Proceedings of the 2005 IEEE International Workshop on
Robot and Human Interactive Communication
pp. 217-222, August 2005

Learning to Comprehend Deictic Gestures in Robots and Human Infants

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*— In this paper, I discuss how visual information about deictic gestures influences learning that enables these gestures to be comprehended. It has been suggested that the ability of human infants to comprehend deictic gestures depends on the physical appearance of gestures, the movement of gestures, and the distance between gestures and the indicated targets. To understand the mechanisms involved, I apply a model that enables a robot to recognize the static orientation of a gesture as an edge image and movement as optical flow. Experiments using a robot reveal that (1) learning to comprehend reaching gestures with all fingers extended is more accelerated than learning to comprehend pointing with the index finger, and that (2) downward tapping movement facilitates learning more than pointing movement along the direction of the gesture. These results suggest that (1) the quantitative difference in the edge features of reaching and pointing that correspond to the directions of gestures influences learning speed, and that (2) the optical flow of tapping movement that offers qualitatively different information from that provided by the edge image makes learning easier than the optical flow of pointing movement.**

***Index Terms*— deictic gesture, pointing, learning, image feature, infant development**

I. Introduction

Investigating how humans recognize gestures presented by others can provide ideas for understanding our communication mechanisms and for designing robots that interact naturally with us. I have an interest in the ability of joint visual attention and have been investigating how human infants acquire this ability from the viewpoint of cognitive developmental robotics [1]. Joint visual attention [2], [3] is a social communicative function to comprehend deictic gestures, such as pointing and gazing, presented by another person and to look at the same target as the person. Understanding the mechanism underlying joint visual attention can lead to an understanding of the developmental mechanisms for higher cognitive functions, e.g., language use and theory of mind. I have proposed some developmental models for joint visual attention and examined how the models enabled a robot to acquire this ability through interactions with humans [4]–[6]. The experimental results offered novel perspectives for understanding the developmental mechanisms of infants.

In this paper, I discuss how visual information about deictic gestures is recognized in learning to comprehend gestures. It is suggested that a robot is able to detect a deictic gesture presented by a human as an edge image and as optical flow. The former provides information about the static orientation of a gesture while the latter yields motion information about the gesture. What meanings the visual inputs have and how they are processed in learning to establish joint visual attention are analyzed in comparison with such developments in human infants. The next section first describes current knowledge about the development of gesture comprehension in infants. The roles of visual information are then analyzed using a robotic model, and the mechanism by which a robot reproduces the development of infants is explained. Finally, experimental results that verify the analysis are reported along with a discussion.

II. Comprehension of Deictic Gestures by Human Infants

The ability of infants to comprehend deictic gestures has been suggested to be influenced by the physical appearance of gestures, the movement of gestures, and the distance between gestures and targets to be indicated.

Woodward and her colleagues [7]–[9] investigated infants' understanding of the link between a deictic gesture and a target. They studied at what age infants came to understand gestures, such as grasping, pointing, and gazing, as object-directed actions. Their studies revealed that 6-month-old infants understood the relationship between a grasping gesture and the target [7], while infants came to understand pointing and gazing as object-directed actions between 9 and 12 months [8], [9]. These abilities involve an understanding of the intention of the person making the gesture. However, their results suggest that infants' ability to comprehend gestures as signal values is also influenced by the appearance of the gestures. Woodward [9] pointed out that infants' ability to comprehend depends on the physical connection between gestures and targets.

