Proceedings of the Third International Symposium on Imitation in Animals and Artifacts (AISB'05 Convention) pp. 87-96, April 2005

Joint Attention Development in Infant-like Robot based on Head Movement Imitation

Yukie Nagai
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289 Japan
yukie@nict.go.jp

Abstract

The ability to imitate others enables human infants to acquire various social and cognitive capabilities. Joint attention is regarded as a behavior that can be derived from imitation. In this paper, the developmental relationship between imitation and joint attention, and the role of motion information in the development are investigated from a viewpoint of cognitive developmental robotics. It is supposed in my developmental model that an infant-like robot first has the experiences of visually tracking a human face based on the ability to preferentially look at salient visual stimuli. The experiences allow the robot to acquire the ability to imitate head movement by finding an equivalence between the human's head movement and the robot's when tracking the human who is turning his/her head. Then, the robot changes its gaze from tracking the human face to looking at an object at which the human has also looked at based on the abilities to imitate a head turning and gaze at a salient object. Through the experiences, the robot comes to learn joint attention behavior based on the contingency between the head movement and the object appearance. The movement information which the robot perceive plays an important role in facilitating the development of imitation and joint attention because it gives an easily understandable sensorimotor relationship. The developmental model is examined in learning experiments focusing on evaluating the role of movement in joint attention. Experimental results show that the acquired sensorimotor coordination for joint attention involves the equivalence between the human's head movement and the robot's, which can be a basis for head movement imitation.

1 Introduction

Neonatal imitation is a remarkable capability in human development. Such behavior might tell us that infants can associate their own action with others' action they see. The ability to imitate enables infants to acquire social identification and further social and cognitive capabilities (Meltzoff and Moore, 1997). Through the experiences of reproducing others' action, infants come to be able to understand the meaning of the action and the others' intention. Joint attention (Scaife and Bruner, 1975; Butterworth and Jarrett, 1991; Moore and Dunham, 1995) is one of the capabilities that can be derived from imitation. It is defined as a behavior to look where someone else is looking by following his/her gaze. In other words, joint attention is regarded as a type of imitative behavior that one turns one's own head and eyes towards the same side as another turns his/hers.

In this paper, the developmental relationship between imitation and joint attention, and the role of movement information in the development are discussed. Many researchers in cognitive science and developmental psychology have been investigating the capabilities of imitation and joint attention as the basis for infant development (Moore and Dunham, 1995). However, it is difficult to find the study in which the developmental relationship between the two abilities was examined. As described above, joint attention is an imitative behavior to copy others' head and eyes turning. It can emerge in infant-caregiver interactions when either of them, mostly a caregiver, introduces an object into their dyadic interactions based on the imitation. Considering the developmental progress from the dyadic to the triadic interaction, i.e. joint attention, is important for understanding the social and cognitive development in infants. This paper presents the developmental progress by which an infant-like robot incrementally learns to imitate and establish joint attention through interactions with a human caregiver. It is discussed from a standpoint of cognitive developmental robotics (Asada et al., 2001)

what capabilities a robot should be equipped with for interacting with an environment and learning the experiences, and how a caregiver should encourage and support the robot's development. As a key for the consecutive development from imitation to joint attention, a robot employs movement as its perceptual information. It is known in infant development that the motion information facilitates the development of the two abilities, e.g. (Vinter, 1986; Moore et al., 1997). Infants are more able to imitate others' action and comprehend others' gaze when they are presented with the behavior with the movement rather than without the movement. On the basis of the knowledge, a learning model by which a robot acquires joint attention ability through the experiences of head movement imitation by using motion information is proposed.

The rest of the paper is organized as follows. First, the findings about imitation and joint attention in infants are referred, in which the role of movement in the development is suggested. The finding of head movement imitation is also indicated, which is considered as a basis for joint attention development. Then, the current robotics models of imitation and joint attention are reported. Various models have been proposed with the aim of investigating infant development and/or constructing intelligent robots. The problems that the models did not deal with the developmental progress between imitation and joint attention and did not utilize motion information are pointed out. Next, a developmental model by which a robot learns joint attention based on head movement imitation is proposed. By utilizing motion information, a robot incrementally learns to imitate head movement and achieve joint attention without any *a priori* or symbolic representation for perceptual information given by a designer. Experiments that examined the validity of the model by using an infant-like robot are then described. Finally, discussion and ongoing work are given.

2 Related work on imitation and joint attention

2.1 Findings from infant studies

Meltzoff and Moore (1977, 1989) investigated the ability to imitate in infants at a few days or a few weeks of age. They found that infants were able to imitate facial and manual gestures and head movements demonstrated by an adult. On the basis of the finding, Meltzoff and his colleagues (Meltzoff and Moore, 1997; Rao and Meltzoff, 2003) proposed an active intermodal mapping model as the mechanism for early facial imitation. According to their model, infants can imitate an action by evaluating the equivalence between the action they see and their own action in a supra-modal representational space. In contrast, Jacobson (1979) suggested that facial and manual gestures of infants could be elicited by the presentation of a moving object. She showed that a moving pen and a ball were as effective as the tongue model of an adult in eliciting tongue protrusion by infants, and that a dangling ring elicited as much hand opening and closing as the adult hand model. This finding suggests that the motion information which infants perceive plays an important role in their early imitation. Vinter (1986) also indicated the significance of motion information in infant imitation. She showed that infants were more likely to imitate facial and manual gestures when they were presented with the gestures with the movement rather than without the movement. The reason was conjectured that the movement which infants perceive is effective in encoding their perceptual information.

