

Acquisition of the Head-Centered Peri-Personal Spatial Representation Found in VIP Neuron

Sawa Fuke, Masaki Ogino, and Minoru Asada, *Fellow, IEEE*

Abstract—Both body and visuo-spatial representations are supposed to be gradually acquired during the developmental process as described in cognitive and brain sciences. A typical example is face representation in a neuron (found in the ventral intraparietal (VIP) area) of which the function is not only to code for the location of visual stimuli in the head-centered reference frame, but also to connect visual sensation with tactile sensation. This paper presents a model that enables a robot to acquire such representation. The proprioception of arm posture is utilized as reference data through the “hand regard behavior,” that is, the robot moves its hand in front of its face, and the self-organizing map (SOM) and Hebbian learning methods are applied. The simulation results are shown and discussions on the limitation of the current model and future issues are given.

Index Terms—Body representation, learning and adaptive system, sensor fusion, ventral intraparietal (VIP) neuron.

I. INTRODUCTION

ACQUIRING BODY representation is the most fundamental issue not only for robotics, in order to accomplish different kinds of tasks, but also for cognitive and brain sciences and related disciplines, since how humans acquire such representation is one of the great unresolved issues of human cognitive development. General consensus of body representation is roughly categorized into two types: “body schema,” an unconscious neural map in which multimodal sensory data is unified, and “body image,” an explicit mental representation of the body and its functions [1], [2]. The body representations in biological systems are apparently flexible and acquired by spatio-temporal integration of different information from different sensory modalities (i.e., [3] and [4]).

Among different modalities, vision is the most representative spatial perception that is expressed in various kinds of reference frames in different brain regions. A typical example is that the visual stimulus of a target is perceived in a retinotopic manner [5]. On the other hand, the adjacent ventral intraparietal (VIP) area includes neurons which encode bimodal sensory information in a head-centered space coordinate system [6]–[8]. They are supposed to play an important role for the avoidance

Manuscript received December 19, 2008; revised August 13, 2009. First published August 25, 2009; current version published October 21, 2009.

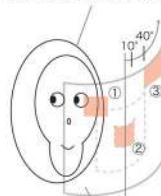
S. Fuke and M. Asada are with the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Osaka 556 0017, Japan (e-mail: sawa.fuke@ams.eng.osaka-u.ac.jp; asada@ams.eng.osaka-u.ac.jp).

M. Ogino was with Asada Synergistic Intelligence Project, ERATO, Japan. He is now with the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Osaka 556 0017, Japan (e-mail: ogino@jeap.org).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TAMD.2009.2031013

The surface of the subject (monkey)' s face is touched



Visual stimuli are displayed on the screen

The examples of receptive fields of VIP neurons

	Somatosensory receptive field	Visual receptive field
Neuron A	Around the right eye	In the area ① on the screen
Neuron B	Around the nose	In the area ② on the screen
Neuron C	Around the left blow	In the area ③ on the screen
⋮	⋮	⋮

Fig. 1. The table shows the examples of visual and somatosensory receptive fields of neurons in VIP area that are observed in [6]. The same VIP neuron is activated when something is shown on the screen's shaded area in front of the monkey and when the face's shaded area is stimulated regardless of where the monkey is fixating.

of obstacles and projectiles. Not only tactile stimuli on the face, but surprisingly also visual stimuli, whose locations can be expressed in a head-centered reference frame regardless of ocular angles, can activate these neurons. Fig. 1 shows examples of the visual and somatosensory receptive fields of the same neuron, which are not affected by gaze directions.

We may hypothesize that in the brain, the locations of visual stimuli on the retina are transformed to the locations in the abstract reference frames by integrating them with the proprioception (i.e., neck and ocular angles) and associated with other sensor information (i.e., tactile sense). Intriguingly, it is suggested that not only the body image (schema), but also this transformation system between reference frames in the visuo-space is also adaptively acquired through experiences (i.e., [9]). However, the way humans acquire such representations in the brain in spite of the changes in body structures and sensitivities has remained an issue to be revealed.

A number of synthetic approaches aiming at understanding the acquisition process of body and visuo-spatial representation in humans have been attempted in cognitive developmental robotics [10], where the self-body, or body parts, are found or identified based on invariance in the sensor data [11], synchronization and contingency of motion and perception [12]–[14], Jacobian estimation [15], and reference frame transformation [16]. In these studies, the representation of invisible body parts, such as a face or a back, cannot be acquired because the robot cannot detect the visual information of the surface directly with their own cameras. Fuke *et al.* [17] proposed a model that acquired the body representation of a robot's invisible face by estimating its hand position from the change of the proprioception while touching its own face. However, these studies assumed that camera positions are fixed or that the coordinate system in visual space is given by the designer.

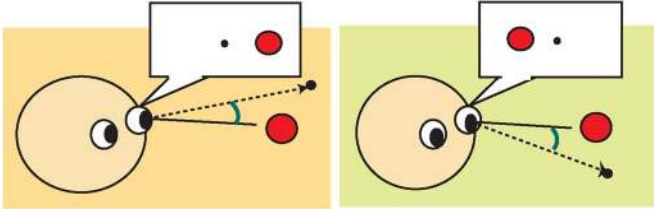


Fig. 2. In these two situations, we humans can recognize that the red objects locate at the same position, though ocular angles and retina image are different when we detect it in the peripheral visual field. Actually, when we see something, there are many different sets of perceived information (i.e., ocular angles and retina image).

As a learning model of visuo-spatial representation, Pouget *et al.* [18] proposed an approach based on a neural network with statistically distributed input data so that multi-modal sensations can be integrated. However, they have not discussed what kind of information can be used to select the signals pertaining to the same location of the target. In fact, as shown in Fig. 2, adult humans can recognize that a stationary object is located in the same position though ocular angles and retina image are different when we detect it in the peripheral visual field.

Aiming at revealing the above issue, through the process of mutual feedback between hypothesis generation and its verification, here we propose a learning model in which a robot acquires not only the head-centered reference frame, but also the multimodal representation of the face based on the knowledge in neurophysiological and cognitive science by focusing on a “hand regard” behavior that infants around 4-months-old often show. Eventually, we hope that the properties of acquired cross-modal representation are similar to the one of VIP neurons found in neuroscience. The proprioception of arm posture is utilized as reference data through the “hand regard” behavior, that is, the robot moves its hand in front of its face, while the self-organizing map (SOM) and Hebbian learning methods are applied. The SOM algorithm was proposed by Kohonen [19] who suggested that cortical maps may self-organize in a nearest-neighbor relationship. Based on this assumption, Aflalo *et al.* [20] actually modeled motor cortex topography using a Kohonen SOM and argued that their maps resembled the actual maps obtained from the lateral motor cortex of monkeys. Here, we used SOM for data compression. The simulation results are shown and discussions on the limitation of the current model and future issues are given.

