The behaviour of the robot in the highlighted stages I, II, and III are discussed in figure 8.

equivalent to the developmental process of infants' joint attention shown in figure 1. Each behaviour of the robot's joint attention in stages I, II and III corresponds to the behaviour of infants at 6–9, 12 and 18 months old, respectively.

### 4.4. *Evaluation of the final task performance in a real environment*

Finally, we evaluated the final task performance of the robot in a real environment. The sensorimotor co-ordination in the learning module acquired through offline learning when the number of objects was set to five was implemented in the actual robot shown in figure 6(a). The robot performed based on the implemented learning module and realized joint attention with a human caregiver.

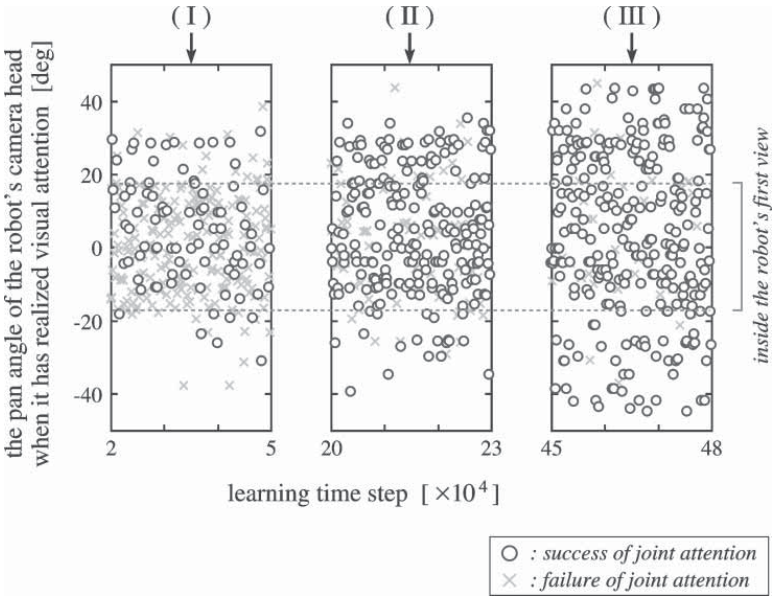
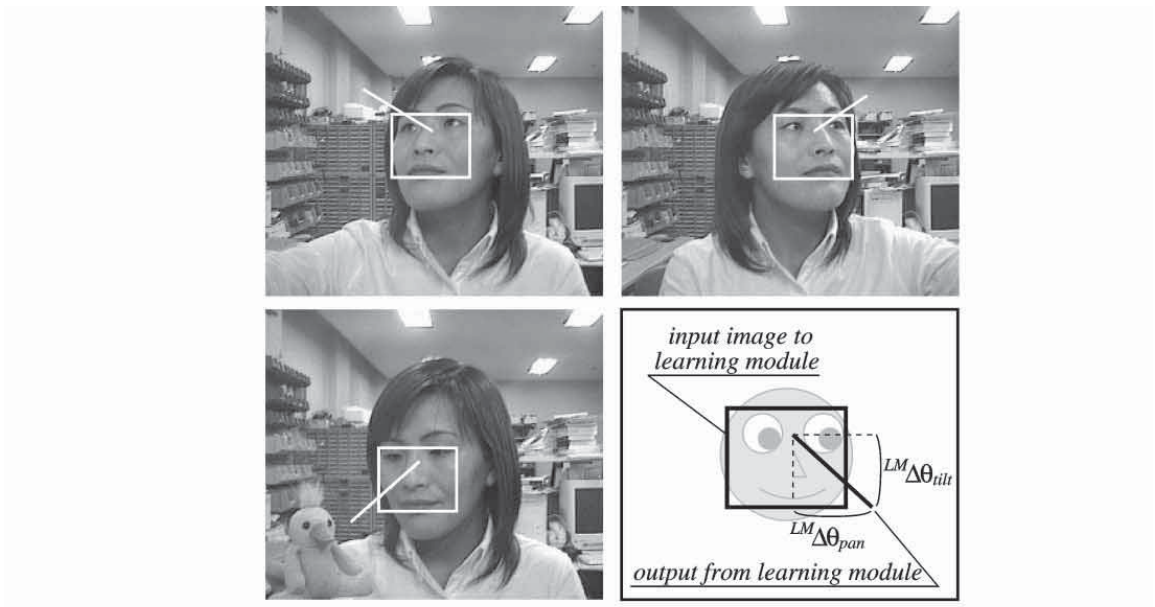
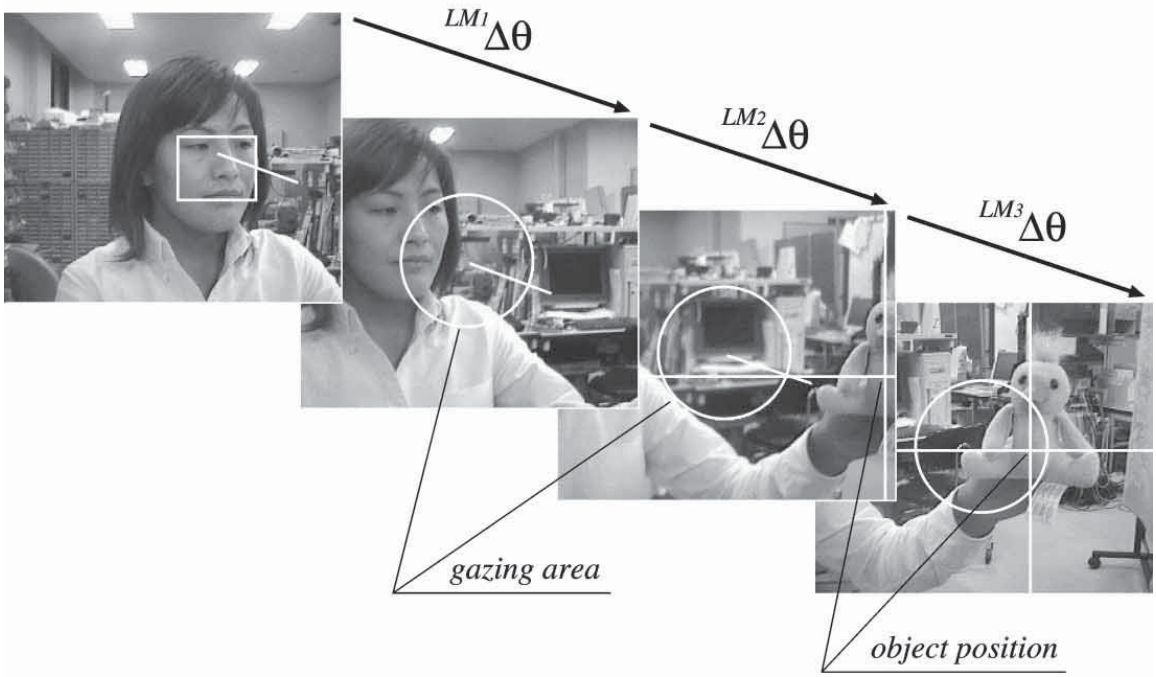



