Acquisition of Joint Attention through natural interaction utilizing motion cues

Hidenobu Sumioka, Koh Hosoda[‡], Yuichiro Yoshikawa, and Minoru Asada[‡]

Department of Adaptive Machine Systems,
[‡]HANDAI Frontier Research Center,
Graduate School of Engineering,
Osaka University, 2-1 Yamadaoka, Suita, Osaka, 565-0871 Japan,
{*sumioka, yoshikawa*}*@er.ams.eng.osaka-u.ac.jp*
{*hosoda, asada*}*@ams.eng.osaka-u.ac.jp*

Abstract

Joint attention is one of the most important cognitive functions for the emergence of communication not only between humans but also between humans and robots. In the previous work [8], we have demonstrated how a robot can acquire the primary joint attention behavior (gaze following) without external evaluation. However, this method needs the human to tell the robot when to shift its gaze. This paper presents a method that does not need such a constraint by introducing an attention selector based on a measure consisting of saliencies of object features and motion cues. In order to realize natural interaction, self-organizing map for real-time face pattern separation and contingency learning for gaze following without external evaluation are utilized. The attention selector controls the robot gaze to switch often from the human face to an object and vice versa, and pairs of a face pattern and a gaze motor command are input to the contingency learning [8]. The motion cues are expected to reduce the number of the incorrect training data pairs due to the asynchronous interaction that affects the convergence of the contingency learning. The experimental result shows the gaze shift utilizing motion cues enables a robot to synchronize its own motion with human motion and to learn joint attention efficiently in about 20 minutes.

keywords: joint attention, motion cues, real time learning, asynchrony, bootstrap learning

1 INTRODUCTION

Joint attention, especially visual joint attention defined as looking at an object that someone else is looking at, is regarded as one of the building blocks for social capabilities such as language communication and mind reading in cognitive science and developmental psychology [1] since it appears to initiates the communication with other persons. For human–robot interaction, joint attention may have an important role to realize smooth communication with humans, and some robotics researchers have focused on this issue as fundamental ability for communication [2, 3]. Instead of explicit coding of joint attention behavior, we may expect such constructive approaches to enable robots to adapt themselves to changes in an environment including humans through the learning process, and also to provide a new understanding of the developmental process of joint attention in an infant [4].

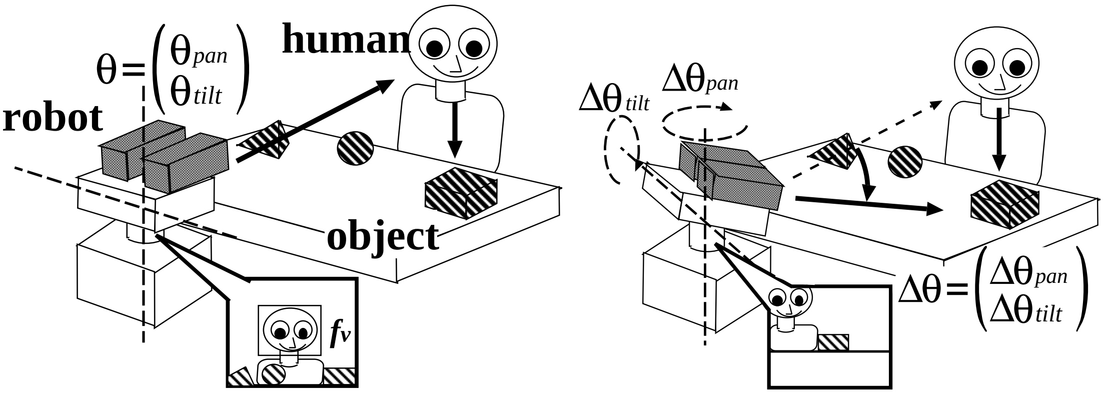
The design issues of joint attention are "where and when to shift the gaze," and the existing approaches have been focusing on only "where" issue by assuming the turn taking of gaze change between a human and a robot. Further, they can be classified into two categories: with and without external evaluation. In the former, reinforcement learning [5] or probabilistic algorithms [6] with task evaluation from a supervisor are utilized. In this category, the supervisor always needs to evaluate the robot's behavior. On the other hand, the second category [7, 8] does not need such evaluation by utilizing the consistency of the relationship between the other agent's gaze direction and the location of a salient object. Especially, Nagai *et al.* introduced a contingency learning mechanism which enabled a robot to acquire joint attention by learning sensorimotor mappings from a human face pattern to own motor command to gaze at an object [8]. However, these approaches are implemented in only computer simulation or not in real-time when applied to real robots. Further, they have not considered "when" issue by synchronizing turn taking of gaze changes between a human and a robot. In order to realize natural interaction between them, real-time interaction without a synchronization assumption should be considered. The issue is how to decide when to shift the gaze to achieve the joint attention with a human.

In this paper, we present a method that solves the issue by introducing an attention selector based on a measure consisting of saliencies of object features and motion information. In order to realize natural interaction that means real-time response without constrained synchronization of gaze shift between a human and a robot, self-organizing map (SOM) for real-time face pattern discrimination [9] and contingency learning for gaze following without external evaluation are utilized. The attention selector controls the robot gaze to switch often from the human face to an object and vice versa, and pairs of a face pattern and a gaze motor command are input to the contingency learning. The motion cues are expected to reduce the number of the incorrect training data pairs due to the asynchronous interaction that affects the convergence of the contingency learning [8].