Lempers et al. [10], [11] investigated the ability of infants 9-, 12-, and 14-month-old to comprehend pointing and another person's line of gaze under several conditions of movement and distance to the target. They found that infants more correctly understood the directions of pointing and gazing when they observed gestures with movement

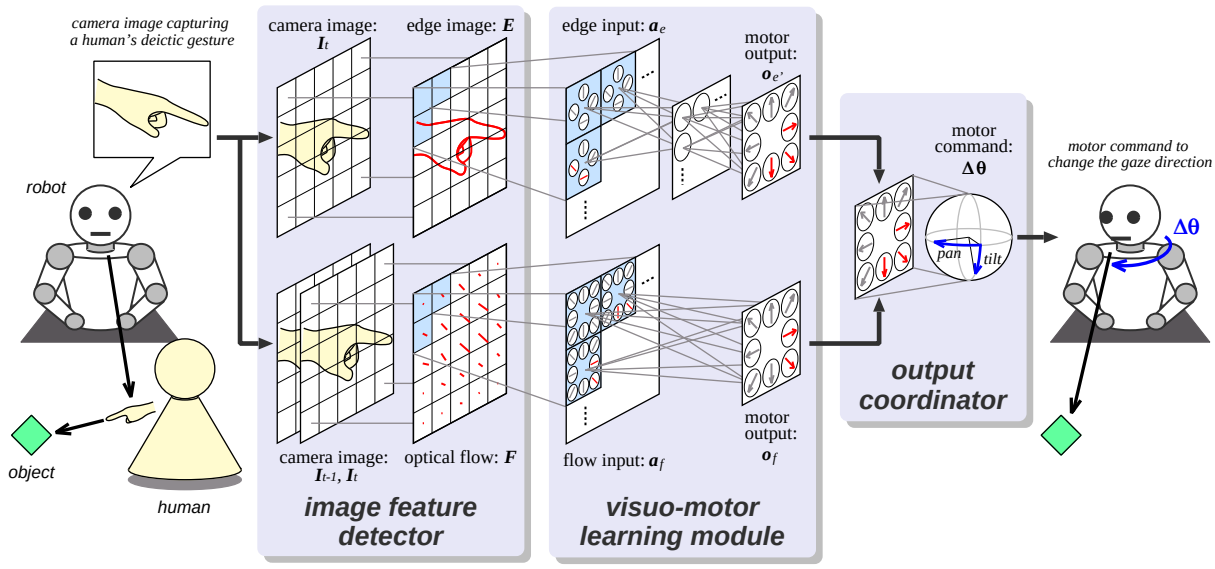


Fig. 1. Learning model for joint visual attention using edge image and optical flow detected when observing deictic gesture. The model enables a robot to follows the direction of a deictic gesture presented by a human. This was originally proposed by Nagai [6].

rather than without it. Moreover, infants appeared to develop the ability to comprehend pointing as a function of the distance between the gesture and the target. Moore et al. [12] also examined the role of movement in infants' learning to follow another's line of gaze. Nine-month-olds were trained to follow the direction of an experimenter's gaze while observing an action of experimenter's gaze sift with head turning movement, without movement, or with movement but with no static head orientation. Their experiment demonstrated that only infants who observed the action with movement, even without static head orientation, could learn to follow the experimenter's direction of gaze. The same effect of movement in learning to follow the direction of pointing was reported by Rohlfing et al. [13].

Butterworth and his colleagues [14], [15] investigated how infants' ability to establish joint visual attention changed with age. In their experiment, 12-month-old infants could follow their mother's gaze into their visual fields and look at the same object as she did whereas 6-month-olds were likely to look at a nearby salient object in the direction of her gaze. Only 18-month-olds, on the other hand, turned around to find an object when the mother was looking behind them. These results suggest that the ability to comprehend another person's gaze develops as a function of the distance between the gaze shift gesture and the target. Especially whether both gestures and targets can be simultaneously observed in the infants' field of view is a significant factor in their ability to establish joint visual attention.

In the next section, I present the learning model and describe how it enables a robot to acquire the ability to comprehend deictic gestures like infants. The mechanisms for how comprehension ability is influenced by the appearance of gestures, movement, and the distance between gestures and targets are explained using the model.