Joint attention development has also been suggested to be facilitated by motion information. Moore et al. (1997) compared the infants' ability to learn gaze following when infants were presented with the final static state of an adult's head turning and the ability when infants were presented with the head turning with the movement. Their comparison showed that only infants presented with the movement were able to learn to establish gaze following. Lempers (1979) studied the developmental change in the ability to comprehend deictic gestures of infants at 9 to 14 months of age. His observational results showed that motion information helped younger infants to understand others' pointing and gaze. Corkum and Moore (1998) investigated the origin of joint attention and found that infants have a developmental stage at which they respond sensitively to the movement of an adults' gaze shift. They also examined the learning performance of joint attention in infants by presenting the infants with unnatural situations in which an interesting target appeared in the opposite side to the direction of an adult's head turning. Their examination showed that infants did not acquire the behavior to look at the object by turning to the opposite side of the head turning, but acquired the behavior to follow the adult's head turning although they could not find any object. This means that the learning mechanism of joint attention is not only based on the contingency between the adult's head turning and the object activation but also facil-

itated by the physical characteristics of the adult action, i.e. the direction of the head movement. I suppose from the result that infants learn the relationship between their own action and others' action before learning to find an object based on the others' cue.

2.2 Computational and robotic models

In order to investigate infant development and/or construct intelligent robots, computational and robotic models of imitation and joint attention have been proposed based on the findings from infant studies. Demiris and his colleagues (Demiris and Hayes, 1996; Demiris et al., 1997) constructed a model of head movement imitation based on the scheme of the active intermodal mapping proposed by (Meltzoff and Moore, 1997). Their model enabled a robot to imitate a human's head movement by establishing an equivalence between the human's head posture, which was estimated from the movement detected as an optical flow, and the robot's posture, which was given as encoder values. Scassellati (1999) built a humanoid robot that could imitate yes/no nods of a human. In his model, a robot recognized the yes/no nods by detecting the cumulative displacement of a human face in the robot's vision and then drove the fixed-action patterns for moving the robot's head as an imitative behavior.

The author (Nagai et al., 2002, 2003) proposed developmental models by which a robot learned joint attention through interactions with a human caregiver. I investigated how a robot with limited and immature capabilities could acquire the joint attention ability based on the evaluation from a caregiver or based on the robot's ability to autonomously find a sensorimotor contingency through its experiences. Triesch and his colleagues (Carlson and Triesch, 2003; Lau and Triesch, 2004) introduced the scheme of reward-based learning for a computational developmental model of gaze following. They suggested that the infant abilities of preferential looking, habituation, and reward-based learning, and an environmental setup in which a caregiver looks at an object that an infant prefers to look at can be a basic set for the emergence of gaze following. Shon et al. (2004a,b) constructed a model by which a robot acquired the ability to establish joint attention based on the imitation of a human's head movement. In their model, the imitation was achieved based on the scheme of the intermodal equivalence mapping (Meltzoff and Moore, 1997). In other words, a robot could imitate a head movement by turning its head to the same posture as that of the human, which was estimated from an image pattern of the human head. Then, the imitation of the head movement enabled a robot to achieve joint attention by finding an object at which the human was looking based on a probabilistic model.

However, these models of robotic imitation and joint attention have problems that they did not utilize motion information detected from visual perception and that they learned the mechanism to estimate the posture of a human head by using the exact posture which could not be detected by a robot. The following section presents a developmental model by which a robot consecutively learns to imitate and establish joint attention by utilizing both static and motion information detected by itself.

3 Joint attention development based on head movement imitation

3.1 Developmental progress

The developmental progress of joint attention via head movement imitation is shown in Figure 1. The development is based on the infant abilities to interact with an environment and learn the experiences and encouragement by a caregiver.

An infant is supposed to have the capability to preferentially look at salient visual stimuli, such as a bright colored object and a human face. This basic capability enables an infant to interact with an envi-

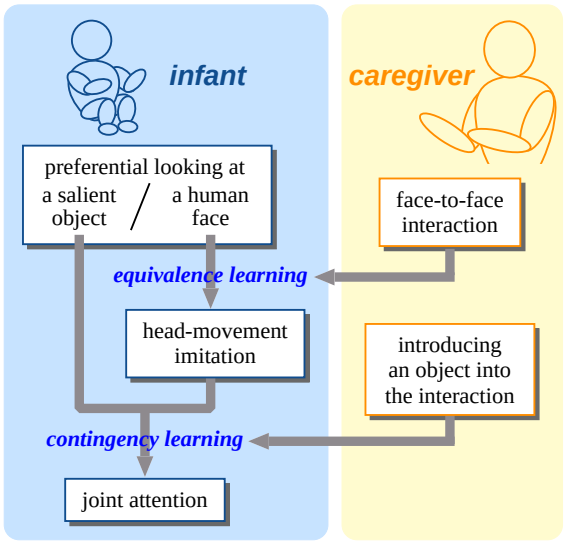


Figure 1: The developmental progress of joint attention via head movement imitation.