II. FINDINGS IN DEVELOPMENTAL SCIENCE CONCERNING THE VISUO-SPATIAL REPRESENTATION

Observation study suggests that the visual abilities of human infants develop dramatically from the age from 3- to 7-months-old. The 3-month-old infants tend to plan saccades based on the retinocentric reference frame, ignoring the target shift due to eye movements. On the other hand, the 7-month-old infants do not ignore it [21]. This implies that human infants do not seem to have the visuo-spatial representation within certain reference frame systems from the beginning, but acquire it through their experiences while their strength of muscles, and sensitivity and placement of sensory organs continue to change. While these visual abilities develop, a typical infant behavior called “hand

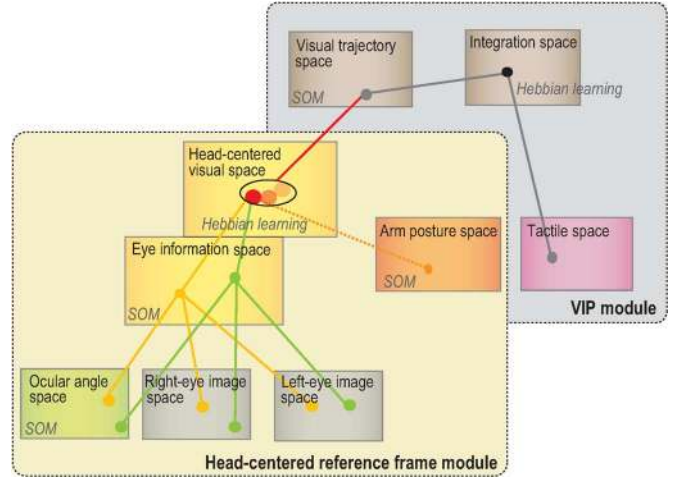


Fig. 3. An overview of the proposed model.

regard” [22] is observed. “Hand regard” is the phenomenon in which 3- or 4-month-old infants often gaze at their own hands in front of their faces. Among many interpretations of this phenomenon, Rizzolatti *et al.* [23] suggest that it is probably to be ascribed to the necessity of calibrating “peri-personal space” (defined as the space within reach of the arm [24]) around a body by combining the motor and visual information. The VIP area in the parietal cortex is known as the region that contains this peri-personal visuo-spatial representation.

Considering these observations of infants, we propose a learning model that starts from the retinocentric representation, to head-centered representation, through “hand regard” behavior as shown Table I. Then, we approach the issue of identifying what actual mechanism leads to the development of the visuo-spatial representation development in the infant’s brain. In our simulation, a robot learns the association between the tactile representation of the face and the learned visuo-spatial representation, which enables the robot to show the reflexive behavior like the VIP neurons.

III. VIP NEURON MODEL

An overview of the proposed model is shown in Fig. 3, where two modules are involved. First, the robot acquires the head-centered reference frame module. It has many sets of ocular angles and the retinotopic images (camera images) that are represented in the eye information space in Fig. 3. In order to construct a head-centered reference frame, the robot associates the ocular angles and camera images by regarding the proprioception of its own body (joint angles of the arm) as the reference information.

Next, in the VIP module, the robot integrates the tactile sensation with the patterns of visual stimuli computed in the head-centered reference frame in the former trained module when it touches its own face with its hand. Finally, the robot can acquire the cross-modal representation of its own face. The details of the robot simulator used and the details of each module are given in the following Sections III-A, III-B, and III-C, respectively.

A. Robot Simulator

In order to validate the model, computer simulations were conducted with a dynamic simulator based on the method of Feath-

TABLE I
 THE FINDINGS THAT SUPPORT THE CONDITIONS OF SIMULATION

Month	Observed behavior (infant)	Situation of the robot experiment
3	Saccadic movement in the retinocentric reference frame	Perceiving ocular angles and camera image data
3,4	"Hand regard" behavior	Moving its own hand and watching it
7	Saccadic movement in the body-centered reference frame	Integrating ocular angles and camera image data

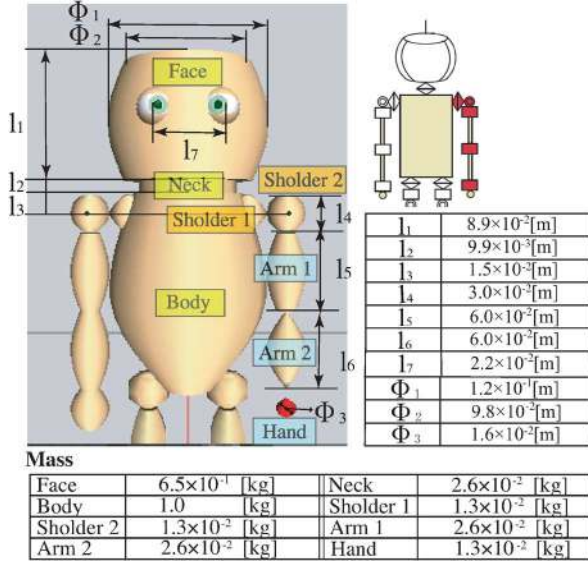
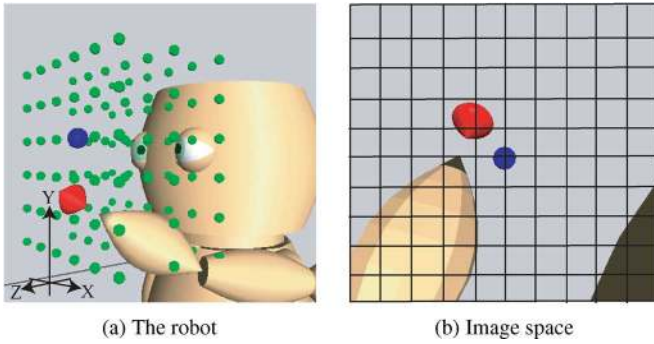


Fig. 4. Robot model and its specifications used in the experiments.