III. Robotic Learning Model for Comprehension of Deictic Gestures

Fig. 1 shows the model that was proposed in [6] by which a robot learns to establish joint visual attention with a human by comprehending his or her deictic gesture. The validity of the model was verified in a gaze following task, and Fig. 1 illustrates a situation where pointing is comprehended. Using the model, a robot learns sensorimotor mapping between a camera image $\boldsymbol{I}$ capturing a deictic gesture and a motor command $\boldsymbol{\Delta\theta}$ to follow the direction of the gesture. The input image features detected from $\boldsymbol{I}$ are an edge image $\boldsymbol{E}$ and optical flow $\boldsymbol{F}$. The former is used for estimating the static orientation of a gesture, while the latter is for estimating the motion direction. The sensorimotor mapping between the two image features and motor output is acquired by using two neural networks (NNs). Noteworthy characteristics of the model are:

- the NN for edge input consists of three layers while the NN for flow input consists of two layers,
- edge features and flow vectors are encoded in orientation or direction selective neurons when input to the NNs, and
- output from the NNs is represented in motion direction selective neurons.

Using input and output neurons that are selective to orientation or motion direction, sensorimotor mapping can be represented as understandable, e.g., one-to-one mapping between neurons with the same selectivity. The appropriate

numbers of layers enable the NNs to acquire accurate sensorimotor mapping for edge input and to learn fast mapping for flow input. Refer to Nagai [6] for more detailed explanations of the mechanism.

This section describes how the image features — an edge image and optical flow — are recognized in comprehending the directions of deictic gestures. I explain what mechanisms enable a robot to develop the ability to comprehend like infants.

A. Role of Visual Information in Comprehending Pointing

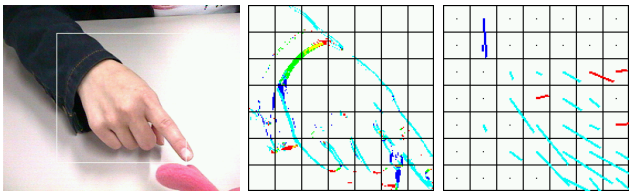
Fig. 2 shows an example of input image features detected when a robot is looking at a human gesture of pointing to an object in the lower right of the robot's view, in which (a), (b), and (c) show the camera image, the edge image detected from the center area of the camera image, and optical flow.

1) Edge Image: An edge image of pointing offers information to estimate the exact direction of the gesture. As confirmed from Fig. 2 (b), a robot can determine the direction in which the person is pointing by interpreting the contour of her hand. The information obtained from an edge image is the major orientation of edge features and the spatial dispersion of the features in the image. The former provides alternative directions of a pointing gesture, and the latter determines the correct direction. For example, the major orientation "$\diagdown$" detected in Fig. 2 (b) indicates that the person is pointing in the direction of "$\diagdown$" or "$\diagdown$," and then the spatial dispersion in the edge features determines the exact direction as "$\diagdown$." Note that the model in Fig. 1 does not process edge information in clearly separate ways as described here, but the NN is expected to acquire such well-organized recognition ability.

2) Optical Flow: The optical flow of pointing provides a rough but easily understandable motion direction of the gesture. Here, it is assumed that humans often move their hands from the axis of their bodies outward when pointing to a target. In Fig. 2 (c), the person is moving her hand from the front of her chest ahead to her left. This assumption enables the robot to find rough correspondence between the direction of optical flow and that of pointing. However, the correspondence is less accurate than the relation between an edge image and the direction of pointing. In contrast, optical flow is much more easily transformed into a motor command to follow the pointing direction because it has the same direction of motion. Thus, optical flow is expected to accelerate the learning of sensorimotor mapping to achieve joint visual attention.

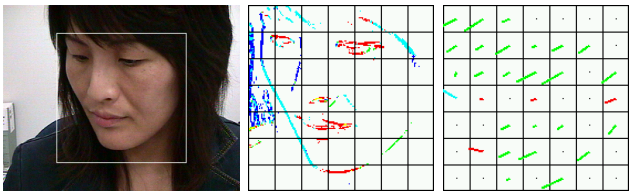
B. Role of Visual Information in Comprehending Gaze Direction

Fig. 3 shows an example of image features detected when a robot is looking at a human who is changing her gaze from looking straight at the robot's camera to looking at an object at her lower right.