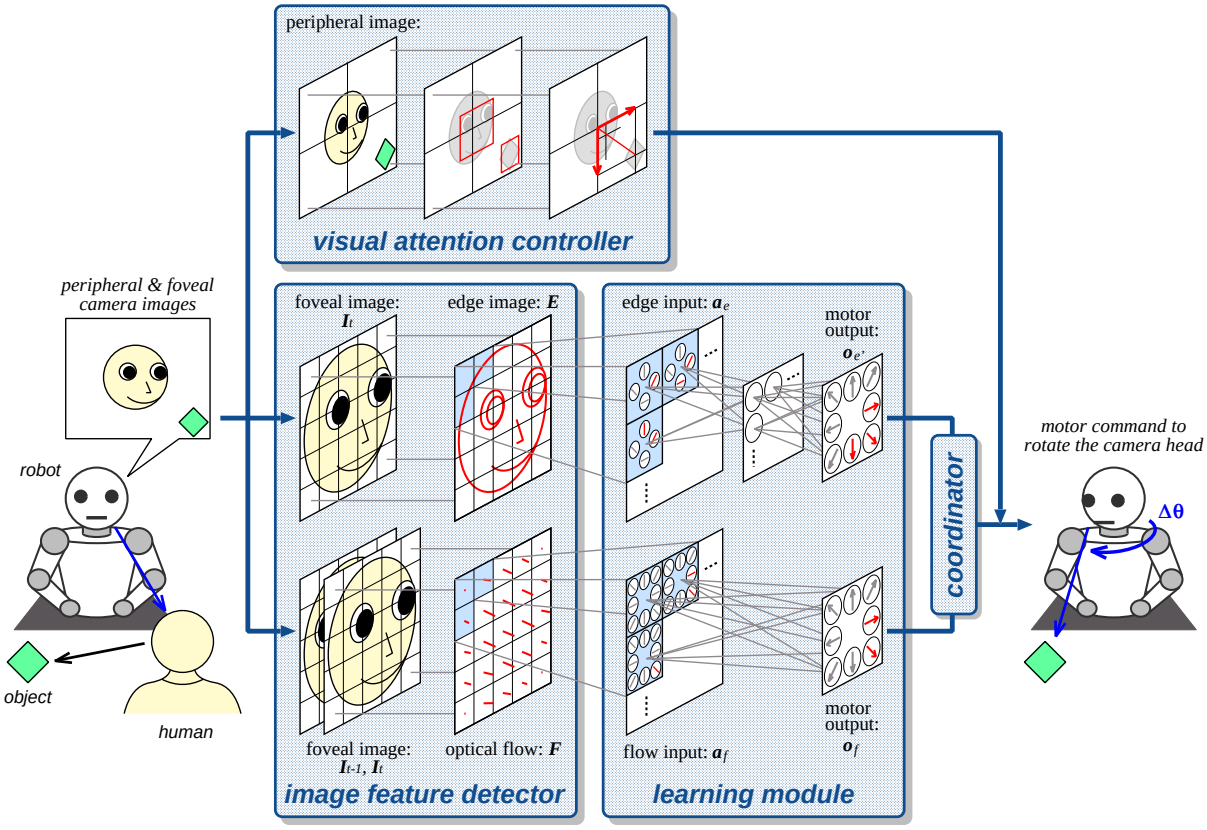


Figure 2: A learning model of joint attention based on head movement imitation. The visual attention controller enables a robot to have experiences of preferentially looking at a human face and a salient object. Through the experiences, the robot learns the sensorimotor coordination to imitate a head movement and achieve joint attention through the lower three modules.

ronment and have experiences for learning to imitate and establish joint attention. In early development, an infant often has dyadic interactions with a caregiver because of the caregiver's encouragement. A caregiver attempts to involve an infant in face-to-face interactions and emotionally communicate with the infant by showing facial expressions and head movements. The caregiver's movement drives the infant to visually track the caregiver's face as an interesting target, which provides experiences for learning to imitate head movement. In other words, when the caregiver turns his/her head vertically or laterally, the infant also turns his/her head to almost the same direction by tracking the caregiver's face. As the result, the infant finds an equivalence between the movement of the caregiver's head and that of the infant's and consequently acquires the ability to imitate head movements.

In parallel with or following the learning of head movement imitation, an infant starts to learn to achieve joint attention. A caregiver introduces an object, at which an infant prefers to look, into their dyadic interactions by presenting the infant with the object near the line of the infant gaze. The caregiver attempts to control the infant attention by moving the object and shifting the caregiver's own gaze to the object. The caregiver's encouragement drives the infant to change his/her attention target. The infant shifts his/her gaze from looking at the caregiver to looking at the object based on the abilities to imitate the caregiver's head movement and preferentially look at a salient object. This provides an experience for learning joint attention. The infant can acquire the sensorimotor coordination for joint attention by finding a contingency between the caregiver's gaze shift and the appearance of the object.