 Fig. 5. A simulation model: 108 green points in (a) are given by the designer as reaching targets during random hand movements and placed at 0.02 [m] intervals in the x, y, and z directions. The blue ball represents the gaze point of the two eyes. (b) Image space is the actual camera image divided into 10×10 units. The winner unit is the one in which the center of the hand is included. (a) The robot and (b) image space.

erstone [25]. The robot model used in this experiment and its specifications are shown in Fig. 4. It has arms with five degrees of freedom. Furthermore, it has a binocular vision system and each eye has two degrees of freedom (pan and tilt). The left hand is colored red so that the robot can easily detect its position in the camera image. Color range is tuned by trial and error so that it cannot be influenced by illumination changes caused by arm movements. There are tactile sensor units on its face. A total of 108 ($6 \times 6 \times 3$) green points in Fig. 5(a) are given by the designer as reaching targets during random hand movements and placed at 0.02 m intervals in the x, y, and z directions. The blue ball in Fig. 5(a) and (b) represent the gaze point of the two eyes.

B. Head-Centered Reference Frame Module

1) *Arm Posture Space*: Five joint angles of the left arm, which are colored red in Fig. 4, constitute the arm posture space. First, the robot selects one of the green points in front of the face randomly. The points are selected randomly, but they approximately cover the space in front of the face. The position of the selected point in the global reference frame (3-D Cartesian reference frame) is denoted as $\mathbf{X}_{\text{green}}$ and the hand position in the global reference frame is denoted as \mathbf{X}_{hand} . The following force \mathbf{F}_{hand} is applied to the center of the hand

$$\mathbf{F}_{\text{hand}} = a_1(\mathbf{X}_{\text{green}} - \mathbf{X}_{\text{hand}}) \quad (1)$$

where, a_1 is a positive constant and set to 75 N/m here. Since the initial position of the hand is near the waist, as shown in Fig. 4, and all joints between hand and shoulder are free (no force is applied), an arm posture for each reaching target is uniquely determined. Therefore, correlations between 5 DOF postures and the corresponding target points might be caused as a result. 5 DOF arm postures are similar to each other if the corresponding target points are near to each other. When the hand reaches $\mathbf{X}_{\text{green}}$, the robot selects another point again.

Here, we focus on an SOM [19] algorithm that is a kind of unsupervised neural network used in order to reduce the dimensions and classify the multidimensional input data. The joint angle data are recorded and used as training data to construct the SOM as an arm posture space.

Each unit in the map contains a representative vector whose dimension is the same with the one of input. The representative vector Θ_i for the i th unit is

$$\Theta_i = (\theta_1^i, \theta_2^i, \dots, \theta_n^i) \quad (2)$$

where n is the number of joint angles (here, $n = 5$). When the current arm posture Θ is given

$$\Theta = (\theta_1, \theta_2, \dots, \theta_n) \quad (3)$$

the mapping for the i th unit on the space is updated depending on the distance from the winner $c_{\text{arm}i}$ th unit

$$\Delta \Theta_i = \kappa(t) \exp(-\|i - c_{\text{arm}i}\|/\gamma) (\Theta - \Theta_i) \quad (4)$$

$$c_{\text{arm}i} = \arg \min_i \|\Theta - \Theta_i\|. \quad (5)$$

$\kappa(t)$ and γ are a learning rate that decays as the learning proceeds and a scaling factor, respectively. For example, in this case, $\kappa(t)$ is computed as following:

$$\kappa(t) = 0.4 \times \exp(1.0 - t/500) \quad (6)$$

and the γ is set to 0.01. The size of the SOM is 10×10 and the learned map is shown in Fig. 6(a). The number of learning

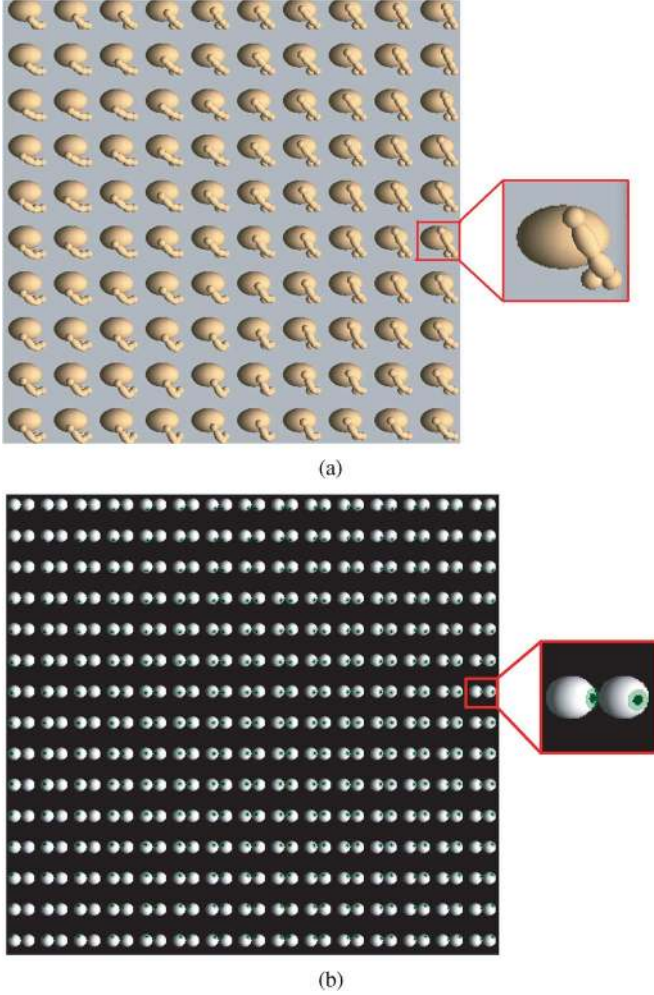


Fig. 6. Acquired maps of arm posture space and ocular angle space. (a) Arm posture space and (b) ocular angle space.

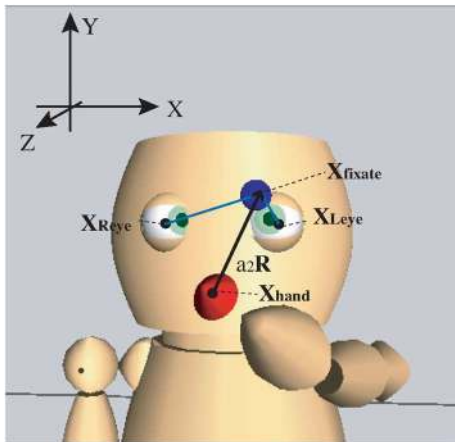


Fig. 7. Variables for the calculation of ocular angles.

steps is 500. The map holds the similar representative vectors in neighboring units.

