Proceedings of the 2005 IEEE
International Conference on Robotics and Automation
pp. 2081-2086, April 2005

The Role of Motion Information in Learning Human-Robot Joint Attention

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*— To realize natural human-robot interactions and investigate the developmental mechanism of human communication, an effective approach is to construct models by which a robot imitates cognitive functions of humans. Focusing on the knowledge that humans utilize motion information of others' action, this paper presents a learning model that enables a robot to acquire the ability to establish joint attention with a human by utilizing both static and motion information. As the motion information, the robot uses the optical flow detected when observing a human who is shifting his/her gaze from looking at the robot to looking at another object. As the static information, it extracts the edge image of the human face when he/she is gazing at the object. The static and motion information have complementary characteristics. The former gives the exact direction of gaze, even though it is difficult to interpret. On the other hand, the latter provides a rough but easily understandable relationship between the direction of gaze shift and motor output to follow the gaze. The learning model utilizing both static and motion information acquired from observing a human's gaze shift enables the robot to efficiently acquire joint attention ability and to naturally interact with the human. Experimental results show that the motion information accelerates the learning of joint attention while the static information improves the task performance. The results are discussed in terms of analogy with cognitive development in human infants.**

***Index Terms*— human-robot joint attention, learning, motion information, optical flow**

I. Introduction

To design artificial models that imitate abilities of human beings or other animals is an effective methodology to develop intelligent and adaptive robots. Especially in studies on human-robot communication, implementing human-like cognitive models into a robot helps the robot and a human to understand each other's internal states. This understanding can lead to the emergence of natural human-robot communication. In addition, it is interesting to investigate how a robot develops and learns such cognitive capabilities through interactions with a human from a constructivist viewpoint [1]. The author [18], [19] has focused on joint attention as a form of nonverbal communication between a robot and a human (see Fig. 1). Joint attention is defined as a process to look at an object that someone else is looking at by following his/her gaze [5]. Through this interaction, human beings are able to infer others' internal states, i.e. desire, intention, knowledge and so on, and to naturally interact with others. Many researchers [4], [17], [20] in cognitive science and developmental psychology place importance on joint attention ability for social development in human infants. The ability enables infants to interact with adults and learn from adults. As it is important for infants, joint attention ability could play an important role for a robot to achieve natural interactions and acquire knowledge from humans.

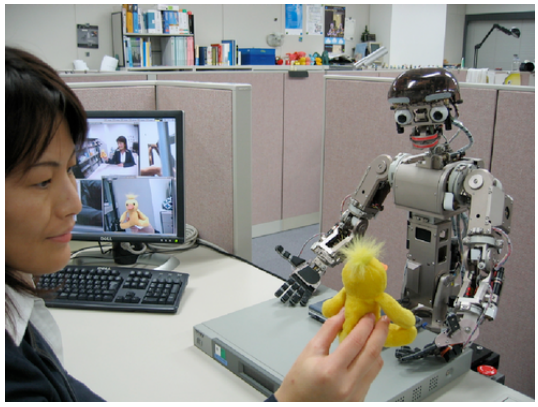


Fig. 1. Human-robot joint attention, in which an infant-like robot, called *Infanoid* [11], is looking at a stuffed toy that a human is holding in her hand by following the direction of her gaze.

The author [18], [19] proposed learning models by which a robot acquired joint attention ability through interactions with a human. On the basis of cognitive developmental findings, I have been investigating how a robot with limited and immature capabilities, like those of infants, acquires the ability to follow human gaze. As a related study, Triesch and his colleagues [6], [13] have been investigating joint attention development in infants by taking a computational approach in closely cooperating with cognitive developmental research. There are a number of studies aiming at facilitating human-robot interactions based on joint attention [3], [8] and discussing further cognitive development based on joint attention [12], [21]. Joint attention mechanisms not only between a human and a robot but also between two robots have been developed [10]. This recent work related to human-robot/robot-robot joint attention demonstrates the significance of joint attention in communication. However, the joint attention models described in the recent work have utilized only static information from another agent, e.g. the posture and/or the face direction of the agent, but did not use any visually perceived motion information from another agent.