3.2 Learning model of joint attention based on head movement imitation

Figure 2 shows a proposed model by which a robot incrementally learns to imitate head movements and

establish joint attention. The model consists of four modules: a visual attention controller, an image feature detector, a learning module, and a coordinator. The visual attention controller enables a robot to have experiences of looking at salient visual stimuli. The latter three modules enable the robot to learn the sensorimotor coordination for imitation and joint attention through the above experiences.

3.2.1 Visual attention controller

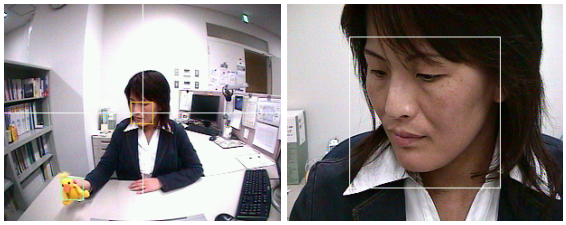
The visual attention controller enables a robot to have fundamental experiences for the development. This module enables a robot to preferentially look at salient visual stimuli, such as a human face and a bright colored object, in an environment. A human face and a salient object are respectively detected by template matching and using color information from a peripheral camera image. Figure 3 (a) shows an example of the peripheral image, in which a human face and a yellow object are indicated by rectangles. In this case, the robot is controlling its gaze to look at the human face at the center of the image. A motor command to look at the object can be generated by multiplying the horizontal and vertical displacement between the object and the center of the image by scalar values.

3.2.2 Image feature detector

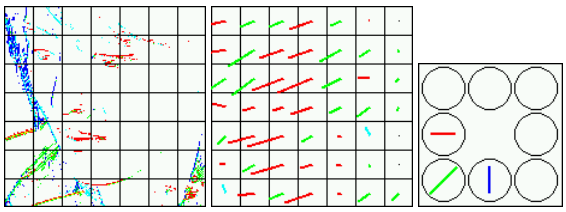
The image feature detector extracts visual information needed to achieve imitation and joint attention. The detector extracts the edge image $\boldsymbol{E}$ of a human face and the optical flow $\boldsymbol{F}$ of the human's gaze shift from foveal camera images $\boldsymbol{I}_{t-1},\ \boldsymbol{I}_t$. An example of the detected features is shown in Figure 3 (b)-(d), in which (b) shows the foveal camera image when the robot is gazing at the human face as shown in (a), and (c) and (d) show the edge image and the optical flow detected from the center area (168 $\times$ 168 pixels) enclosed by a rectangle in (b). The position of the enclosed area is fixed at the center of the foveal image. The foveal and peripheral cameras are mechanically fixed and controlled to gaze at a visual target at the center of the peripheral image.

The edge image $\boldsymbol{E}$ is generated by orientation selective filters. Four filters that are selective with respect to four orientations $(e_1,\ e_2,\ e_3,\ e_4) = (-,\ \diagdown,\ |,\ \diagup)$ extract edge images $\boldsymbol{E}_n$, where $n = 1, \ldots,\ 4$, each of which includes one oriented edge. The value of each pixel $E_n(x,\ y)$ is calculated as

$$E_n(x,\ y) = \begin{cases} 1 & \text{if } \epsilon_n(x,\ y) > \epsilon_{\text{threshold}} \\ 0 & \text{otherwise,} \end{cases}$$



(a) peripheral camera image (b) foveal camera image: $\boldsymbol{I}_t$



(c) edge image: $\boldsymbol{E}$ (d) optical flow: $\boldsymbol{F}$ (e) motor output: $\boldsymbol{o}_{e'f}$

Figure 3: An example of input-output datasets, in which (a) and (b) show a peripheral and a foveal camera image when the robot is looking at the human; (c) and (d) show the edge image and the optical flow detected from the center area in (b); (e) shows motor output to follow the human gaze, which is encoded in motion direction selective neurons.

where

$$\epsilon_n(x,\ y) = \left| \sum_{i=-1}^{1} \sum_{j=-1}^{1} \alpha_n(i,\ j) I(x+i,\ y+j) \right| - \left| \sum_{i=-1}^{1} \sum_{j=-1}^{1} \beta_n(i,\ j) I(x+i,\ y+j) \right|. \quad (1)$$

$(x,\ y)$ indicate a position in a camera image, and the coefficients, $\alpha_n(i,\ j)$ and $\beta_n(i,\ j)$, are given as

$$\boldsymbol{\alpha}_1 = \boldsymbol{\beta}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{bmatrix},\ \boldsymbol{\beta}_1 = \boldsymbol{\alpha}_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix},$$

$$\boldsymbol{\alpha}_2 = \boldsymbol{\beta}_4 = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix},\ \boldsymbol{\beta}_2 = \boldsymbol{\alpha}_4 = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix},$$

where

$$\boldsymbol{\alpha}_n = \begin{bmatrix} \alpha_n(-1,-1) & \alpha_n(0,-1) & \alpha_n(1,-1) \\ \alpha_n(-1,\ 0) & \alpha_n(0,\ 0) & \alpha_n(1,\ 0) \\ \alpha_n(-1,\ 1) & \alpha_n(0,\ 1) & \alpha_n(1,\ 1) \end{bmatrix}. \quad (2)$$

Figure 3 (c) shows the edge image $\boldsymbol{E}$ combining $\boldsymbol{E}_n$ ($n = 1, \ldots, 4$), in which edges with one of the four orientations, —, ╲, |, and ╱, are colored red, cyan, blue, and green, respectively. The edge image provides information to estimate the static direction of the human head and allows the robot to acquire the accurate sensorimotor coordination to achieve head movement imitation and joint attention.