After learning, in each step, the Euclidean distance between the representative vector of the i th unit and the actual arm posture is calculated. Then, using the winner unit (here the c_{arm} th

unit) with the smallest Euclidean distance, activity α_i^{arm} of the i th unit is given by

$$\alpha_i^{arm} = \exp\left(-\beta\|\Theta_i - \Theta_{c_{arm}}\|^2\right). \quad (7)$$

β is a constant parameter which decides the range of activation.

2) *Ocular Angle Space*: To collect the sets of the ocular angles and the location of the visual stimuli in the camera image, the robot records the ocular angles (pan-tilt angles of each eye) while simultaneously recording the arm joint angles. First, the position \mathbf{X}_{fixate} in the global reference frame is defined as

$$\mathbf{X}_{fixate} = \mathbf{X}_{hand} + a_2\mathbf{R} \quad (8)$$

$$\mathbf{X}_{fixate} = (X_{fixate}x, X_{fixate}y, X_{fixate}z) \quad (9)$$

$$\mathbf{R} = (Rx, Ry, Rz). \quad (10)$$

Rx and Ry are selected among the values from -1 to 1 randomly. Rz equals to 0 and a_2 is 0.05 . We adopt this random noise \mathbf{R} in order to duplicate a behavior of infants who cannot move eyeballs toward an object correctly. We denote the vector of the actual ocular angles Φ and the position of eyes on the face ($\mathbf{X}_{Reye}, \mathbf{X}_{Leye}$) are given by

$$\Phi = (\phi_{right-pan}, \phi_{right-tilt}, \phi_{left-pan}, \phi_{left-tilt}) \quad (11)$$

$$\mathbf{X}_{Reye} = (X_{Reye}x, X_{Reye}y, X_{Reye}z) \quad (12)$$

and

$$\mathbf{X}_{Leye} = (X_{Leye}x, X_{Leye}y, X_{Leye}z). \quad (13)$$

Then, ocular angles are given by

$$\phi_{right-pan} = \arctan\left(\frac{X_{fixate}x - X_{Reye}x}{X_{fixate}z - X_{Reye}z}\right) \quad (14)$$

$$\phi_{right-tilt} = \arcsin\left(\frac{X_{fixate}y - X_{Reye}y}{\|\mathbf{X}_{fixate} - \mathbf{X}_{Reye}\|}\right) \quad (15)$$

$$\phi_{left-pan} = \arctan\left(\frac{X_{fixate}x - X_{Leye}x}{X_{fixate}z - X_{Leye}z}\right) \quad (16)$$

$$\phi_{left-tilt} = \arcsin\left(\frac{X_{fixate}y - X_{Leye}y}{\|\mathbf{X}_{fixate} - \mathbf{X}_{Leye}\|}\right). \quad (17)$$

On the other hand, the robot cannot move its eyeballs to an object voluntarily based on the positions in the camera reference frame, which is the same situation as infants.

Recorded ocular data are used as training data to construct an SOM and the size is 15×15 as shown in Fig. 6(b). The number of learning steps is 1000. After training, the winner unit, whose ID is c_{eye} , is computed in the same manner as for the arm posture space

$$c_{eye} = \arg \min_i \|\Phi - \Phi_i\| \quad (18)$$

where the representative vector is given by

$$\Phi_i = (\phi_{right-pan}^i, \phi_{right-tilt}^i, \phi_{left-pan}^i, \phi_{left-tilt}^i). \quad (19)$$

3) *Image Space*: While recording the ocular data, the robot simultaneously detects its hand position in the camera reference frame. The right(left)-eye image space is the actual camera image divided into 10×10 units as shown in Fig. 5(b). The winner unit whose ID is $c_{rightimage}(c_{leftimage})$ is the one in

which the center of the hand area is included. We adopt the demarcated parts instead of coordinates in the camera image in order to decrease the amount of information to deal with.

4) *Eye Information Space*: In the next step, the eye information space is prepared to combine the activating patterns in the three spaces of the ocular angle space, the right-eye space, and the left-eye image space. An SOM [19] is constructed by utilizing the IDs of the winner units in these three spaces, $\mathbf{C} = (c_{\text{eye}}, c_{\text{rightimage}}, c_{\text{leftimage}})$, as the representative vector in the same way shown in Section III-B1). The size is 20×20 and the number of learning steps is 1000. The winner unit whose ID of this space is denoted as c_{eyeinfo} , and the activity $\alpha_i^{\text{eyeinfo}}$ of the eye information space are defined in the same manner as (7).

5) *Head-Centered Visual Space*: Finally, in the head-centered visual space, the robot learns the association of these combinations of ocular angles and image information to code the same location in the head-centered reference frame by using the proprioception (arm joint angles) as a reference information. The units of the head-centered visual space connect to the units of the arm posture space in a one-to-one correspondence. Then, activity α_i^{space} of the head-centered visual space is

$$\alpha_i^{\text{space}} = \alpha_i^{\text{arm}}. \quad (20)$$

The robot hand is moved toward the green points and its gaze point is around the hand in the same way as learning the ocular angle space and arm posture space in Sections III-B1 and III-B2. Meanwhile, the robot learns the association between head-centered space and the eye information space based on Hebbian learning [26], which is modeled after the synaptic connections in the brain. It is basically an unsupervised training algorithm in which the strength of a connection (weight between units) is increased if both neurons (units) are active at the same time. The original hebbian rule itself has no mechanism for connection weights to get weaker and no upper bound for how strong they can get, and is therefore unstable. Therefore, some modified approaches were suggested. In this model, we use Von Der Malsburg's method [27] that maintains a constant integration of all connection strengths to the same neuron through normalization.

All units of two spaces are connected to each other and the connection weight between the i th unit in the eye information space and the j th unit in the head-centered visual space, w_{ij}^{space} , is updated based on (21)–(23)

$$\bar{w}_{ij}^{\text{space}}(t+1) = \frac{w_{ij}^{\text{space}}(t+1)}{\sum_{j=0}^{N_1} w_{ij}^{\text{space}}(t+1)} \quad (21)$$

where

$$w_{ij}^{\text{space}}(t+1) = w_{ij}^{\text{space}}(t) + \Delta w_{ij}^{\text{space}} \quad (22)$$

and

$$\Delta w_{ij}^{\text{space}} = \epsilon_1 \alpha_i^{\text{space}} \alpha_j^{\text{eyeinfo}}. \quad (23)$$

N_1 and ϵ_1 are the number of units of the head-centered visual space (here, 100) and a constant learning rate (here, 0.2), respectively. It is given by the designer and is smaller to reduce the oscillations while learning.