The rest of this paper is organized as follows. First, we describe the task of joint attention between a human and a robot, and the problem addressed in this paper. Next, we give a learning architecture with an attention selector. Then, experimental results on a real robot are given. Finally, we discuss future issues and conclude the paper.

2 JOINT ATTENTION BETWEEN A HUMAN AND A ROBOT

The task environment of joint attention between a human and a robot is shown in Figure 1. The robot is sitting in front of the human, and there are some objects between them. The robot looks at the human's face pointed to an object and then captures the face pattern $\boldsymbol{f_v}$ from its camera image (see Figure 1 (a)). According to the face pattern $\boldsymbol{f_v}$, the robot calculates its head motion $\boldsymbol{\Delta\theta} = (\Delta\theta_p, \Delta\theta_t)$ to turn its head to the object (see Figure 1 (b)). Note that a face pattern $\boldsymbol{f_v}$ does not directly indicate an orientation of the face. To achieve joint attention, therefore, the robot needs to learn the sensorimotor mappings from $\boldsymbol{f_v}$ to $\boldsymbol{\Delta\theta}$.



$\theta = \begin{pmatrix} \theta_{pan} \\ \theta_{tilt} \end{pmatrix}$
human
robot
object
f_v
$\Delta\theta_{tilt}$
$\Delta\theta_{pan}$
$\Delta\theta = \begin{pmatrix} \Delta\theta_{pan} \\ \Delta\theta_{tilt} \end{pmatrix}$

(a) The human looks at an object, and the robot captures a face image pattern, $\boldsymbol{f_v}$.

(b) Based on $\boldsymbol{f_v}$, the robot outputs a motor command $\boldsymbol{\Delta\theta}$ to gaze at the same object the human is looking at (the success of joint attention).

Figure 1: Joint attention between a robot and a human.

In the previous work [8], this sensorimotor mapping was learned through interactions where the timing of gaze shift between the robot and a human was constrained to ensure consistency of the relation between a human face pattern and positions of the object that the human is looking at: that is, the human needs to tell the robot when to shift its gaze. Since we aim at more natural interactions between a human and a robot, we like to relax such a constraint. If each other's gaze shift is asynchronous, the relationship between a human face pattern and the robot's motor command is not always consistent. This means that it becomes difficult for the robot to learn the sensorimotor mappings because the number of pairs of the incorrect training data increases. To learn the sensorimotor mapping to perform joint attention, the robot needs to shift its gaze when the consistency of the relation between a human face pattern and positions of objects is ensured.

3 THE LEARNING ARCHITECTURE UTILIZING MOTION CUES

Instead of human instructor to tell the robot when to shift its gaze, we utilize motion cues to synchronize the turn taking of gaze change between the human and the robot. The proposed architecture is shown in Figure 2, where two key components are 1) an attention selector that decides which face or one of objects to gaze at and when to turn its head utilizing motion information, and 2) an online contingency learning module that enables to acquire joint attention by a spatial contingency within a certain time period [9].

Saliency filters extract different features from the captured camera images. Based on these features (including motion information), an attention selector decides where and when to gaze at. The position of a target (x, y) in the robot's view is sent to the visual feedback module (VFM) that outputs a motor command to gaze at the object. At the same time, an online contingency learning module (LM) outputs another motor command based on similarities between the captured face pattern and pre-categorized face patterns contained in a SOM, and on the robot's posture θ at that time. A gate selects one of these commands and then the robot behaves according to the selected motor command $\Delta\theta$.

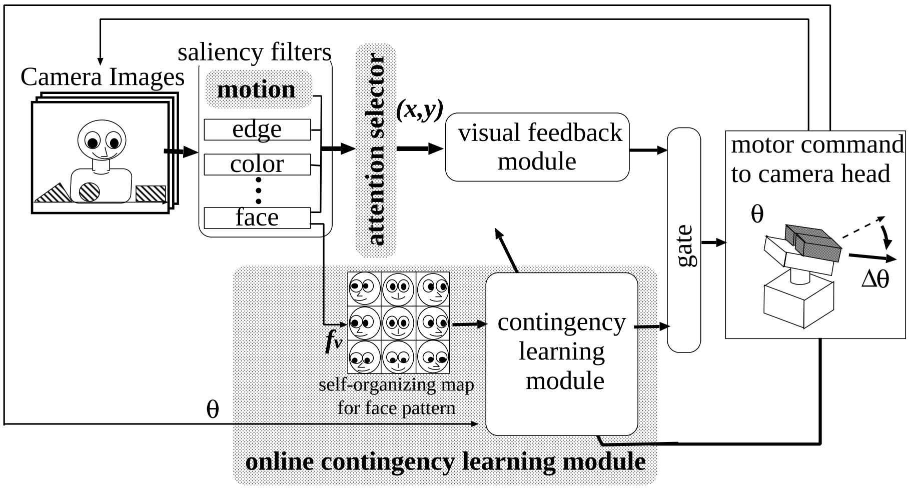

Camera Images
saliency filters
motion
edge
color
face
attention selector
(x,y)
visual feedback module
gate
motor command to camera head
θ
Δθ
contingency learning module
f_v
self-organizing map for face pattern
θ
online contingency learning module

Figure 2: An architecture for learning of joint attention through natural interactions based on motion cues.