The image feature detector also extracts the optical flow $\boldsymbol{F}$. The center area of the foveal image is divided into small areas called receptive fields (24 × 24 pixels). The optical flow $\boldsymbol{F}^k$ in the k-th receptive field is calculated as the cumulative displacement of the image feature in the receptive field over ten image frames:

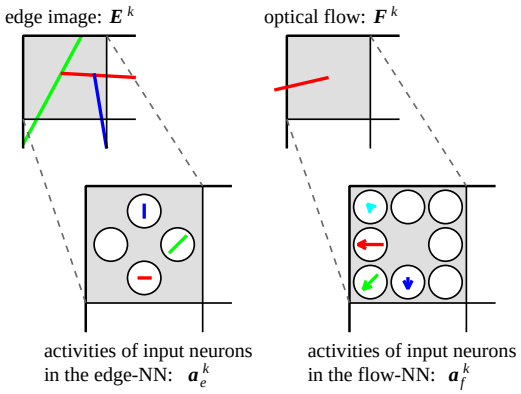
$$\boldsymbol{F}^k = \begin{bmatrix} \sum^{10\text{frames}}(x_k - px) \\ \sum^{10\text{frames}}(y_k - py) \end{bmatrix}, \qquad (3)$$

where $(x_k,\ y_k)$ and $(px,\ py)$ are the center position of the k-th receptive field in $\boldsymbol{I}_t$ and that of the corresponding image area detected by template matching in $\boldsymbol{I}_{t-1}$, respectively. Figure 3 (d) shows the optical flow detected when the human changes her gaze from looking straight at the robot's camera to looking at the yellow object shown in (a). Like the edges, the flows are drawn with four colors. Although the optical flow cannot provide enough information to infer the exact direction of the human head compared with the edge information, it gives a rough but easily understandable relationship with the movement direction of the human's head turning. Therefore, the flow information should enable the robot to quickly acquire rough sensorimotor coordination for head movement imitation and joint attention.

In addition, the flow information is utilized as a cue for the robot to control the timing of its own head turning. The temporal change in the amount of the optical flow indicates the start and end of a human's head turning. In other words, when the flow becomes zero after exceeding an upper threshold, this means that a human has shifted his/her head direction from looking at one location to looking at another and is gazing at a certain location. Based on this mechanism, the robot obtains the input data of the optical flow when the flow has a maximum value and the edge image when the flow becomes zero. This enables the robot to immediately follow a human's head turning without any explicit cue.

3.2.3 Learning module

This module learns the sensorimotor coordination between the edge input and motor output and between the optical flow and motor output through two independent neural networks (see Figure 2). The neural network for the edge input (the edge-NN) consists of three layers: input, hidden, and output layers, because edge information is difficult to interpret into the human's head direction. In contrast, the neural network for the optical flow input (the flow-NN) has two layers: input and output layers, because flow information gives an easily understandable relationship with the motor output to imitate the human's head movement and achieve joint attention.



(a) the encoding of edge input (b) the encoding of flow input

Figure 4: The encoding of detected image features into the input neurons, in which (a) and (b) show the encoding of edge and flow inputs into the four orientation selective neurons and the eight directions selective neurons, respectively. The length of a line in each circle denotes the activity of the neuron.

Input to the edge-NN is represented as activities of four kinds of neurons that are selective to four orientations. Figure 4 (a) shows edge input encoding into the selective neurons. The activities of the four neurons $a_{e_n}^k$ $(n = 1, \ldots, 4)$ in the k-th receptive field are calculated as

$$a_{e_n}^k = E_n^k / \max_k \sum_{m=1}^{4} E_m^k,$$
$$\text{where } E_n^k = \sum_{x_k} \sum_{y_k} E_n(x,\ y). \qquad (4)$$

$E_n(x,\ y)$ is given by (1), and E_n^k means the amount of the edge e_n in the k-th receptive field. In the bottom of Figure 4 (a), the length of a line in each circle shows the activity of each neuron. No line means that the activity is zero.

Like the encoding of edge input, the optical flow

is encoded in eight kinds of neurons that are selective to eight directions $(f_1,\ f_2, \ldots,\ f_8) = (\leftarrow,\ \nwarrow, \ldots,\ \swarrow)$ as shown in Figure 4 (b). The activities of the eight neurons $a^k_{f_n}$ $(n = 1,\ 2, \ldots,\ 8)$ in the k-th receptive field are calculated as

$$a^k_{f_n} = \begin{cases} \boldsymbol{F}^k \cdot \boldsymbol{u}_n / \max_k \|\boldsymbol{F}^k\| & \text{if } \boldsymbol{F}^k \cdot \boldsymbol{u}_n \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad (5)$$

where $\boldsymbol{F}^k$ is given by (3), and $\boldsymbol{u}_n$ are unit vectors in eight directions. The activities of the eight neurons are also drawn as the length of the arrows as shown in Figure 4 (b). The methodology of coding edge and flow information is based on physiological evidence that the visual cortex in some animates has orientation selective neurons (Hubel and Wiesel, 1959) and motion direction selective neurons (Barlow and Hill, 1963). The similarity in the representation of edge and flow inputs leads to the possibility that the robot translates a well-acquired sensorimotor coordination in the edge-NN or the flow-NN into the other.