(a) camera image (b) edge image (c) optical flow

Fig. 2. Input image features detected when looking at human pointing to lower right. The edge features in (b) and the flow vectors in (c) are colored according to the orientation or motion direction.

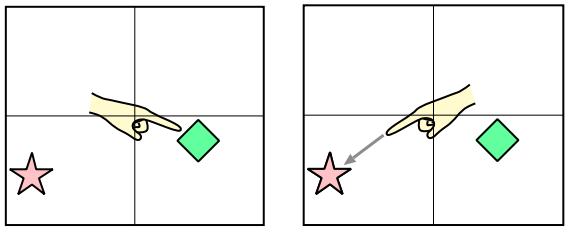


(a) camera image (b) edge image (c) optical flow

Fig. 3. Input image features detected when observing person's gaze shift to lower left.

1) Edge Image: An edge image of a human face offers information about the direction of the person's gaze. A robot can estimate the person's direction of gaze using the image in Fig. 3 (b), in which the contours of her face, eyes, nose, and mouth are extracted. The spatial dispersion of edge features especially provides position information about facial features, e.g., eyes and mouth, which enables a robot to infer the face direction. In contrast, the orientations of edge features cannot provide any useful information because the edge orientations of facial features do not change with respect to face direction. For example in Fig. 3 (b), the spatial dispersion of edge features extracted mostly in the left side of the image indicates that the person is looking to the robot's left. The gaze direction as well as the face direction can be estimated using an edge image with sufficient resolution.

2) Optical Flow: Optical flow detected when observing human's gaze shift provides information about the motion direction of the person's head and eyes turning. As confirmed from Fig. 3 (c), the flow yields a rough correspondence with the direction of the person's head. Here, it is assumed that humans often make eye contact with others before shifting their gaze. This assumption enables a robot to detect optical flow of which direction clearly corresponds to the gaze direction. However, the correspondence is less accurate than the relationship between an edge image and gaze direction. One reason is that optical flow does not include information about how much a person has



(a) pointing to nearby object (b) pointing to distant object

Fig. 4. Pointing to nearby/distant object. The large rectangle denotes the field of a robot's view when it is gazing at human pointing.

turned her head. Therefore, optical flow is only expected to accelerate the start-up time of learning by providing an approximate but clear correspondence with the gaze direction.

C. Effect of Distance between Deictic Gesture and Target

Why do infants' abilities to comprehend deictic gestures depend on the distance between the gesture and the target? This section explains what mechanisms enable a robot to go through the same developmental process as infants.

A robot is assumed to be embedded with the following mechanisms, which are based on knowledge obtained from infant studies (e.g., [16], [17], and [3]):

(i) *preferential looking* at motion, human faces, and salient colored objects detected in the field of the robot's view,
(ii) *gaze shift* to another target after observing a target for a certain period of time or after detecting constant motion of a target,
(iii) *target selection* with priority on how close the target and current gazing position are,
(iv) *sensorimotor learning* based on self-evaluation of experiences of looking at a preferred target, and
(v) *use of sufficiently acquired sensorimotor mapping* to follow the direction of a deictic gesture.

These mechanisms enable a robot to acquire the ability to comprehend pointing to a nearby object and to a distant object incrementally.

1) Pointing to Nearby Object: Fig. 4 (a) outlines a robot's camera image capturing a gesture of human pointing to a nearby square object. In this situation, the robot first looks at the pointing gesture using mechanism (i), and then shifts its gaze to the square object using (ii) and (iii). Through experience, the robot autonomously learns mapping between visual information about pointing and the motor command used when gazing at the square object based on (iv). It consequently acquires the ability to comprehend pointing to a nearby object.