3.1 LEARNING PROCESS

The robot shifts its gaze to the human's face or a salient object selected by the attention selector (described in the next section). Note that the robot is not programmed to direct its gaze alternately to the human's face and one of objects. Instead, the attention selector decides both which the human face or one of objects to gaze at and when to shift the robot's gaze. It is designed to regard the face as the most salient object because infants are supposed to have innate preference to human faces [10].

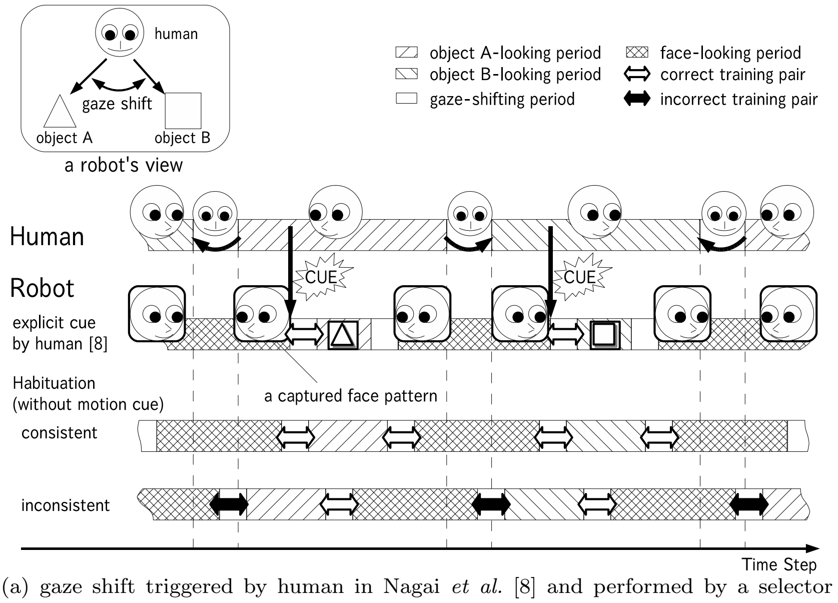
Consequently, the robot more often shifts its gaze between the human's face and an object.

The robot learns the sensorimotor mapping from the face patterns to the motor commands in an almost the same manner as in the previous work [8] but with an attention selector. Now, let the robot gaze at the human's face, and capture the face pattern. Then, it turns its head whenever triggered by the attention selector that utilizes motion cues as one of triggers to shift its gaze. The gate decides whether the robot adopts output from the online contingency learning module (LM) as motor command or not. We use a predetermined sigmoid function as the gate to represent the selecting rate of LM. At the beginning of learning, the gate selects the output from the visual feedback module (VFM) as the robot's motor command and the robot will turn its head to the most salient object that is determined by the attention selector. When it succeeds in gazing at the object around the center of the view, it strengthens the connection between the last face pattern obtained before shifting its gaze and the motor command to gaze at the object regardless of which output the gate selects. Here, this process also occurs in the case of gaze shift from an object to the face to have double chances to obtain the number of the training data pairs and, as a result, it is expected to accelerate learning of joint attention. As learning proceeds, the gate gradually comes to adopt the output from the LM more than one from the VFM.

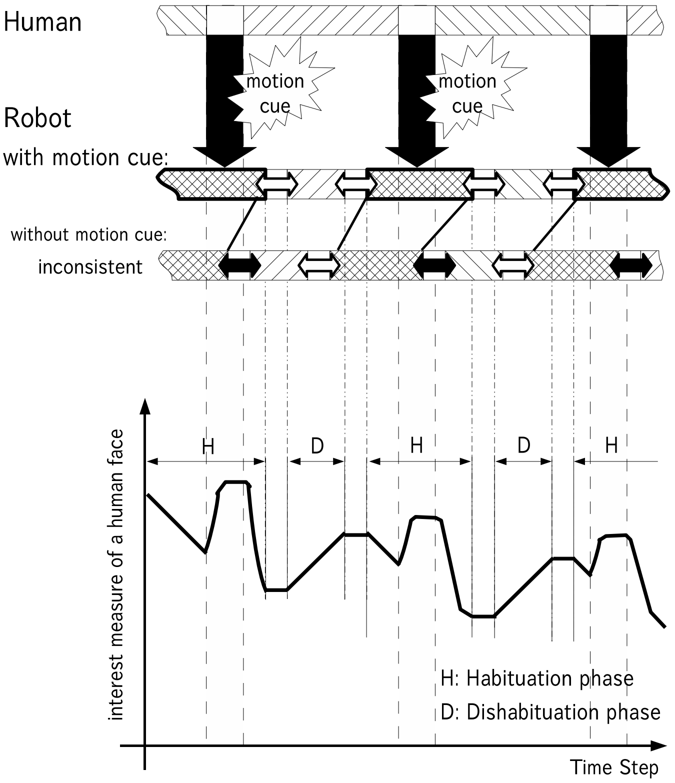
3.2 AN ATTENTION SELECTOR

(a) How it works

Although the previous architecture [8] included a mechanism to shift its gaze to the most interesting object, a robot was not able to shift its gaze to another object automatically without any cue from a human after gazing at the object (see Figure 3 (a)). In order to realize unconstrained interaction, we introduce an attention selector that is designed based on the phenomenon called habituation in developmental psychology. Habituation can be explained such that human infants lose the interest when they perceive the same stimulus for a while. Therefore, infants change their gaze directions to another stimulus. Some robotics researchers also point out that it is needed for the development of joint attention [7, 13, 14]. We define an interest measure for each object based on image features to model the habituation. The attention selector selects an object according to its selection probability that depends on the interest measure for the object. The higher the probability is, the more often the object is selected to gaze at. As the robot gazes at it, the measure gradually decreases, and then the robot shifts its gaze to another object that has higher interest measure than the current object.