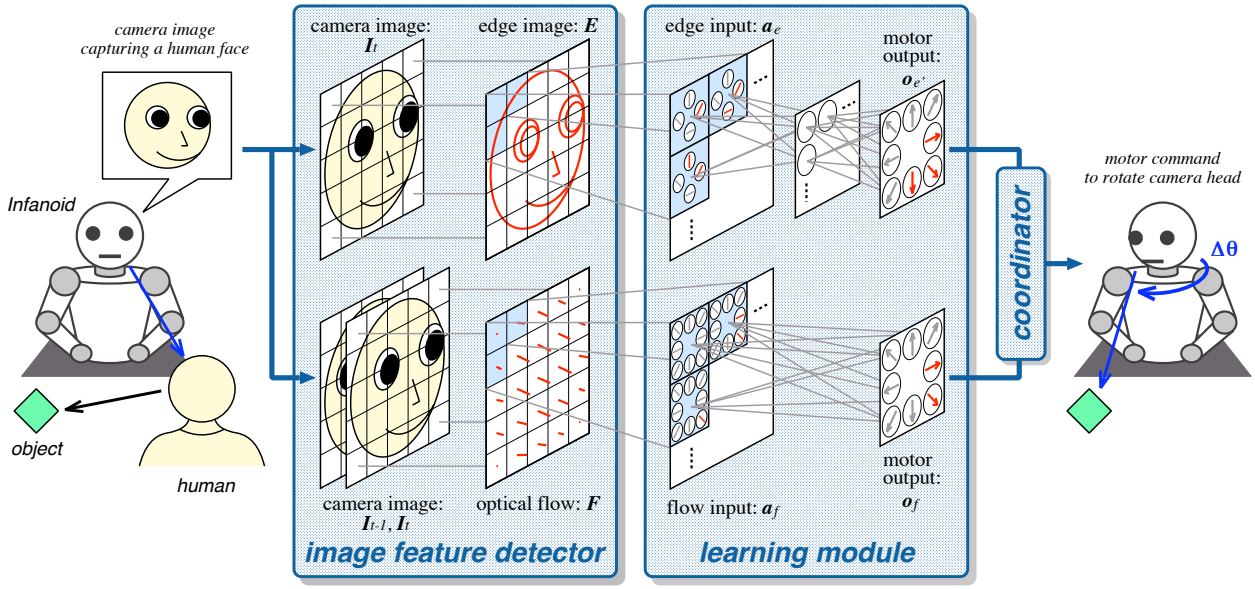


Fig. 2. A learning model of joint attention utilizing the edge image of a human face as static information and the optical flow of the human's gaze shift as motion information.

Human beings clearly utilize motion information from others' action. That is, we receive cues from others' movements from which we infer their desires and intentions. Movement provides novel information unlike static information. For instance, it has been suggested that motion information facilitates infants' learning of joint attention [14], [16]. Studies on neonatal imitation, which precedes the development of joint attention, indicated the importance of movement in eliciting facial and manual imitations by newborns [9], [22]. In addition, physiological evidence indicates that some animates have selective neurons to motion directions in the visual cortex [2]. These findings from cognitive science, developmental psychology, and neuroscience support the validity of utilizing motion information acquired from others for designing communication mechanisms for a robot.

This paper presents a learning model by which a robot acquires the sensorimotor coordination to establish joint attention with a human by utilizing both motion and static information acquired from observing a human's gaze shift. The motion information is the optical flow detected when the robot is observing a human who shifts his/her gaze from looking at the robot to looking at another object. The static information is the edge image extracted when the robot is looking at the human while he/she is gazing at the object. These two kinds of information have complementary characteristics. The former provides a rough but easily understandable relationship between the direction of a gaze shift and motor output to follow the gaze. The latter gives the exact gaze direction, even though it is difficult to interpret. The learning model utilizing both static and motion information enables a robot to efficiently acquire joint attention ability and to establish natural interactions with a human. The validity of the model was examined using an infant-like humanoid robot, called *Infanoid* [11], shown in Fig. 1. The experimental results show that the motion information accelerates the learning of joint attention while the static information improves the task performance.