Outputs from the edge- and flow-NNs are represented as the activities of eight neurons, $o_{e'_n}$ and o_{f_n} $(n = 1, \ldots,\ 8)$, which are selective to eight motion directions $(e'_1, \ldots,\ e'_8) = (f_1, \ldots,\ f_8) = (\leftarrow, \ldots,\ \swarrow)$, respectively. Figure 3 (e) shows an example of the activities of the output neurons. The representation is similar to that of encoded optical flow data. The activities of the output neurons are decoded into a motor command $\boldsymbol{\Delta\theta}$ to move the robot's head by the coordinator described in the next section.

3.2.4 Coordinator

This module coordinates motor outputs from the edge- and flow-NNs. In the experiments discussed here, the robot uses a simple method that generates a motor command $\boldsymbol{\Delta\theta}$ by decoding the mean value of the two outputs:

$$\boldsymbol{\Delta\theta} = \begin{bmatrix} \Delta\theta_{pan} \\ \Delta\theta_{tilt} \end{bmatrix} = \begin{bmatrix} g_{pan} \sum_n u_{n_x} o_{e'f_n} \\ g_{tilt} \sum_n u_{n_y} o_{e'f_n} \end{bmatrix}, \quad (6)$$

where g_{pan} and g_{tilt} are scalar gains; u_{n_x} and u_{n_y} are the horizontal and vertical components in $\boldsymbol{u}_n$; $o_{e'f_n}$ is the mean value of $o_{e'_n}$ and o_{f_n}. A motor command to move the robot's head is represented as displacement angles in the pan and tilt directions.

3.3 Learning processing

Employing the model, a robot has two-staged learning. First, a robot learns the sensorimotor coordination to imitate head movements. As a human turns his/her head vertically and laterally in front of the robot, the robot also turns its head to almost the same direction by tracking the human face based on the visual attention controller. Through the experiences, when the robot detects simultaneous activation of the input and output neurons that are selective to the same directions in the flow-NN, it learns the equivalence of the movement by multiplying the connecting weights between the neurons. This leads to the ability to imitate head movements. Next, the robot learns the sensorimotor coordination for joint attention. The human starts to introduce an object into the human-robot dyadic interactions. When the human shifts his/her gaze direction from the robot to the object by turning his/her head, the robot first imitates the head movement based on the acquired sensorimotor coordination and then changes its gaze from looking at the human to looking at the object based on the visual attention controller. This provides a sensorimotor experience of joint attention. The robot learns the sensorimotor coordination in the edge- and flow-NNs by back propagation based on the input-output dataset obtained in the above process and consequently acquires joint attention ability.

4 Preliminary experiment

4.1 Experimental setup

As a preliminary experiment, the model was evaluated with a focus on the role of movement in learning joint attention. The model was implemented into an infant-like robot, called *Infanoid* (Kozima, 2002), shown in Figure 5, which was developed by our group

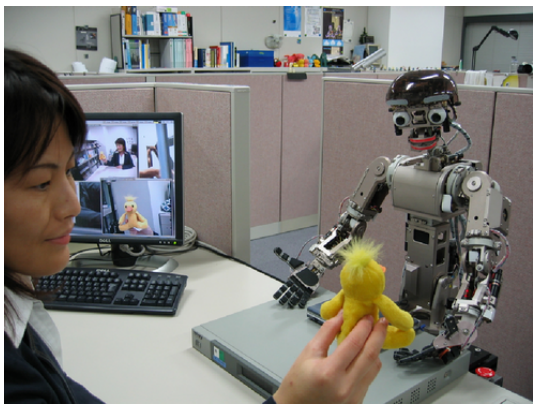


Figure 5: Human-robot joint attention, in which an infant-like robot, called *Infanoid*, is looking at the stuffed toy that the human is looking at.

as a tool for investigating the cognitive development in human infants. Infanoid has a stereo-vision head with three degrees of freedom (DOFs) in its neck (one for the pan and two for the tilt directions) and three DOFs in its eyes (two for the each pan and one for the common tilt directions). Each eye has two color CCD cameras: a peripheral camera and a foveal camera, and the two camera images from the left eye were used in the experiment. The three DOFs in the neck were used to move the robot's head while the three DOFs in the eyes were fixed at the center positions. The displacement angle $\Delta\theta_{tilt}$ derived from (6) was equally divided into the two tilt DOFs in the neck.

A human sat face to face with Infanoid and interacted with it by using a salient object. In every trial, the human replaced the object at random positions and then changed her gaze from looking at the robot to looking at the object by turning her head. The human always looked at the object in front of her face.