(a) gaze shift triggered by human in Nagai *et al.* [8] and performed by a selector without utilizing motion cues.



(b) gaze shift performed with motion cues and transition of the interest measure of the human face.

Figure 3: Effects of motion information on the time periods of the robot's gaze shifts.

Habituation enables a robot to shift its gaze automatically. However, at the same time, it is necessary for the robot to find the periods when the human is gazing at an object and when to shift its gaze to the human or the object to learn joint attention through natural interaction with a human. "Natural interaction" means real-time response with unconstrained synchronization of gaze shift between the human and the robot. This is important especially in the case that the human moves the object that he/she is looking at to a different location: in such a case, since the robot needs the time to change its gaze, the robot may miss the timing to capture the correct pair of the human face pattern and the position of object. For example, Figure 3 (a) shows a simple example to indicate the difference between the gaze shifts triggered by a human and by an attention selector based on habituation but without motion cues.

In Figure 3, it is assumed that a human shifts the gaze alternately to the objects A and B at a constant frequency and a robot looks at the object A, a face, and the object B in order. Note that the robot captures a human face pattern both before its gaze shift from the human face to an object and after from an object to the face [1]. With the attention selector, there are two cases; the case where only correct pairs of the face pattern and the motor command are input to the learning system and the case where incorrect pairs are included: if the robot shifts its gaze during gaze shift by a human, it cannot learn the correct relation between the last face pattern obtained before shifting its gaze and the motor command output to gaze at an object (see solid both–side arrows).

To solve this problem, we construct the interest measure including not only object–specific image features, such as color and edge, but also motion information such as a human head turn or motions of objects manipulated by the human. In developmental psychology, there are some observations that an infant shifts its gaze utilizing an adult's head turn or the moving hand as well as motion of objects [11, 12] as one of cues of gaze shift. Therefore, this implementation is appropriate as a human infant model. Shifting the gaze based on motion cues enables a robot to change the timing of the gaze shift depending on the timing when the human shifts the gaze and picks up an object. Then, we designed the parameters of attention selector in a such a way that motion cue causes the rapid increase of the interest measure of a moving object or a turning face. If a robot gazes at the moving object or face, the robots gaze at it longer. If the robot does not, the robot shifts its gaze to it immediately. Figure 3 (b) indicates a simple example in the case where a robot gaze at the turning face. The top shows changes of robot's gaze shift based on motion cues. The bottom shows transition of the interest measure of the human face, where H and D indicate habituation and dishabituation phases, respectively. Note that interest measures between phases do not change because they are not calculated when the robot rotates its head. Motion cues about a human head turn increase the interest measure of the face, and the robot keeps gazing at the face until the human stops turning the head. As a result, an attention selector with motion information can provide a robot with more chances to obtain the correct training data pairs in the inconsistent case of Figure 3 (a) than one without motion information, and acceleration of learning joint attention is expected.

[1] It captures a face pattern only before it shifts its gaze from the face to an object in previous work [8].

(b) The mechanisms of the attention selector

The robot can extract a human's face image by detecting a face-like area, and extract objects by detecting object-specific features such as color and edge. These image features, including the face-like one, are candidates for the robot to gaze at.

Let n be the number of candidates for objects to be looked at in the robot's camera image. The interest measure $I_i(t)$ of each candidate is defined as

$$I_i(t) \quad = \quad M_i(t)S_i(t) \quad (i = 1, 2, \cdots, n, n+1), \tag{1}$$

where t is the sampling time, and the $(n+1)$-th candidate shows the interest measure of the human's face. $I_i(t)$ consists of the motion saliency, $M_i(t)(> 0)$, and the object-specific saliency, $S_i(t)(> 0)$. $M_i(t)$ denotes a value that is influenced by how long the i-th candidate moves until the sampling time t and is defined as

$$M_i(t) \quad = \quad g(m_i(t)), \tag{2}$$

where $m_i(t)$ represents the degree of motion and is defined as follows:

$$m_i(t) = \begin{cases} m_i(t-1) + 1 & (|\boldsymbol{f}_i| > \epsilon_i) \\ max\{m_i(t-1) - 1, 0\} & (|\boldsymbol{f}_i| \le \epsilon_i) \end{cases}, \tag{3}$$

where the flow vector $\boldsymbol{f}_i$ for the i-th candidate is calculated by optical flows, and ϵ_i is a small positive constant. Motion detection is prohibited when the robot rotates its head to avoid the confusion of motion detection due to its own motion or independent object motions.