The following section explains the learning model of joint attention utilizing static and motion information. Experiments are then described, and their results are discussed in terms of analogy with cognitive development in human infants. Conclusions and future work are given at the end.

II. Learning Model of Joint Attention Utilizing Edge Image and Optical Flow

A learning model of joint attention utilizing both static and motion information is shown in Fig. 2. The model consists of three modules: an image feature detector, a learning module, and a coordinator. Utilizing the model, a robot learns the sensorimotor coordination between the camera images, $\boldsymbol{I}_{t-1}$ and $\boldsymbol{I}_t$, from which the edge image and the optical flow of human gaze are detected, and the motor output $\boldsymbol{\Delta\theta}$ to follow the gaze. The mechanisms of the three modules are explained below.

A. Image Feature Detector

The image feature detector extracts the edge image $\boldsymbol{E}$ of a human face and the optical flow $\boldsymbol{F}$ of the human's gaze shift from the camera images $\boldsymbol{I}_{t-1}$, $\boldsymbol{I}_t$. The edge image provides static information while the flow provides motion information. An example of input-output datasets is shown in Fig. 3, in which (a) and (b) show a peripheral and a foveal camera image; (c) and (d) show the edge image and the optical flow detected from the center area (168 $\times$ 168 pixels) enclosed with a rectangle in (b); (e) shows the output to follow the human gaze. The position of the enclosed area is fixed at the center of the foveal image. The robot controls the directions of the peripheral and foveal

cameras, which are mechanically fixed, so that it looks at the human face 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

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

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

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

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

where

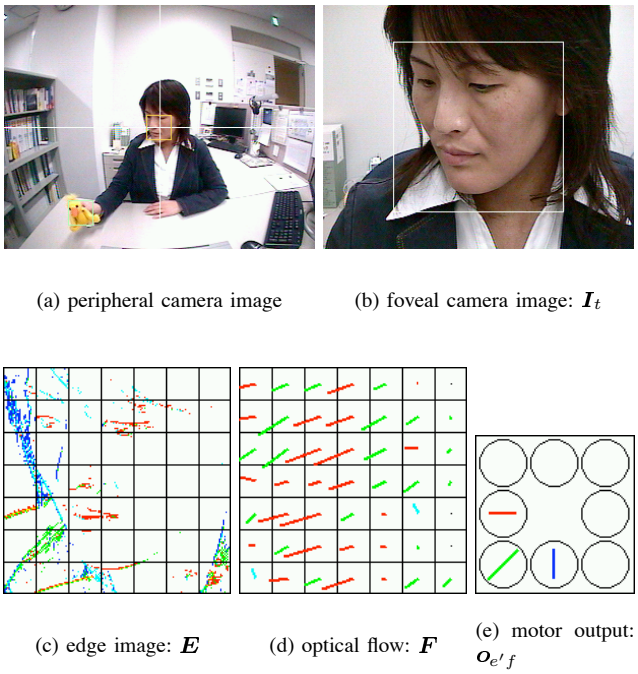
$$\boldsymbol{\alpha}_n = \begin{bmatrix} \alpha_n(-1,-1) & \alpha_n(0,-1) & \alpha_n(1,-1) \\ \alpha_n(-1,\ 0) & \alpha_n(0,\ 0) & \alpha_n(1,\ 0) \\ \alpha_n(-1,\ 1) & \alpha_n(0,\ 1) & \alpha_n(1,\ 1) \end{bmatrix}. \quad (2)$$

Fig. 3 (c) shows the edge image $\boldsymbol{E}$ combining $\boldsymbol{E}_n$ $(n = 1, \ldots,\ 4)$. Edges with one of the four orientations, —, $\diagdown$, |, and $\diagup$, are colored red, cyan, blue, and green, respectively. The edge image provides static information to estimate the direction of human gaze and allows the robot to acquire the accurate sensorimotor coordination for joint attention.