After learning this association, the robot records the $c_{\text{act-space}}$ th unit that is most strongly connected to the c_{eyeinfo} th unit.

C. VIP Module

In the VIP module, the robot integrates the tactile stimuli of the face and the visual stimuli that are specified in the head-centered reference frame through tactile experience.

1) *Visual Trajectory Space*: This space is prepared for classifying the historical data of approaching visual stimuli whose positions can be computed in the head-centered reference module. First, the robot repeatedly moves its hand toward a random position on the surface of its face from the front. In this case, the gaze point is moved in the same way as before. At that time, the robot computes $c_{\text{act-space}}$ that has the strongest connection to c_{eyeinfo} by using the input data of the ocular angles and the positions in the camera reference frame in every step. Then, the trajectory of the last three steps ($c_{\text{act-space}}(t-2)$, $c_{\text{act-space}}(t-1)$, $c_{\text{act-space}}(t)$) is achieved and used as the representative vector to construct another SOM (visual trajectory space). t is the time when the hand gets within 0.02[m] of the face. The size of the map is 10×10 .

After acquiring SOM, activity α_i^{traj} of the visual trajectory space is calculated when the hand touches the face.

2) *Tactile Space*: The sensor units on the surface of the face correspond to units in tactile space. If the robot perceives tactile stimuli within period t_{const} after t , the ID of the actual activated c_{tac} th unit in the tactile space is recorded as a winner unit. The position of the i th unit on the tactile space is expressed as

$$\mathbf{T}_i = (T_{x_i}, T_{y_i}) \quad (24)$$

where T_{x_i} and T_{y_i} correspond to its coordinates at x-direction and y-direction respectively

$$\alpha_i^{\text{tac}} = \exp\left(-\zeta \|\mathbf{T}_i - \mathbf{T}_{c_{\text{tac}}}\|^2\right) \quad (25)$$

ζ is a constant parameter which decides the range of activation.

3) *Integration (VIP) Space*: In this case, the tactile space units are connected to those in the integration (VIP) space as a one-to-one correspondence. Activity α_i^{vip} in the latter space is given by

$$\alpha_i^{\text{vip}} = \alpha_i^{\text{tac}}. \quad (26)$$

The robot learns the association between the visual trajectory space and the integration (VIP) space based on Hebbian learning. The connection weight between the i th unit in the visual trajectory space and the j th unit in the integration (VIP) space, w_{ij}^{vip} , is updated based on (27)–(29)

$$\bar{w}_{ij}^{\text{vip}}(t+1) = \frac{w_{ij}^{\text{vip}}(t+1)}{\sum_{i=0}^{N_2} w_{ij}^{\text{vip}}(t+1)} \quad (27)$$

where

$$w_{ij}^{\text{vip}}(t+1) = w_{ij}^{\text{vip}}(t) + \Delta w_{ij}^{\text{vip}} \quad (28)$$

and

$$\Delta w_{ij}^{\text{vip}} = \epsilon_2 \alpha_i^{\text{traj}} \alpha_j^{\text{vip}}. \quad (29)$$

N_2 is the number of units of visual trajectory space and set to 100. ϵ_2 is the learning rate and set to 0.5. It is given by the designer.

Finally, by calculating the $c_{\text{act-vip}}$ th unit that is most strongly connected to c_{traj} th unit, the robot can estimate the tactile sensor units that are going to be hit by the hand.

IV. EXPERIMENTAL RESULTS

A. Head-Centered Reference Frame Module

The proposed neural learning architecture described above is applied to the simulation model. First, to evaluate the learning maturation of Hebbian learning in the head-centered visual space, the averaged variance of weights w_{ij}^{space} of the connection between one unit of the eye information space and all units of the head-centered visual space is calculated. In this case, the robot learns the association between two spaces when the hand touches a green point. Initially, one unit of the former space is associated with all units of the latter space equally, therefore the variance is still large. However, during learning, the stronger the connection becomes between one unit of the former space and the appropriate unit of the latter space, the smaller the averaged variance becomes.

The averaged position on the head-centered visual space, $\bar{\mathbf{r}}^i$, which is connected from the i th unit of the eye information space is calculated as

$$\bar{\mathbf{r}}^i = \frac{\sum_{j=1}^{N_1} w_{ij}^{\text{space}} \mathbf{r}_j}{\sum_{j=1}^{N_1} w_{ij}^{\text{space}}} \quad (30)$$

where \mathbf{r}_j denotes the position vector of the j th unit on the head-centered visual space. Furthermore, the variance of connection weights, \hat{r}^i , is calculated as

$$(\hat{r}^i)^2 = \frac{\sum_{j=1}^{N_1} w_{ij}^{\text{space}} \|\mathbf{r}_j - \bar{\mathbf{r}}^i\|^2}{\sum_{j=1}^{N_1} w_{ij}^{\text{space}}}. \quad (31)$$

Then, the connection-weight evaluation is performed with

$$R_1 = \frac{\sum_{i=1}^{N_3} \hat{r}^i}{N_3} \quad (32)$$

where N_3 is the number of units of eye information space and set to 400. The result of 6000 steps during learning is shown in Fig. 8. As learning proceeds, the variance converges and the connections between the units seem potentiated.