In equation (2), the function g is a kind of threshold function. Here we use the following function:

$$g(x) \quad = \quad 1 + \frac{a}{1 + \exp\{(d - x)/T\}}, \tag{4}$$

where a, d, T are positive real numbers. The parameter a decides influence of motion information on the interest measure. The larger a is, the higher the probability of selection for the i-th candidate is when it moves. The parameter d is set to absorb noise about the flow vector and T decides the sensitivity to motion information. We set each parameter in terms that the function enables a robot to detect both human face motion and objects'.

The motion saliency, $M_i(t)$, changes the gaze duration of a robot, such as "object-A-looking period", "object-B-looking period" and "face-looking period" in Figure 3 significantly. If a human turns the head when the robot is gazing at the face, an attention selector with motion cues detects the timing of a human head turn and the motion saliency about the human face $M_{n+1}(t)$ increases. As a result, the robot keeps looking at the face until the head turn stopping because the interest measure about the human face $I_{n+1}(t)$ is increasing. This increase realizes the motion synchronization of shifting the gaze between the robot and the human to obtain the correct training data.

$S_i(t)$ shows the object-specific saliency of the i-th candidate. We set an initial value of $S_i(t)$ as

$$S_i(0) \quad = \quad C_i \quad (C_i > 0), \tag{5}$$

where C_i is a weighting constant to decide the basic bias for the robot to select the i-th candidate, that is, a preference to the candidate. We initialize the larger value of the human face than other objects' so that the robot can simulate the innate preference of infants to human faces [10]. $S_i(t)$ is defined as follows:

$$S_i(t+1) = \begin{cases} \alpha_i S_i(t) & \text{if the i-th candidate is attended} \\ max\{C_i, \beta_i S_i(t)\} & \text{else} \end{cases}, \tag{6}$$

where α_i ($0 < \alpha_i < 1$) is a decay factor while β_i (> 1) is a growth factor. Equation (6) means the object-specific saliency $S_i(t)$ gradually decreases during the robot continues to gaze at the i-th candidate and vice versa. The decay and growth factors for a candidate influence habituation and dishabituation phases, respectively as shown in the bottom of Figure 3 (b). It is expected that the robot shifts its gaze more frequently than the human because it can have the more opportunities to learn training data pairs in the situation where the consistency of the relationship between the other agent's gaze direction and the location of a salient object is ensured. The robot also needs to experience shifting the gaze alternately to the human's face and one of objects as much as possible to learn joint attention. Therefore, the interest measure of the human face should be designed to decrease and recover faster than the measures of other objects.

The robot calculates the interest measure $I_i(t)$ for each candidate. According to the interest measures, the selection probability $\Pr(i,t)$ for the i-th candidate is calculated as follows:

$$\Pr(i,t) = \frac{I_i(t)}{\sum_{j=1}^{n+1} I_j(t)}. \tag{7}$$

Note that the human's face and objects are not distinguished in the target selection process though they are different in learning process. Therefore, the robot sometimes shifts the gaze from one object to another or keeps gazing at the same target.

3.3 AN ONLINE CONTINGENCY LEARNING MODULE

An online contingency learning module strengthens the connection between a face pattern and own motor command to turn its head to an object. The point of this learning process is that the human does not provide the robot with any evaluation whether or not the connection is appropriate to acquire joint attention. In addition, the robot cannot explicitly find which object the human looks at. That is, through the learning process, the contingency learning module strengthens not only relevant connections but also irrelevant ones. Nagai *et al.* [8], however, shows if positions of objects change randomly and the human gazes at objects, the relevant connections to acquire joint attention are more strengthened than irrelevant ones because there exists a contingency between a face pattern and the position of the object that the human looking at. As a result, the robot can acquire joint attention based on this contingency. We leave the details to Nagai *et al.* [8].

Instead of the high-dimensional face image matching [8] that consume a large amount of computation, we utilize a SOM of face patterns [9]. In advance, we make a robot learn the SOM to categorize face patterns in which each neuron represents a vector of a gray scale face image. As inputs of learning of joint attention, we utilize the activations of each neuron calculated based on the similarity with a face image that the robot is gazing at. Figure 4 shows a network, where two-layered perceptron with an SOM input layer is learned through backpropagation by utilizing the robot's motor command as reference signal. The compression of input dimension by the SOM enables the robot to discriminate face pattern in real time.

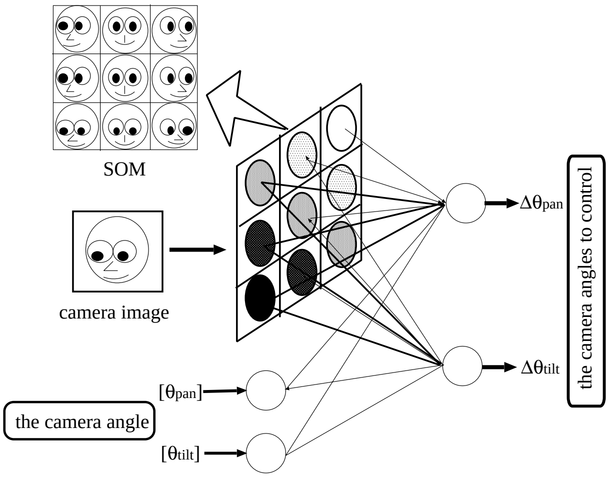


Figure 4: Online contingency learning module: a robot learns the relation between the activations of individual neurons in the SOM calculated based on the similarity with the captured face pattern and its motor command.