Figure 8. An experimental result of the staged learning process of joint attention. This graph plots the pan angle of the robot's camera head when the robot has achieved visual attention, in which '○' and '×' indicate the success of joint attention and the failure of joint attention, respectively. In stage I, the robot can realize joint attention only at a chance level and has a tendency to look at an object inside the field of the robot's first view. This is because the output from the visual feedback controller is mainly selected by the gate module. In stage II, the robot comes to realize joint attention within the robot's first view. Finally, in stage III the robot realizes joint attention at every position. This learning process of the robot's joint attention is considered to be equivalent to the developmental process of the infants' one shown in figure 1.

Figure 9 shows the experimental results in which the final task performance of the robot is indicated in its camera image. Figure 9(a) shows the robot's camera images when the robot is gazing at the caregiver who is looking at an object at various positions. In each image, a caregiver's face image enclosed in a rectangle indicates the input to the learning module, and a vector on the face shows the output from the module. Note that a vector does not mean the gaze of the caregiver but means the motor command of the robot. In other words, the horizontal component and the vertical one of a vector indicate the pan and the tilt angles of the motor command, respectively. The robot rotates its camera head based on the motor command and tries to find the object that the caregiver is looking at. Figure 9(b) shows the change of the robot's camera image when it shifts its gaze direction from the caregiver's face to the object based on the output from the learning module. The robot generates motor commands by the learning module using the caregiver's face image enclosed in a rectangle in the top-left image and the angle of the camera head until finding any object in the centre of the camera image. A circle and a cross line in each image show the gazing area of the robot and the object's position detected by the salient feature detector, respectively. In this trial, the robot incrementally outputs the motor commands ${}^{\mathrm{LM}_1}\boldsymbol{\Delta\theta}$, ${}^{\mathrm{LM}_2}\boldsymbol{\Delta\theta}$ and ${}^{\mathrm{LM}_3}\boldsymbol{\Delta\theta}$ at each step, and consequently finds the object that the caregiver is looking at. The success rate of joint attention in a real environment was 85% ($=17/20$ [trials]) under the condition that the caregiver was the same person as the learned one, and the objects were set at different positions from the learned ones. From the results of figure 9(a, b), it is confirmed that the learning




(a)




(b)

Figure 9. Experimental results that show the final task performance of the robot in a real environment. Each result shows the robot's camera image, in which the sensorimotor co-ordination acquired in the learning module based on the proposed model is indicated. (a) The robot's camera images in which the sensorimotor co-ordination in the learning module acquired based on the proposed model is indicated. In each image, a caregiver's face image enclosed in a rectangle shows the input to the learning module, and a vector on the face shows the output from the module. Note that a vector does not mean the gaze of the caregiver but means the motor command of the robot. (b) The change of the robot's camera image when the robot shifts its gaze direction from the caregiver's face to the object that the caregiver is looking at based on outputs from the learning module. This result shows that the learning module has acquired adequate ability to realize joint attention in a real environment.

module has acquired the adequate ability to realize joint attention in a real environment based on the proposed model.

## 5. Discussion

This paper has proposed a constructive model for the development of joint attention. The experimental results showed that:

- the proposed model enables a robot to acquire the ability of joint attention without any task evaluation from a caregiver even if the environment includes multiple objects;
- the proposed model makes a robot reproduce the staged developmental process of infants' joint attention.

In cognitive science, the developmental phenomena of infants' joint attention have been explained (e.g. Butterworth and Jarrett 1991), however the developmental mechanisms have not yet been revealed. The proposed model is designed based on knowledge about infants that has been found in cognitive science and demonstrates the developmental process of infants' joint attention. The capabilities embedded in the robot, i.e. visual attention and learning with self-evaluation, have been indicated to be inherent also in human infants (Bremner 1994). The staged learning process of the robot's joint attention shown in figure 8 seems to be similar to the staged developmental process of infants' joint attention shown in figure 1. These similarities in both the mechanism and the developmental phenomenon suggest that the proposed model could be one of the models to explain how infants acquire the ability of joint attention.

In the future, a more efficient learning mechanism should be developed so that real-time learning can be conducted. It is expected that the experiments of real-time learning in a real environment expose true difficulties in the development of communication. To address such difficulties is challenging for the researchers in both cognitive science and robotics. In addition, an adaptive gate module should be designed. Infants are conjectured to shift their behaviours according to the performance of each behaviour. The robot should be designed to shift its behaviours by using any performance of itself, e.g. the success rate of visual attention by the learning module. The acquired learning module should be analysed for explaining how the robot represents the ability of joint attention in the sensorimotor co-ordination. For the purpose of explanation, the learning experiment should be conducted using several caregivers in turn. The learning of various experiences allows the robot to acquire the well-understood ability of joint attention. The explanation of the robot's mechanism will help us to understand the infants' mechanisms of joint attention. To resolve these issues would make the proposed model much more valuable.

## Acknowledgement

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

## References

Asada, M., MacDorman, K.F., Ishiguro, H., and Kuniyoshi, Y., 2001, Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, **37**, 185–193.

Baron-Cohen, S., 1995, *Mindblindness*, (Cambridge MA: MIT Press).

Breazeal, C., and Scassellati, B., 2000, Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, **8** (1): 49–74.

Bremner, J. G., 1994, *Infancy* (Oxford: Blackwell).

Brooks, R. A., Breazeal, C., Irie, R., Kemp, C. C., Marjanović, M., Scassellati. B., and Williamson, M. M., 1998, Alternative essences of intelligence. In *Proceedings of the American Association of Artificial Intelligence*, pp. 961–968.

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

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K., 1996, *Rethinking Innateness: A Connectionist Perspective on Development* (Cambridge, MIT Press).