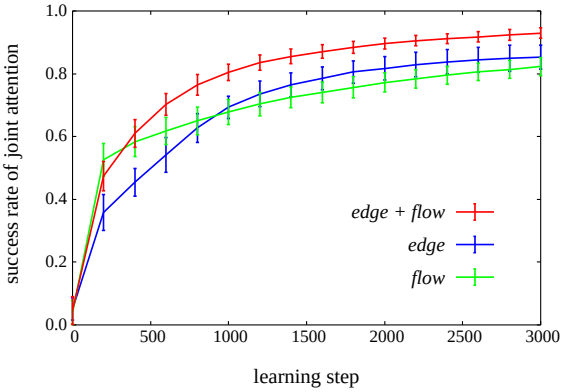


Figure 6: The change in the task performance of joint attention over the learning period. The red, blue, and green lines indicate the results when the model utilized both the edge and flow inputs, only the edge input, and the flow input, respectively.

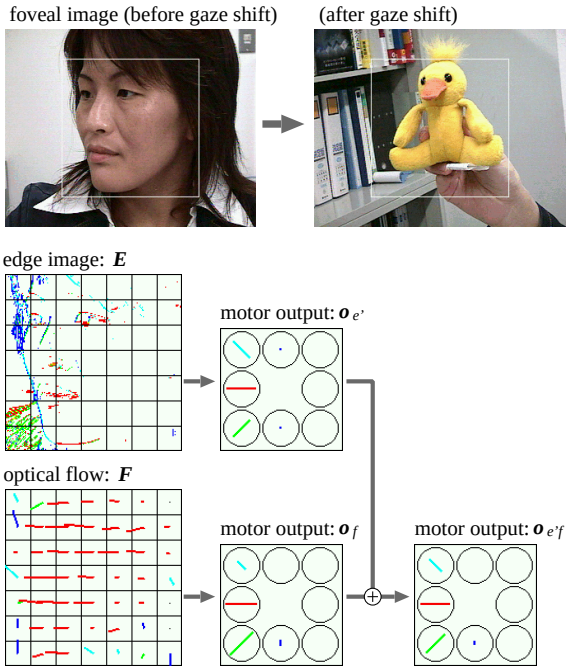
4.2 Evaluating the role of movement in learning joint attention

The role of motion information in learning joint attention was evaluated. In this experiment, Infanoid learned to establish joint attention without learning to imitate. In other words, the robot learned a contingency between the human's head turning and the object appearance to acquire the sensorimotor coordination for joint attention through the edge- and flow-NNs without using any pre-acquired sensorimotor coordination to imitate head movements.

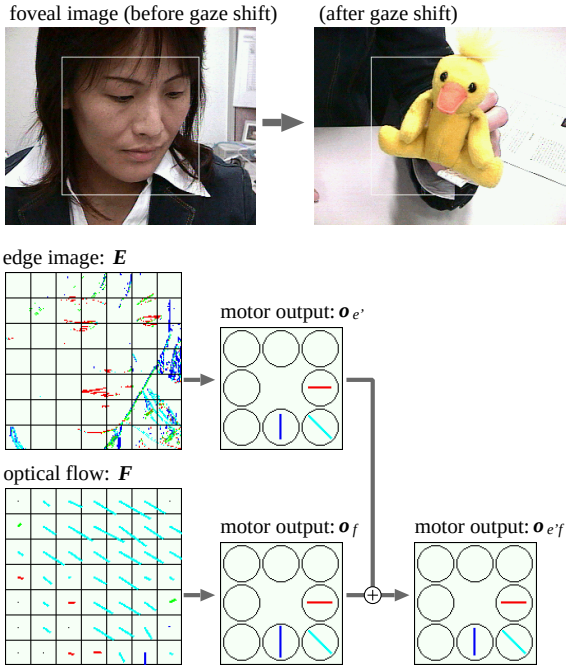
Figure 6 shows the changes in joint attention performance over the learning period, in which the horizontal and vertical axes respectively denote the learning step and the success rate of joint attention. The success of joint attention means that the robot looks at the object at which the human is looking within ± 8 degrees of error. The learning experiment was conducted off-line by repeatedly using 200 input-output datasets acquired beforehand, and the sensorimotor coordination acquired through learning was evaluated in joint attention experiments every 200 learning steps. The red line plots the result when the model used both the edge and flow inputs. The blue and green lines plot the results when the model used only the edge or the flow input, respectively. The graph shows the mean result of fifty experiments with different initial conditions and its standard deviation. Comparing the results for when the robot used either the edge or the flow input, it is confirmed that the flow input accelerated the start-up time of learning while the edge input gradually improved the task performance. This complementary result can be expected from the characteristics of the two inputs. By using both the edge and flow inputs, the model enabled the robot to quickly acquire the high performance of joint attention by combining the advantages of the two inputs.

4.3 Joint attention experiment after learning

The acquired sensorimotor coordination was evaluated in joint attention experiments. Figure 7 (a) and (b) show the two cases of input-output datasets when the robot attempted to achieve joint attention based on the acquired NNs. In case (a), the human shifted her gaze from looking at the robot to looking at an object in the outer left side of the foveal image. In case (b), the human shifted her gaze direction from the robot to an object in the outer lower right of the foveal image. The upper side of each figure shows the change in the foveal image when the robot shifted its head direction based on the output from the coordinator shown in the lower side. From these results, we can see that both the edge-NN and the flow-NN generated appropriate output to achieve joint attention. In these two cases, the robot was able to find the object at which the human was looking and establish joint attention. The success rate of joint attention with the same person as in the learning experiment was 90% (18/20 trials). In addition, we can confirm from this result that the flow-NN acquired one-to-one correspondence between the activities of the input and output neurons. The direction of the motor output from the flow-NN is



(a) In the case that the human shifted her gaze direction from the robot to an object in the outer left side of the foveal image.