4 EXPERIMENTS

4.1 ENVIRONMENTAL SETUP

The experimental setup is shown in Figure 5. The robot and the human are seated face-to-face. Throughout the experiment, the distance between a human and a robot is constant. Four objects with different colors are placed on the table between them. The robot head has two degrees of freedoms (DOFs): the pan and tilt. A CCD camera (Firefly produced by Point Grey) on the head provides 320 × 240 color video images at 30 frames per second. Note that the horizontal and vertical angles of view of the camera are about 61.9 and 48.5 degrees, respectively, and these angles are wide enough for the robot to capture both the human face and objects on the table. The template matching method is used as face detector and a 32 × 32 pixel face-like region is extracted. Also, the color areas are extracted as object regions, and an optical flow by the block matching method is detected.

The robot learned the SOM to categorize face patterns within three minutes before it learns joint attention. Figure 6 shows a learned SOM used in the later joint attention learning. The learned SOM consists of 9 × 9 clusters, each of which is constituted by a 32 × 32 pixel gray scale image based on the

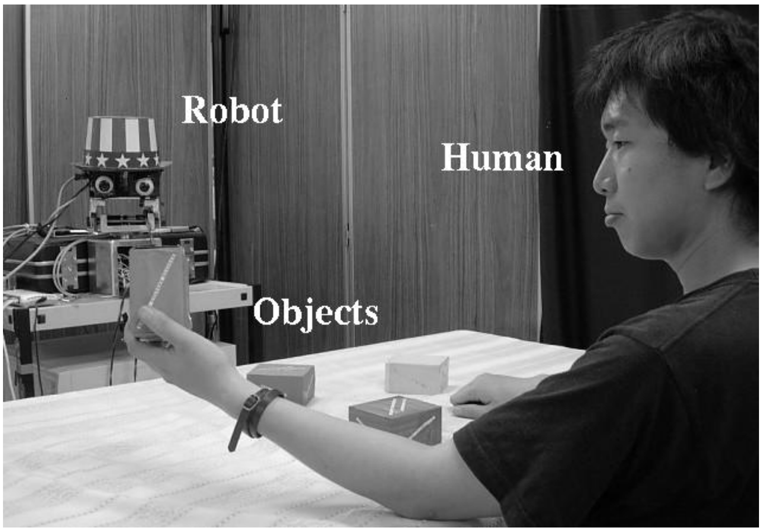


Figure 5: The experimental setting: the robot and the human are seated face-to-face and between them there are four objects with different colors.

face-like region extracted by the face detector.

In the following experiments, we assume the robot can always observe a human face and some objects in the field of view. Table 1 shows parameters used in the attention selector, and parameters in the threshold function $g(x)$ (eq. (4)) were set as $(a, d, T) = (4.5, 20, 1.4)$. The robot took about one second to shift its gaze from one target to another.

In addition, we used a sigmoid function as a gate. The robot decides whether it adopts the output from the online contingency learning module as a motor command according to the probability Pr_g:

$$Pr_g(l) = \frac{1}{1.0 + \exp\{(p-l)/q\}}, \tag{8}$$

where l is the number of learning iteration. As the learning proceeds, Pr_g becomes higher. As a result, the robot gradually comes to adopt the output from the learning module. Each of parameters in the gate function decides learning time. Before experiments, therefore, we performed preliminary experiments to determine the parameters of the gate in an environment where the robot could learn most easily, that is, the timing of gaze shift between the human and the robot is synchronized completely. Based on the result, parameters of Pr_g were set as $(p, q) = (150, 22.5)$ (see Figure 7). This represents selecting rate of learning module's output reaches to 50% at the 150th learning step.

Table 1: parameters of attention selector.

Candidate	C_i	α_i	β_i
the human's face	1500	$\exp(-2.0 \times 10^{-2})$	$\exp(1.2 \times 10^{-2})$
the object: A (red), B (yellow), C (blue), D (green)	800	$\exp(-1.0 \times 10^{-2})$	$\exp(2.0 \times 10^{-3})$

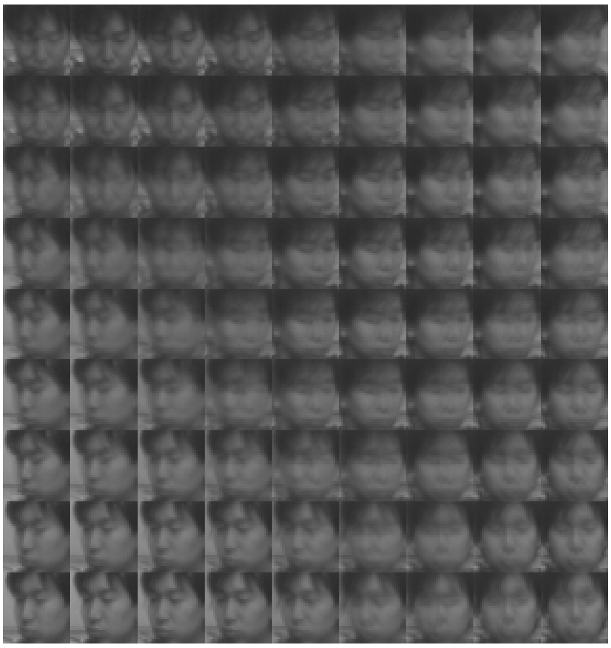


Figure 6: A learned SOM of the face patterns.