2) Pointing to Distant Object: Fig. 4 (b) outlines a situation where a human is pointing to a distant star-shaped object. As in the previous situation, the robot first looks at the pointing gesture and then shifts its gaze to an object. If only the star-shaped object is observed in the robot's view, the robot looks at the object and learns to comprehend pointing to a distant object. On the other hand, if both the star-shaped object and the square are detected in the robot's view, the robot is likely to look at the square object using mechanism (iii). In such cases, the robot cannot learn to comprehend pointing because of the inconsistency between visual input and motor output. However, after acquiring the ability to comprehend pointing to nearby objects in every direction, the robot can apply the ability for comprehending pointing to a distant object based on (v). Thus, the robot develops the comprehension ability as a function of the distance between a pointing gesture and a target as infants. The ability to comprehend pointing outside the robot's view is also acquired based on the same strategy.

IV. Learning Experiments for Comprehension of Pointing

I conducted experiments to evaluate the roles of an edge image and optical flow in learning how to comprehend pointing. The evaluation of learning to comprehend another person's line of gaze was reported in [6]. Fig. 5 shows the experimental environment, in which a robot, called *Infanoid* [18], is looking at an object a human is pointing to. The robot is able to detect the pointing gesture by using the foveal camera in its left eye and is able to control its gaze direction using the six degrees of freedom in its eyes and neck. I compared learning performance with this experimental setup when the robot was presented with one of three gestures:

- *pointing* with the index finger moving in the indicated direction,
- *reaching* with all fingers extended moving in the indicated direction, and
- *tapping* with the index finger moving downward.

Pointing and reaching gestures were used to evaluate the effect edge features had, and pointing and tapping gestures were used to study the effect optical flow had.

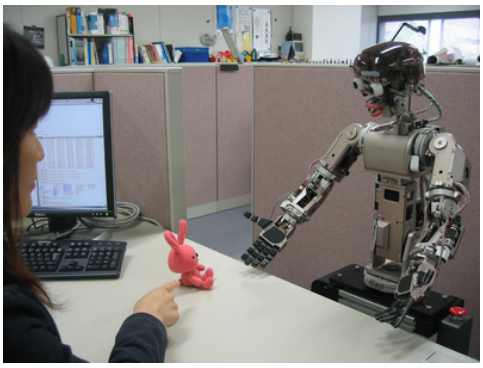
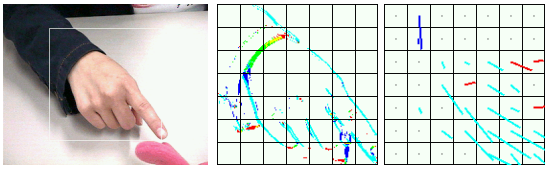
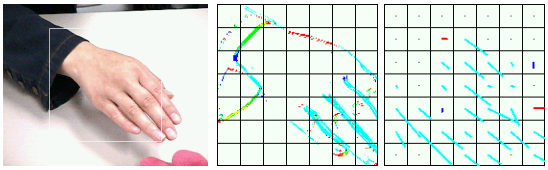


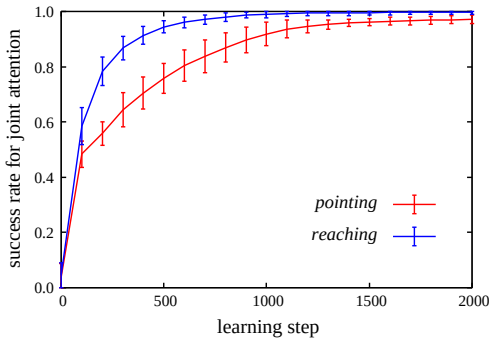
Fig. 5. Overview of experimental environment.