(b) In the case that the human shifted her gaze direction from the robot to an object in the outer lower right of the foveal image.

Figure 7: The input-output datasets when the robot attempted to achieve joint attention based on the acquired NNs. The robot was able to establish joint attention in these two cases.

clearly corresponding with the same direction of the optical flow. This means that the sensorimotor equivalence, which should be acquired through imitation learning, is also utilized in joint attention.

5 Discussion and ongoing work

This paper has presented a developmental model by which a robot learns joint attention based on head movement imitation. The preliminary experiments showed that the model accelerated the learning of joint attention by using movement information and that the equivalence of self and other movement was utilized to achieve joint attention. This result supports the idea that joint attention emerges through the experiences of head movement imitation. Ongoing work is to examine that learning to imitate head movements facilitates the development of joint attention. This is expected to lead to the possibility to reveal the role of other neonatal imitation, such as tongue protrusion and hand opening-closing, in the development of social and cognitive capabilities of infants. Another issue to be solved is to develop a mechanism that enables a robot to recognize not only head directions but also gaze directions. It was assumed in the experiments that a human shifted his/her gaze by turning his/her head and looked at an object in front of his/her face. This assumption is likely in joint attention by infants. However, infants can acquire the ability to recognize gaze directions. To solve the problem, I will apply a mechanism that changes the resolution of the receptive fields in the NNs as learning proceeds. Such mechanism will increase the resolution around the image area including important facial features, e.g. eyes and mouth, and consequently enable a robot to acquire the ability to recognize gaze directions.

References

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

H. B. Barlow and R. M. Hill. Selective sensitivity to direction of movement in ganglion cells of the retina. *Science*, 139:412–414, 1963.

George Butterworth and Nicholas Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy.

British Journal of Developmental Psychology, 9: 55–72, 1991.

Eric Carlson and Jochen Triesch. A computational model of the emergence of gaze following. In *Proceedings of the 8th Neural Computation and Psychology Workshop*, 2003.

Valerie Corkum and Chris Moore. The origins of joint visual attention in infants. *Developmental Psychology*, 34(1):28–38, 1998.

J. Demiris, S. Rougeaux, G. M. Hayes, L. Berthouze, and Y. Kuniyoshi. Deferred imitation of human head movements by an active stereo vision head. In *Proceedings of the 6th IEEE International Workshop on Robot and Human Communication*, pages 88–93, 1997.

John Demiris and Gillian Hayes. Imitative learning mechanisms in robots and humans. In *Proceedings of the 5th European Workshop on Learning Robot*, pages 9–16, 1996.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, 148:574–591, 1959.

Sandra W. Jacobson. Matching behavior in the young infant. *Child Development*, 50:425–430, 1979.

Hideki Kozima. Infanoid: A babybot that explores the social environment. In K. Dautenhahn, A. H. Bond, L. Canamero, and B. Edmonds, editors, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, chapter 19, pages 157–164. Amsterdam: Kluwer Academic Publishers, 2002.

Boris Lau and Jochen Triesch. Learning gaze following in space: a computational model. In *Proceedings of the Third International Conference on Development and Learning*, 2004.

Jacques D. Lempers. Young children's production and comprehension of nonverbal deictic behaviors. *The Journal of Genetic Psychology*, 135:93–102, 1979.

Andrew N. Meltzoff and M. Keith Moore. Imitation of facial and manual gestures by human neonates. *Science*, 198:75–78, 1977.

Andrew N. Meltzoff and M. Keith Moore. Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. *Developmental Psychology*, 25(6):954–962, 1989.

Andrew N. Meltzoff and M. Keith Moore. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, 6:179–192, 1997.

Chris Moore, Maria Angelopoulos, and Paula Bennett. The role of movement in the development of joint visual attention. *Infant Behavior and Development*, 20(1):83–92, 1997.

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

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

Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

Rajesh P. N. Rao and Andrew N. Meltzoff. Imitation learning in infants and robots: Towards probabilistic computational models. In *Proceeding of Artificial Intelligence and Simulation of Behaviour: Cognition in Machines and Animals*, 2003.

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

Brian Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In C. Nehaniv, editor, *Computation for Metaphors, Analogy and Agents*, volume 1562 of LNCS, pages 176–195. Springer-Verlag, 1999.

Aaron P. Shon, David B. Grimes, Chris L. Baker, Matthew W. Hoffman, Shengli Zhou, and Rajesh P. N. Rao. Probabilistic gaze imitation and saliency learning in a robotic head. Technical report, University of Washington CSE, 2004a.

Aaron P. Shon, David B. Grimes, Chris L. Baker, and Rajesh P. N. Rao. A probabilistic framework for model-based imitation learning. In *Proceedings of the 26th Annual Meeting of the Cognitive Science Society*, 2004b.

Annie Vinter. The role of movement in eliciting early imitations. *Child Development*, 57:66–71, 1986.