The image feature detector also extracts the optical flow $\boldsymbol{F}$ as motion information. The center area of the foveal image is divided into small image 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 field over ten image frames:

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

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. Fig. 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 gaze direction compared with the edge information, it gives a rough but easily understandable relationship with the motor output to follow the gaze. Therefore, the flow information should enable the robot to quickly acquire the rough sensorimotor coordination of joint attention.



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

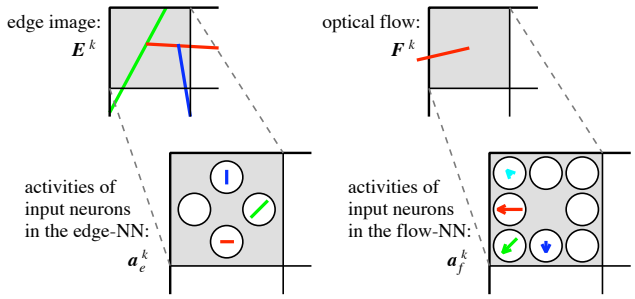
(c) edge image: $\boldsymbol{E}$ (d) optical flow: $\boldsymbol{F}$ (e) motor output: $\boldsymbol{o}_{e'f}$

Fig. 3. An example of input-output datasets: (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.

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

B. 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 Fig. 2). The neural network for 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 gaze direction. In contrast, the neural network for optical flow input (the flow-NN) has two layers: input and output layers, because flow information



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

Fig. 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-direction selective neurons, respectively. The length of a line in each circle denotes the activity of the neuron. No line means there is zero activity.

gives an easily understandable relationship with the motor output to follow the human's gaze shift.

Input to the edge-NN is represented as activities of four kinds of neurons that are selective to four orientations. Fig. 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

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

$E_n(x,\ y)$ 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 Fig. 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 Fig. 4 (b). The activities of the eight neurons $a_{f_n}^k$ $(n = 1,\ 2, \ldots,\ 8)$ in the k-th receptive field are calculated as

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

where $\boldsymbol{F}^k$ is given by (3), and $\boldsymbol{u}_n$ are unit vectors in eight directions. The activities of the neurons are also drawn as the length of the arrows in the circles as shown in Fig. 4 (b). The methods for codings edge and flow information are based on physiological evidence that the visual cortex in some animates has orientation selective neurons [7] and motion direction selective neurons [2]. The similarity in the representation of edge and flow inputs leads to the possibility that the robot may be able to translate a well-acquired sensorimotor coordination in the edge-NN or the flow-NN into the other.

Output from the edge- and flow-NNs is 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. The representation of the output neurons 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 rotate the robot's head by the coordinator.

C. Coordinator

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

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

where g_{pan} and g_{tilt} are scalar gains; u_{n_x} and u_{n_y} are the horizontal and the 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 rotate the robot's head is represented as displacement angles in the pan and tilt directions.

D. Learning Processing

The robot acquires the sensorimotor coordination to achieve joint attention with the edge- and flow-NNs through supervised learning. The learning processing assumes that the object the human is looking at can be detected in the peripheral image by using a given color definition as shown in Fig. 3 (a), and that the robot can gaze at the object to obtain the correct output. Note that the peripheral image cannot be used in joint attention experiments conducted after learning; that is, the robot cannot detect the position of the object that the human is looking at in the joint attention experiment. In learning processing, the robot encodes the motor command $\boldsymbol{\Delta\theta}$ obtained when looking at the object into the eight-direction selective neurons by using the inverse method to (6) and then independently learns the two NNs by back propagation. Fig. 3 (e) shows the motor output obtained when the robot changed its gaze from looking at the human to looking at the object detected in (a). The output data is used as the teacher signal for learning. The independent learning of the two NNs enables the robot to achieve joint attention using only one input, either edge or flow input.