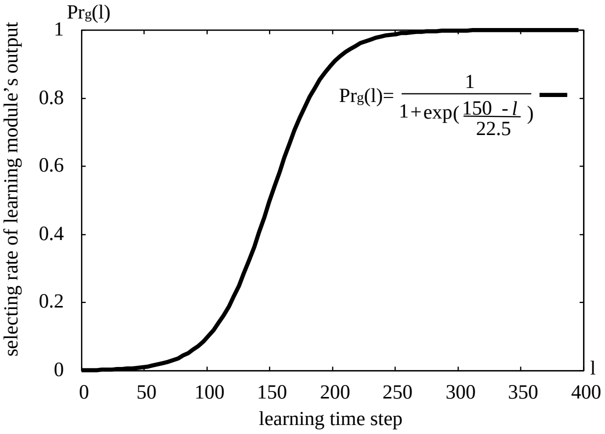


Figure 7: The gating function used in the experiments.

4.2 HUMAN BEHAVIOR

The task for the human is the object-transfer task: a person (here, a male) randomly selects one object and directs his gaze to it. Next, he picks it up and observes it for about two seconds, and then, puts it somewhere on the table, gazes at it for about two seconds, and selects another object. Note that the object manipulated by him is arranged in different positions of the table as evenly as possible and a moving object is only what he is manipulating. The robot also shifts its gaze to one of the objects and the person's face according to the decision of its attention selector.

4.3 LEARNING JOINT ATTENTION

We investigated whether the robot could acquire joint attention through human–robot natural interaction. To validate an effect of motion cues, we compared performances between the architectures with and without motion cues five times. Note that the architecture without motion cues utilizes only object-specific features to select a target. In each session, we counted whether the robot was able to perform joint attention with the human or not when it directed its gaze from his face to an object. Each session lasted approximately 26 minutes. The average number of the robot's gaze shift was 302.0 times with motion cues and 279.6 times without them. The standard deviations were 6.87 and 8.36, respectively. Also, the average numbers of success of joint attention were 199.6 and 124.4, respectively and the standard deviations were 6.25 and 19.26, respectively.

Figure 8 shows the averages and standard deviations in five sessions of moving averages of the success rate of joint attention in terms of with/without motion cues. Each moving average at a given time t minutes, $mov\_ave(t)$, in one experiment was calculated as follows:

$$mov\_ave(t) = \frac{\text{number of joint attention from t-1 to t+1}}{\text{number of robot's gaze shift from t-1 to t+1}}. \tag{9}$$

In Figure 8, '×' and '+' indicate the results with and without motion cues, respectively. The vertical bar at each point represents the standard deviation of five sessions. Note that the success rate at the beginning of learning includes the success of joint attention by visual feedback. While, the success rate at the end of learning indicates the performance by the online contingency learning module. We can see the gaze shift with motion cue significantly improves the performance over without motion cue, and the success rate of joint attention by the proposed architecture reaches 80% after about 20 minutes. Most of failures happened when the robot gazes at a distractor very close to the target. Here, distractors mean the other objects that the subject does not gaze at.

Although the subject did not exactly behave in the same manner, we observed the same tendencies in five sessions in spite that a robot experienced different timing and frequency of human gaze shift. The results with other subjects also showed the same tendencies. Therefore, the proposed architecture may have the validity in experimental environment.

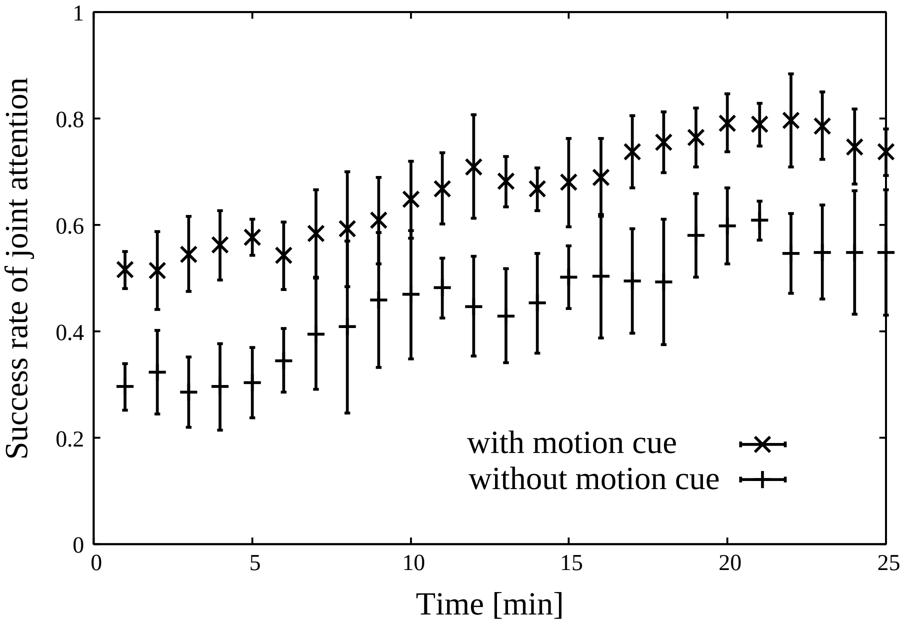


Figure 8: The time courses of success rate of joint attention through interaction with a human.