Fasel, I., Deák, G. O., Triesch, J., and Movellan, J., 2002, Combining embodied models and empirical research for understanding the development of shared attention. In *Proceedings of the 2nd International Conference on Development and Learning* (Boston, USA), pp. 21–27.

Imai, M., Ono, T., and Ishiguro, H., 2001, Physical relation and expression: joint attention for human–robot interaction. In *Proceedings of the 10th IEEE International Workshop on Robot and Human Communication* (Bordeaux and Paris, France).

Johnson, M. H., 1997, *Developmental Cognitive Neuroscience* (Oxford: Blackwell).

Kozima, H., and Yano, H., 2001, A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics* (Lund, Sweden).

Moore, C., and Dunham, P. J. (eds), *Joint Attention: Its Origins and Role in Development* (Hillsdale, NJ: Lawrence Erlbaum).

Morales, M., Mundy, P., and Rojas, J., 1998, Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development*, **21** (2), 373–377.

Nagai, Y., Asada, M., and Hosoda, K., 2002, Developmental learning model for joint attention. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems* (Lausanne, Switzerland), pp. 932–937.

Scaife, M., and Bruner, J. S., 1975, The capacity for joint visual attention in the infant. *Nature*, **253**, 265–266.

Scassellati, B., 2002, Theory of mind for a humanoid robot. *Autonomous Robots*, **12**, 13–24.

## Appendix: The mechanism to find a correlation in sensorimotor co-ordination for joint attention

This appendix illustrates how a robot finds a correlation for joint attention in its sensorimotor co-ordination based on the proposed model with a simple example. Figure A1 shows three situations, each of which includes a robot, a caregiver and two objects. The environment changes in every situation. The robot is expected to acquire the ability of joint attention by finding a correlation in the sensorimotor co-ordination through learning based on the experience of visual attention.

- *Situation I:* First, the environment includes objects 1 and 2. The robot as well as the caregiver have two choices to look at an object based on the mechanism of visual attention. The robot learns its sensorimotor co-ordination based on the mechanism of learning with self-evaluation when visual attention has succeeded. In this situation, the robot acquires the connections of C1:R1 (the caregiver is looking at object 1, and the robot is looking at object 1), C1:R2, C2:R1 and C2:R2. Since the learning of these connections is executed in equal proportions, these connections have equal strengths.
- *Situation II:* Next, the environment changes into including objects 2 and 3. The four connections acquired in situation I are maintained as they are. The robot learns the sensorimotor co-ordination in the same manner as in situation I. In this situation, the robot acquires the connections of C2:R2, C2:R3, C3:R2 and C3:R3. Note that the learning of the connection of C2:R2 is for the second time, therefore the strength of the connection is doubled compared with other connections.
- *Situation III:* Finally, the environment changes into including objects 1 and 3. The robot learns the sensorimotor co-ordination in the same manner as in situations I