We also investigated how the robot adapts itself to situations in which its hand position in the head-centered reference frame is the same although the sets of ocular angles and positions in the camera image are different as shown in Fig. 2. As indicated in Fig. 9(a), the robot places its hand at the fixed point and moves its gazing point for 300 steps as plotted with blue lines. Concretely speaking, in each step, the robot calculates c_{eyeinfo} using the perceived sensation of c_{eye} , $c_{\text{rightimage}}$, $c_{\text{leftimage}}$ that are obtained from the actual gaze angles and camera images. Then, in the head-centered visual space, $c_{\text{act-space}}$ is determined. Moreover, by assigning the representative vector $\Theta_{\text{cact-arm}} = (\theta_1^{\text{cact-arm}}, \dots, \theta_n^{\text{cact-arm}})$

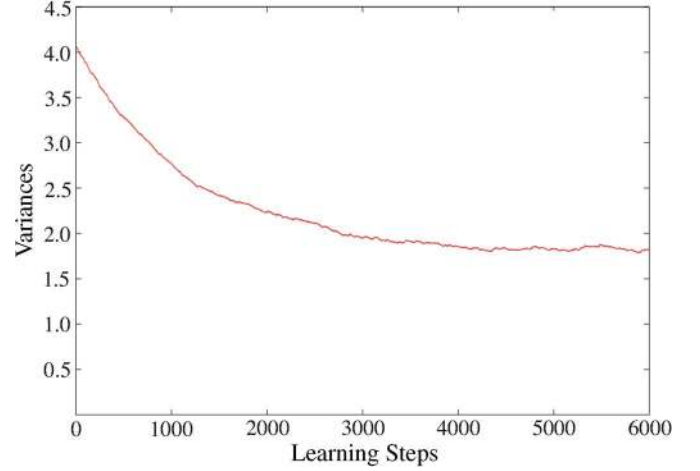


Fig. 8. Variances of the weights during the Hebbian learning of the association between the eye information and the head-centered visual spaces.

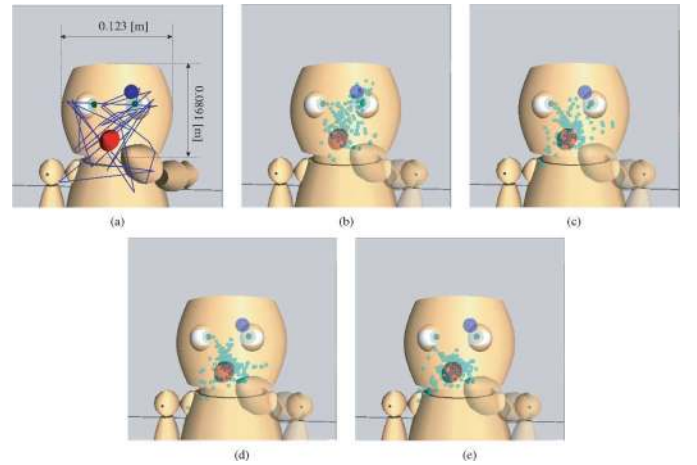


Fig. 9. Estimated hand positions while the robot randomly moves its gaze point around the hand by using the weight values acquired in 1000th, 2000th, 3000th, and 6000th step: blue lines show trajectory of 200 gaze points (blue ball) and light blue points show estimated hand positions. (a) Trajectory of gazing points; (b) the estimated position (1000th step); (c) the estimated position (2000th step); (d) the estimated position (3000th step); (e) the estimated position (6000th steps).

of the $c_{\text{act-arm}}$ th unit in the arm posture space that is interlinked to $c_{\text{act-space}}$ to (33), the position of the hand $\mathbf{X}_{\text{cact-arm}} = (x_{\text{cact-arm}}, y_{\text{cact-arm}}, z_{\text{cact-arm}})$ in the global reference frame (3-D Cartesian reference frame) is calculated as following:

$$\mathbf{X}_{\text{cact-arm}} = f(\Theta_{\text{cact-arm}}) \quad (33)$$

where f is a function that transforms joint angles and link lengths into the hand position in the global reference frame. It is given just to examine the learning results by the designer. In this case, the x , y , and z directions are shown in Fig. 5(a). In order to investigate how the learning proceeds over the time, the robot records the connection weight values at learning steps = 1000, 2000, 3000, and 6000. Then, after learning, we make the robot compute $c_{\text{act-space}}$ and the hand positions $\mathbf{X}_{\text{cact-arm}}$ subsequently based on these four recorded connection weight values. In each step, the moving average of

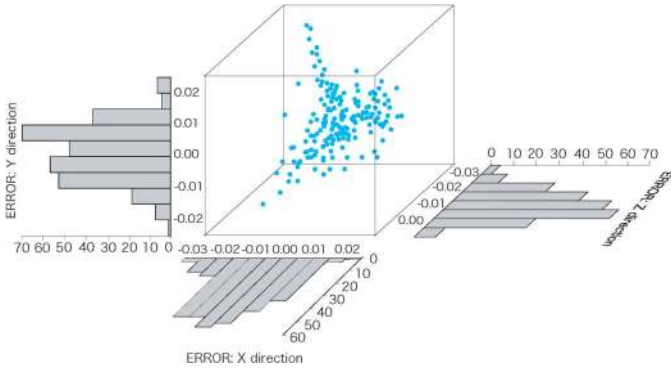


Fig. 10. Histogram of differences between actual and estimated hand positions for Fig. 8(e).

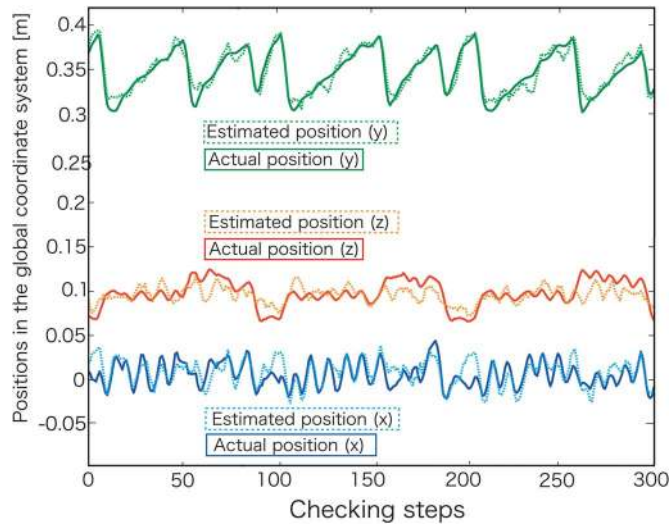


Fig. 11. Difference between the actual and estimated hand positions (while the robot is moving its hand).

$\mathbf{X}_{c_{act-arm}}$ in the last four steps is computed and indicated as the light blue point in Fig. 9(b)–(e) and in Fig. 10. The robot can approximately recall the arm posture that resembles the actual one from the ocular angles and the positions in the camera image that are different from one observation to another. As the learning proceeds, these estimated positions seem to slowly converge to the hand position.

In addition, the histogram of differences (errors) between $\mathbf{X}_{c_{act-arm}}$ and the positions of the actual hand in Fig. 9(e) is shown in Fig. 10. The average values of 300 errors for the three directions are 0.01034 m (x-axis), 0.01057 m (y-axis), and 0.01289 m (z-axis), and the mean error of the z direction is bigger than the others. One reason could be that the number of units in the eye information space is insufficient to cover a large amount of training data.