5 DISCUSSION AND CONCLUSION

In our approach, we utilized the motion information expecting to accelerate learning of joint attention, and we obtained successful results. In developmental psychology, it is suggested as one of precursors of joint attention that 3– and 4– month-olds, who do not have an ability of joint attention with adults, often shift their gaze to adults' moving hand and/or an object in their hand [12]. Therefore, it is plausible that shifting the gaze based on motion information increases the chances to obtain the consistent training data pairs and helps infants to acquire joint attention.

As mentioned in the previous section, most failures are caused by distractors near by the target in the image. If they were distant from the target in 3–D space, these failures might have been avoided by using the depth cues from binocular vision system. In addition, if the robot knows the human attention strategy model through interactions with him/her, the robot might be able to find the target correctly. Shon *et al.* propose a model that enables a robot to learn instructor–specific saliency models by performing joint attention with a human but they need the evaluation for robot's behavior [6]. Without such evaluation, we should build a learning model that can acquire both joint attention and an ability to infer other's preference.

In our experiments, the designer specified the parameters of habituation such as how long the robot gazes at an object. These parameters should be estimated for the robot to be synchronized with human behaviors (head turn and object transfer) through real interactions. Carlson and his colleagues propose a simulation model that can synchronize the gaze shift by a reinforcement learning method [7, 13], and we may apply their method to estimate the parameters for the synchronization.

The attention selector directly utilized the motion cues in the object–transfer task. In the behavior such as the object manipulation, it is supposed that coordination of eye and hand movements has a temporal structure [15]. Therefore, such a structure might be useful for more accurate synchronization due to the capability of prediction of motion sequences. Furthermore, if the pace of each motion can be estimated through interactions, more adaptive synchronization might be possible depending on situations. Actually, the caregivers may change the paces of their motions to adapt themselves with children's behaviors [16].

In Figure 3 (b), the motion cue is used in one way from the human to the robot, but actually the human caregiver is also affected by the robot behavior. In Figure 8, the performance without motion cues appears to have slightly improved by accident due to this effect. We need to observe human–robot interaction with and without this effect and utilize the result to build a robot that can acquire shared joint attention through more natural interaction with a human.

REFERENCES

[1] C. Moore and P. J. Dunham (Eds.), *Joint attention: It's origins and role in development*, Lawrence Erlbaum Associates (1995).

[2] F. Kaplan and V. Hafner, The Challenges of Joint Attention, In *Proc. The Fourth International Workshop on Epigenetic Robotics*, Genoa, Italy, pp.67-74 (2004).

[3] B. Scassellati, Theory of mind for a humanoid robot, In *Proc. the First IEEE-RAS International Conference on Humanoid Robots*, Cambridge, MA. (2000)

[4] 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**, pp.185-193 (2001)

[5] G. Matsuda and T. Omori, Learning of Joint Visual Attention by Reinforcement Learning, In *Proc. International Conference on Cognitive Modeling*, Fairfax, Virginia, USA, pp.157-162 (2001)

[6] A. P. Shon, D. B. Grimes, C. L. Baker, M. W. Hoffman, S. Zhou, and R. P. N. Rao, Probabilistic Gaze Imitation and Saliency Learning in a Robotic Head, In *Proc. IEEE International Conference on Robotics and Automation*, Barcelona, Spain, pp.2876-2881 (2005).

[7] E. Carlson and J. Triesch, A computational model of the emergence of gaze following, In *Proc. the 8th Neural Computation and Psychology Workshop*, Canterbury, England (2003).

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

[9] A. Morita, Y. Yoshikawa, K. Hosoda, and M. Asada, Joint attention with strangers based on generalization through the joint attention with caregivers, *the IEEE/RSJ International Conf. on Intelligent Robots and Systems*, Sendai, Japan, pp3744-3749 (2004).

[10] J. G. Bremner, *Infancy: 2nd Edition*, Oxford: Blackwell (1994).

[11] C. Moore, M. Angelopoulos, and P. Bennett, The Role of Movement in the Development of Joint Visual Attention, *Infant Behavior and Development*, **20**, 83-92 (1997).

[12] S. Amano, E. Kezuka, and A. Yamamoto, Infant shifting attention from an adult's face to an adult's hand: a precursor of joint attention, *Infant Behavior and Development*, **27**, 64-80 (2004).

[13] C. Teuscher and J. Triesch, To Care or Not to Care: Analyzing the Caregiver in a Computational Gaze Following Framework, In *Proc. the 3rd International Conference on Development and Learning*, La Jolla, CA. (2004)

[14] I. Fasel, G. O. Deák, J. Triesch, and J. Movellan, Combining Embodied Models and Empirical Research for Understanding the Development of Shared Attention, In *Proc. of the 2nd International Conference on Development and Learning.*, Cambridge, MA, 21-27 (2002).

[15] J. Pelz, M. Hayhoe, and R. Loeber, The coordination of eye, head, and hand movements in a natural task, *Experimental Brain Research*, **139**, 266-277 (2001).

[16] K. J. Rohlfing, J. Fritsch, B. Wrede and T. Jungmann, How can multimodal cues from child–directed interaction reduce learning complexity in robots?, *Advanced Robotics*, **20**(10), 1183-1199 (2006).