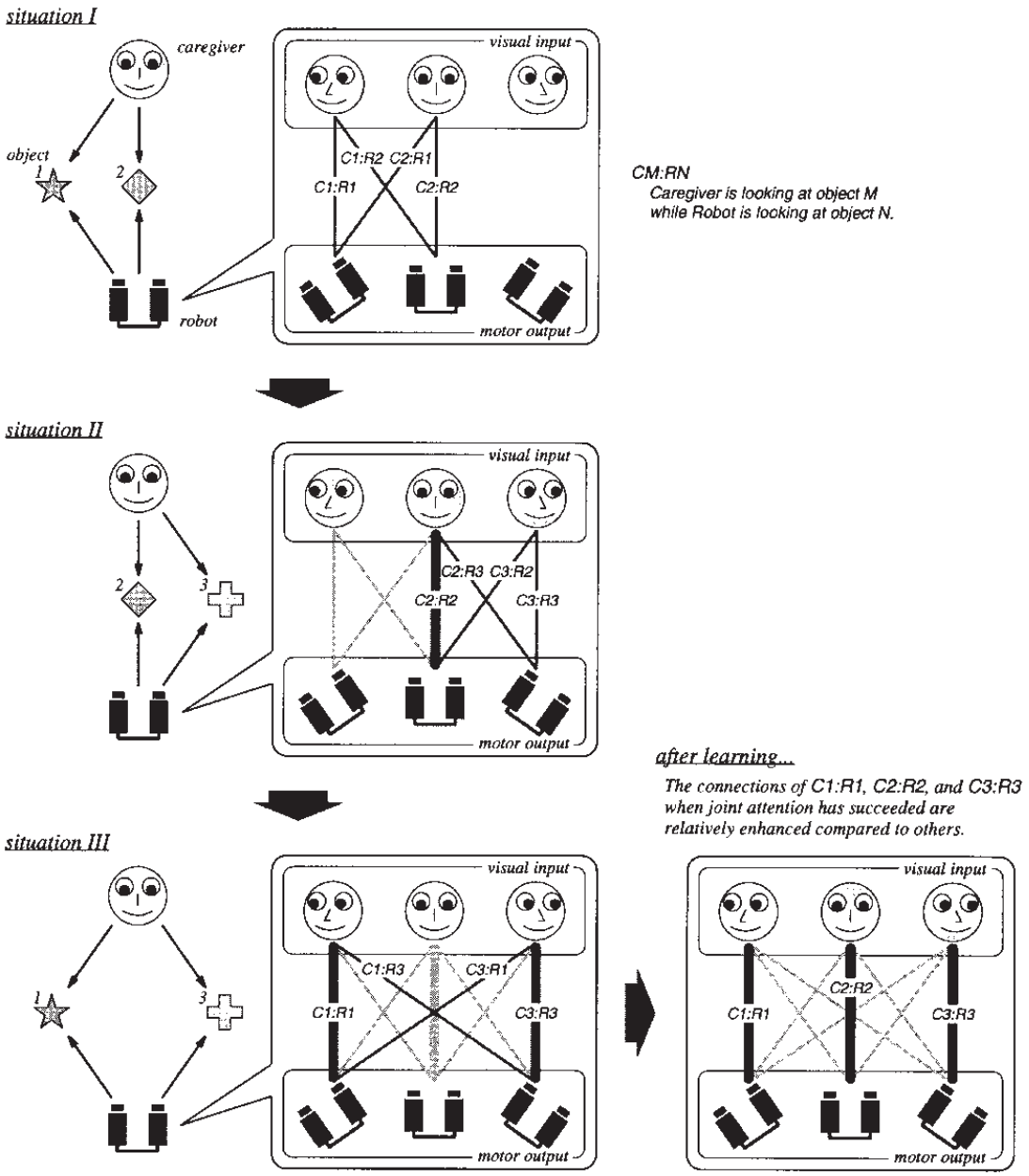



Figure A1. The process of finding a correlation for joint attention in the sensorimotor co-ordination through learning based on the proposed model. In each situation, the environment changes into including two objects, and the robot as well as the caregiver have two choices to look at an object. The robot learns the connection between the sensor input and the motor output when visual attention has succeeded. As a result the connection has strength proportional to the number of learning. In this example, the robot acquires the sensorimotor co-ordination shown on the lower right, in which the connections of C1:R1, C2:R2 and C3:R3 when joint attention has succeeded are relatively enhanced compared with others and show a correlation. This result explains that the proposed model enables the robot to acquire the ability of joint attention by finding the correlation in its sensorimotor co-ordination without any evaluation from the caregiver.

and II. In this situation, the robot acquires the connections of C1:R1, C1:R3, C3:R1 and C3:R3. Since the learning of the connections of C1:R1 and C3:R3 are for the second time, the strengths of these connections are doubled compared with other connections as well as C2:R2.

As a result, the robot acquires the sensorimotor co-ordination shown in the lower right of figure A1, in which the connections of C1:R1, C2:R2 and C3:R3 are relatively enhanced compared with other connections. It can be confirmed that these enhanced connections show a correlation in the sensorimotor co-ordination, and all of the connections have acquired when the robot looked at the same object that the caregiver was looking at, that is, when joint attention succeeded. It can be concluded that the proposed model enables the robot to acquire the ability of joint attention by finding the correlation in the sensorimotor co-ordination. This example shows the case when the number of objects is limited to under three and the positions of objects are fixed, however the real model can deal with more than three objects that are placed at random positions. It can be proved mathematically that the sensorimotor co-ordination when joint attention has succeeded is relatively enhanced compared with that when joint attention has failed under the condition that the number of objects is smaller than the resolution of the environment.