III. Experiments

A. Robot and Experimental Setup

The validity of the model was evaluated using *Infanoid* [11], shown in Fig. 1, which was developed by our group as a tool for investigating the cognitive development of 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. In the experiments described here, Infanoid used two left camera images: a foveal image for extracting

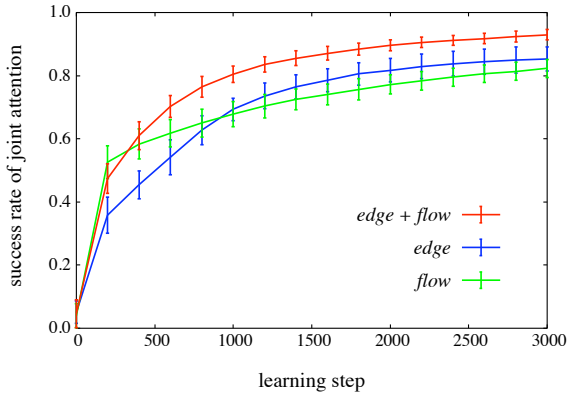


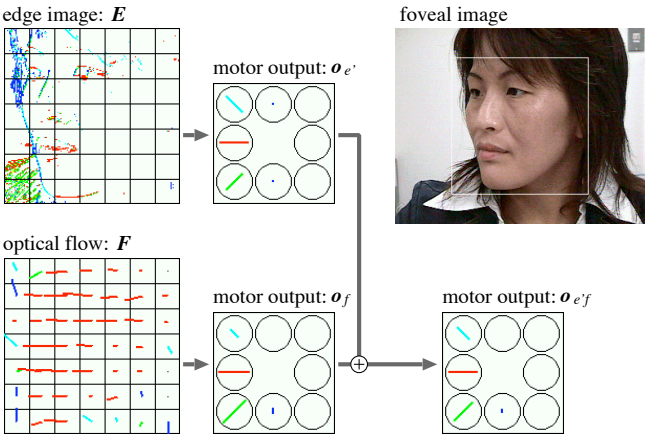
Fig. 5. The change in the task performance of joint attention over the learning period. The red, blue, and green lines indicate the results when the model utilized both edge and flow inputs, only the edge input, and the flow input, respectively.

the edge image and the optical flow of human gaze and a peripheral image for detecting the salient object that the human was looking at during learning processing. The three DOFs in the neck were used to change the robot's gaze direction 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. The human sat face to face with Infanoid and interacted with the robot 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. The human always looked at the object in front of her face.

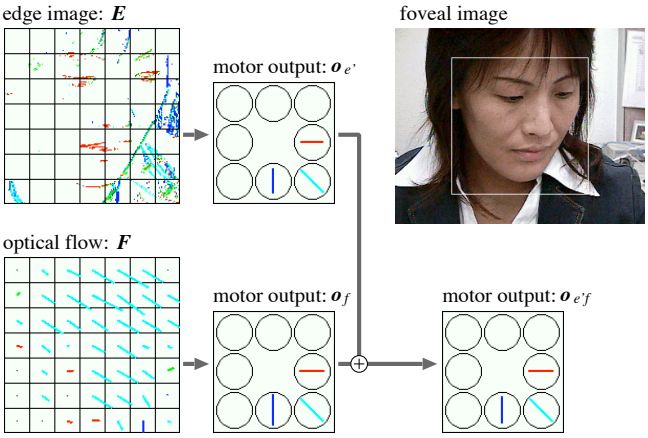
B. Learning Experiment

The model was first evaluated in the learning experiment. The experiment was conducted off-line by using 200 input-output datasets that Infanoid acquired beforehand. The datasets were repeatedly used for learning. Fig. 5 shows the changes in joint attention performance over the learning period, where the horizontal and the vertical axes respectively denote the learning step and the success rate of joint attention. The success of joint attention means that the robot looked at the object that the human was looking at within ± 8 degrees of error by using the acquired model. The red line shows the result when the model used both edge and flow inputs. The blue and green lines show the results when the model used only the edge or the flow input, respectively. Each of the three lines plots 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, we can see that the flow input accelerated the start-up time of learning while the edge input gradually improved the task performance. This complementary result was anticipated from the characteristics of the two inputs. As the result, by using both edge and flow inputs, the proposed model enabled the robot to quickly acquire the high performance of joint attention.