Finally, $\mathbf{X}_{c_{act-arm}}$ and the actual hand positions while the hand is moved toward the green points in order are shown in Fig. 11, where the errors in z direction are bigger than the others also in this case.

B. VIP Module

To check the Hebbian learning maturation in the integration (VIP) space, the averaged variance of the weights of the con-

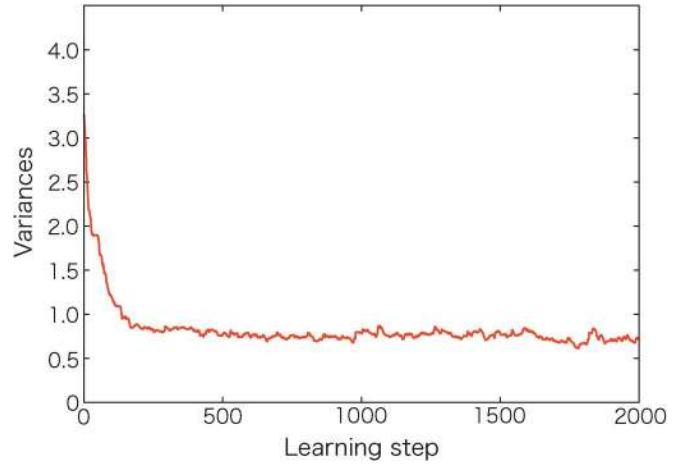


Fig. 12. Variances of the weights during the Hebbian learning of the association between the VIP and the visual trajectory spaces.

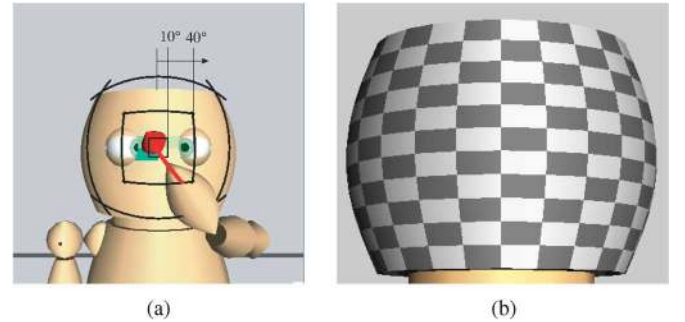


Fig. 13. (a) The robot moves its hand toward the face randomly and estimates the position that is going to be activated by the contact by calculating the unit that has a big weight value between the winner $c_{tra,jth}$ unit in the visual trajectory space. The screen is placed in front of the face and the results are compared with the finding of VIP neurons. (b) Tactile sensor units on the face.

nection between the one unit of the integration (VIP) space and all units of the visual trajectory is computed in the same manner as shown in the last section. The variances of 2000 steps during learning are shown in Fig. 12. As learning proceeds, the connection between the units is evaluated to be potentiated.

Next, we investigated whether the integration space of the VIP module has the same function as VIP neurons themselves, and whether the robot can estimate the tactile units that are going to be activated regardless of the positions of the gaze point. In Fig. 13(a), we placed the screen in front of the robot as seen in the observation of monkeys in Fig. 2 and its center is the midpoint of the two eyes. The tactile sensor units are located as indicated in Fig. 13(b) on the surface of the face. During the hand movement to the face as explained before, we visualize the level of each weight using the green color connected to the $c_{tra,jth}$ unit in the visual trajectory space as shown in Figs. 13(a), 14(III). They are compared with visual and somatosensory receptive fields of VIP neurons. The red arrow indicates the trajectory of the hand. In Fig. 14(I) and (II), some examples of two kinds of receptive fields that VIP neurons have are shown. Moreover, Figs. 14(III) are the activated tactile units at the time when the visual stimuli (the own hand in this case) are shown in each visual receptive field in (I). In Fig. 14(III-f), when the hand is moved toward the bottom of the right eye, an error is

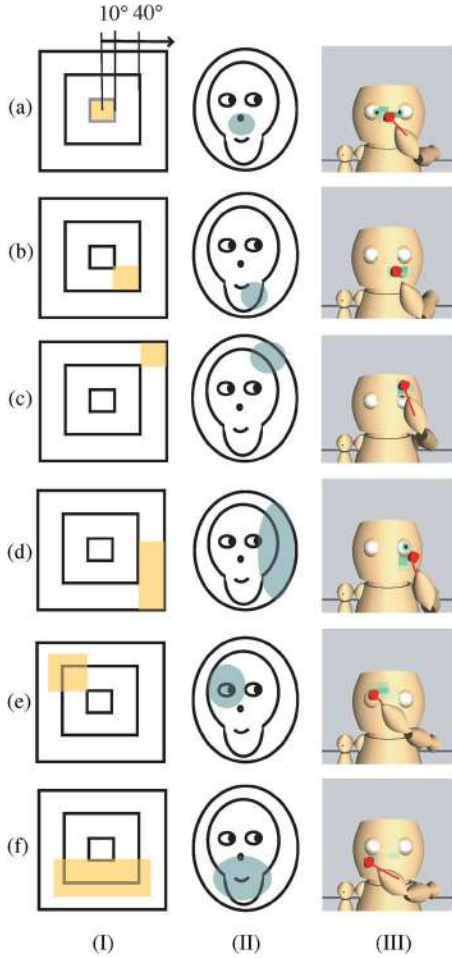


Fig. 14. (I) This square can be thought of as a screen placed in front of the face as seen in Fig. 12. On the square, the patch of orange color corresponds to the visual receptive fields of VIP neurons. (II) The patch of blue color corresponds to the somatosensory receptive field of the same neuron. (III) The activated tactile units at that time when the visual stimuli (the own hand in this case) are indicated in each visual receptive field in (I).

observed. However, the robot can roughly estimate the tactile units that are going to be activated regardless of the position of the gaze point as a result. It seems that the area including the units that connects strongly with the $c_{\text{traj}h}$ unit resembles the corresponding tactile receptive field of the neuron. It could be mentioned that the function of our VIP module is qualitatively similar to actual VIP neurons.