(a) image features of pointing (center: edge image, right: optical flow)

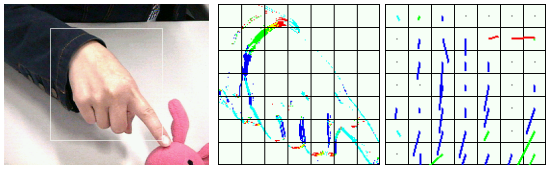


(b) image features of reaching (center: edge image, right: optical flow)

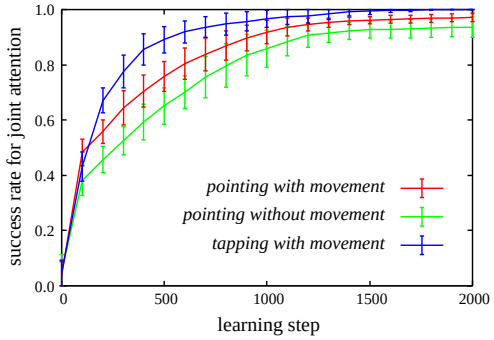


(c) changes in joint attention with learning

Fig. 6. Comparison of joint attention in learning to comprehend pointing and reaching.



(a) image features of tapping (center: edge image, right: optical flow)



(b) change in joint attention with learning

Fig. 7. Comparison of joint attention in learning to comprehend pointing with/without movement and tapping.

A. Evaluation of Role of Edge Features

The role of edge features in learning to comprehend pointing was investigated by using two patterns for gestures: pointing and reaching. Examples of image features for pointing and reaching are shown in Fig. 6 (a) and (b). Comparing the edge images, we can confirm that a reaching gesture provides more edge features that corresponds to the indicated direction than a pointing gesture does. Note that the optical flow for the two gestures was detected as having almost the same pattern.

Fig. 6 (c) plots the results, in which the learning performance of comprehension of pointing (red line) and reaching (blue line) are compared. The horizontal axis denotes the learning step while the vertical axis denotes the robot's success rate in establishing joint visual attention by following the direction of the gesture with sensorimotor mapping acquired through learning. The graph plots the mean value and the variance in 50 experimental results. From the graph, we can conclude that learning to comprehend reaching was more accelerated than learning to comprehend pointing. This may be because the edge images for reaching gestures included more edge features that corresponded to the directions of the gestures, which helped the robot estimate the static orientation of the gestures.

B. Evaluation of Role of Optical Flow

I investigated the role of optical flow in learning to comprehend pointing. The learning performance when the robot was presented with pointing with movement (see Fig. 6 (a)) was compared to those when pointing without movement or tapping (see Fig. 7 (a)) were presented. When learning to comprehend pointing without movement, the robot only used edge images of pointing gestures and not optical flow. Comparing the optical flow in Fig. 6 (a) and Fig. 7 (a), we can see that the movement of a pointing gesture was detected as flow vectors along the edge orientation of the index finger, while the movement of a tapping gesture was detected as downward flow vectors. Note that edge images of all gestures had almost the same features.

Fig. 7 (b) plots the changes in the success rate for joint visual attention over learning. The red, green, and blue lines show the results for the comprehension of pointing with movement, pointing without movement, and tapping, respectively. Comparing these results reveals that learning to comprehend pointing and tapping was facili-

tated by movement information. Both pointing and tapping movements helped the robot estimate the direction that the gestures were indicating. Moreover, the experimental results revealed that tapping movement accelerated learning more than pointing movement did. One reason may be that tapping gestures maintained a consistent positional relationship between the gestures and targets because of the touching actions, while the relationships varied with the pointing gestures. In addition, the optical flow for tapping movement offered different meaningful information from that of edge features. Whereas pointing movement was detected as optical flow with the same orientation as the edge features, tapping movement was detected with a different orientation from that of the edge features. These complementary features were considered to aid the robot in estimating the indicated direction and acquiring the sensorimotor mapping to establish joint visual attention.