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



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

Fig. 6. The input-output datasets when the robot attempted to achieve joint attention by using the acquired model. The robot was able to establish joint attention in these two cases.

C. Joint Attention Experiments after Learning

The model acquired through learning using edge and flow inputs was evaluated in joint attention experiments. Fig. 6 (a) and (b) show the two cases of input-output datasets when the robot attempted to achieve joint attention. In case (a), the human shifted her gaze from looking straight 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 left side of each figure shows the input-output datasets of the edge- and flow-NNs, and the lower right shows the output from the coordinator, which is the mean value of the two outputs. From these results, we can confirm that the two NNs generated appropriate output to achieve joint attention. The success rate of joint attention with the same human in the learning experiment was 90% (18/20 trials) in random object positions.

IV. Analogy to Joint Attention Development in Human Infants

The experimental results showed that motion information can facilitate the learning of joint attention. We can find an analogy between the experimental results and development of joint attention in human infants.

Moore *et al*. [16] found that 9-month-old infants could be trained to follow an adult's gaze shift through trials in which the infant was given experiences of an adult's head turning in association with an interesting sight in the direction of the head turning. In their experiments, only infants presented with the movement of the head turning could acquire the gaze following behavior, whereas infants not presented with the movement could not acquire the behavior. Lempers [14] examined the developmental change in infants' (9- to 14-months-old) capability to comprehend others' pointing and gaze. They compared the capability when an infant was presented with the behaviors with and without movement. Their observational results showed that motion information enabled infants to understand the gaze of others. The importance of movement has also been pointed out in neonatal imitation, which precedes the joint attention development. Vinter [22] found that newborn infants were more likely to imitate tongue protrusion when they observed an adult gesture with movement rather than without movement. Meltzoff and Moore [15] showed that newborn infants turned or moved their heads laterally in response to an adult's head turning. Such head turning imitation by newborn infants could lead to gaze following behavior. All these findings in cognitive science and developmental psychology support the importance of movement in joint attention. The similarity between the cognitive development of infants and my experimental results suggest that the model could be helpful for understanding the developmental mechanism of joint attention in infants.

V. Conclusion and Future Work

This paper has indicated the importance of motion information in learning human-robot joint attention. Human beings utilize movement information detected from observing others in order to infer their desires and intentions and to establish natural communication. Furthermore, movement facilitates the development of joint attention in infants. Based on this knowledge, this paper proposed a joint attention learning model utilizing both motion and static information acquired from observing others' gaze shift. Experimental results demonstrated that motion information accelerated the learning of joint attention while static information improved the task performance.

The coordinator should be refined so that it can produce appropriate output according to a situation. The present model generates output as the mean value of the two outputs from the edge- and flow-NNs. The two NNs have advantages that complement each other. The coordinator will be re-designed so that it can take better advantage of the two NNs. Moreover, the learning experiments should be conducted in real-time and with natural interactions with humans. As infants do, a robot should be able to learn the sensorimotor coordination to achieve joint attention through interacting with several persons and several objects. Such learning will enable a robot to acquire more general and robust joint attention ability.

References

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

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

[3] Cynthia Breazeal and Brian Scassellati. Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, 8(1):49–74, 2000.

[4] G. E. Butterworth. Joint visual attention in infancy. In G. Bremner and A. Fogel, editors, *Handbook of infant development*, pages 213–240. Oxford: Blackwell, 2001.

[5] 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.

[6] 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.

[7] 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.

[8] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. Physical relation and expression: Joint attention for human-robot interaction. In *Proceedings of 10th IEEE International Workshop on Robot and Human Communication*, 2001.

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

[10] Frédéric Kaplan and Verena V. Hafner. The challenge of joint attention. In *Proceedings of the Fourth International Workshop on Epigenetic Robotics*, pages 67–74, 2004.

[11] 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.

[12] Hideki Kozima and Hiroyuki Yano. A robot that learns to communicate with human caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, 2001.

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

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

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

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