V. Discussion

I demonstrated how an artificial model enabled a robot to recognize deictic gestures presented by a human and to learn sensorimotor mapping to achieve joint visual attention. The first experiment showed that the edge features of deictic gestures facilitated learning to comprehend the directions of gestures. Human infants are known to understand grasping gestures as object-directed actions earlier than they do pointing gestures [7], [8]. The experiment using a robot revealed that different forms of gestures, such as pointing and reaching, are quantitatively different in their edge features corresponding to the direction of the gestures, and that the difference influences the learning speed for comprehension of the gestures. The second experiment demonstrated that movement of deictic gestures helped a robot acquire the comprehension ability, and that tapping movement accelerated learning more than pointing movement. This empirically supports the knowledge that human infants correctly follow the directions of deictic gestures when observing gestures with movement rather than without movement [11]–[13]. Movement of deictic gestures detected as optical flow provides useful information for estimating the indicated directions. Moreover, tapping movement offers qualitatively different information from that provided by edge features. This property of movement is considered to make learning easier and the acquired comprehension ability more robust. The characteristic that tapping maintains a consistent positional relationship between the gesture and the target is also considered to assist robot learning.

Future work is to investigate the learning process by which humans also change how they present gestures to a robot. In human caregiver-infant interactions, caregivers often modify deictic gestures into understandable ones so that infants can respond to the gestures appropriately, and they also improve these gestures as infants develop. Interesting work has been done in which an infant chimpanzee was trained to follow the directions of deictic gestures presented by a human [19], [20]. The researchers enabled the chimpanzee to learn to comprehend gestures by presenting tapping, pointing, and gazing in stages. This means that the ability to comprehend gestures is acquired when gestures to be presented are appropriately ordered. This needs to be investigated based on how strategies deictic gestures should be presented and how the strategies are acquired.

References

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

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

[3] C. Moore and P. J. Dunham, editors. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[4] Y. Nagai, M. Asada, and K. 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.

[5] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, 15(4):211–229, 2003.

[6] Y. Nagai. The role of motion information in learning human-robot joint attention. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, pages 2081–2086, 2005.

[7] A. L. Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69:1–34, 1998.

[8] A. L. Woodward and J. J. Guajardo. Infants' understanding of the point gesture as an object-directed action. *Cognitive Development*, 17:1061–1084, 2002.

[9] A. L. Woodward. Infats' developing understanding of the link between looker and object. *Developmental Science*, 6(3):297–311, 2003.

[10] J. D. Lempers, E. R. Flavell, and J. H. Flavell. The development in very young children of tacit knowledge concerning visual perception. *Genetic Psychology Monographs*, 95(1):3–53, 1977.

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

[12] C. Moore, M. Angelopoulos, and P. Bennett. The role of movement in the development of joint visual attention. *Infant Behavior and Development*, 20(1):83–92, 1997.

[13] K. J. Rohlfing, M. R. Longo, and B. I. Bertenthal. Following pointing. does gesture trigger shifts of visual attention in human infants? Poster presented at the 14th Biennial International Conference on Infant Studies, 2004.

[14] G. Butterworth and E. Cochran. Towards a mechanism of joint visual attention in human infancy. *International Journal of Behavioral Development*, 3:253–272, 1980.

[15] G. Butterworth and N. 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.

[16] R. L. Fantz. The origin of form perception. *Scientific American*, 204(5):66–72, 1961.

[17] J. G. Bremner. *Infancy*. Blackwell, 1994.

[18] H. 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.

[19] S. Okamoto, M. Tomonaga, K. Ishii, N. Kawai, M. Tanaka, and T. Matsuzawa. An infant chimpanzee (pan troglodytes) follows human gaze. *Animal Cognition*, 5:107–114, 2002.

[20] S. Okamoto, M. Tanaka, and M. Tomonaga. Looking back: The "representational mechanism" of joint attention in an infant chimpanzee (pan troglodytes). *Japanese Psychological Research*, 46(3):236–245, 2004.