Furthermore, we investigate how the errors in estimating the activating units are being reduced over learning time. The several weight values, $w_{ij}^{\text{vip}}(t = 100, 300, 500, 1000, \text{ and } 2000)$ are recorded in learning phase and used to compute $c_{\text{act-vip}}$ while the robot is moving its hand toward the face randomly for 200 steps. Fig. 15 shows the histogram of the Euclidean distances of the IDs ($c_{\text{act-vip}}$ and c_{tac}) based on each weight value. It appears the accuracy of estimation is enhanced gradually as learning proceeds. Fig. 16(a) shows the final result when utilizing the weight $w_{ij}^{\text{vip}}(2000)$. There are a few errors and they probably happened because the training data of the visual trajectory space, $c_{\text{act-space}}(t-2)$, $c_{\text{act-space}}(t-1)$, and $c_{\text{act-space}}(t)$, were influenced by the errors of the head-centered visual space. Another reason is suggested that the robot some-

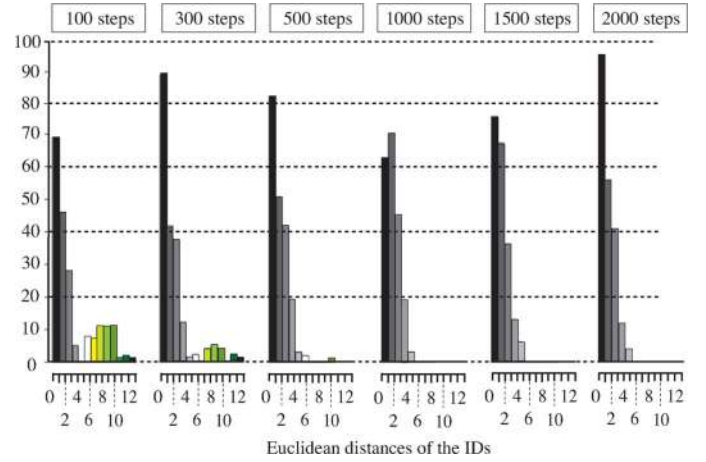


Fig. 15. The activated tactile units are estimated based on the weight values $w_{ij}^{\text{vip}}(100)$, $w_{ij}^{\text{vip}}(300)$, $w_{ij}^{\text{vip}}(500)$, $w_{ij}^{\text{vip}}(1000)$, $w_{ij}^{\text{vip}}(1500)$, $w_{ij}^{\text{vip}}(2000)$. These are histograms of differences between the actual activated and estimated units of tactile space.

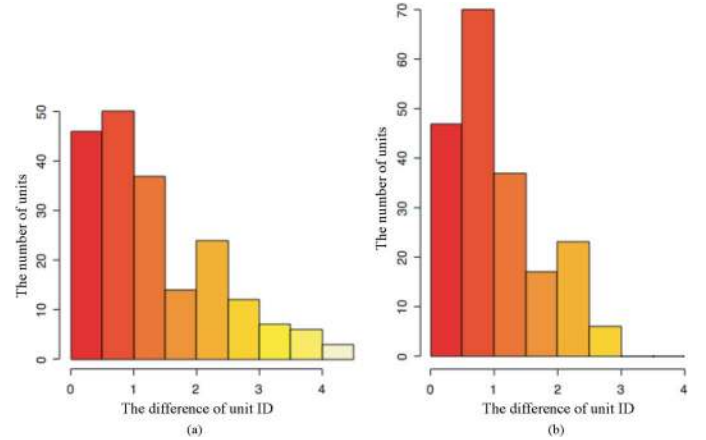


Fig. 16. Histogram of differences between the actual activated and estimated units of tactile space. (a) Indicates the differences which are calculated when the visual information is used. (b) Indicates the differences which are calculated when the proprioceptive information is used.

times loses sight of its hand by moving it outside of the field of view while recording the trajectory. Then, we did an experiment using only the proprioceptive input for VIP module. For the robot, it might be able to ascertain whether it is also possible to predict where the hand will hit without visual information. As stated in the Section III-B5), the units of the head-centered visual space connect to the units of the arm posture space in an one-to-one correspondence as

$$c_{\text{act-space}}(t) = c_{\text{arm}}(t). \quad (34)$$

Therefore, the visual trajectory space is retrained by calculating the historical pattern of the winner c_{arm} units here. In this way, the association between the visual trajectory space and the integration (VIP) space and the active tactile units are also estimated based on the activity determined by actual joint angles originally. In this case, the histogram of distances of IDs of the activated and estimated tactile units is shown in Fig. 16(b). The accuracy of prediction is improved.

V. CONCLUSION AND DISCUSSION

In this paper, we proposed a learning model in which the visuo-spatial and body representations are acquired through “hand regard” behavior that can be observed in the human developmental process. Consequently, the robot acquired a form of perception in which the surrounding space is roughly encoded in a head-centered reference frame. It can also integrate the visual stimuli coded in this reference frame with tactile stimuli on the face, and can acquire the representation whose function is similar to that of VIP neurons by using SOM and Hebbian learning hierarchically.

As a model of acquisition of facial multimodal representation in cognitive developmental robotics, there is a method that Fuke *et al.* [17] proposed so far. In that study, the robot can learn the relative arrangement of tactile sensors on the face. However, because the relationship between these tactile sensors and stimuli in the surrounding space is not considered, it is difficult to utilize the representation for motion generation such as guiding head movements. On the other hand, the representation in this study can be useful to estimate an approaching object. In the future, avoidance behavior might be able to be constructed based on it. In addition, by implementing this method in existing studies, robots can acquire more practical body representations. For example, Hikita *et al.* [28] proposed a method which enables a robot to associate a position of an end effector (hand or tip of tools) in the camera image with the proprioception of its arm. If this “position” can be represented in the head-centered reference frame (or body-centered reference frame in the future) based on our method, it is expected that the robot becomes able to detect the end effectors in a wider area by using the eyeball and neck angles effectively.

There are three issues that we should tackle in the future. First, there are still some errors that can be seen in Figs. 9 and 16. We are going to try to use other algorithm to improve it. There is a possibility that normalizing rules of Hebbian learning negatively affect the maturation. Therefore, for example, we will try to apply the trace rule [29] that is a modified Hebbian rule and changes synaptic weights according to both the current firing rate and the firing rates to recently observed stimuli. This enables neurons to learn to respond similarly to the gradually transforming inputs it receives.

Second, we construct SOMs hierarchically and their sizes are determined by trial and error. If the size is small, the representative vectors become too generalized. Especially in the case of the eye information space, it is not always true that the hand positions in the space are next to each other when the relative vectors of the units of the eye information space are similar. If the size is too big, it is difficult for one unit to encode the visuo-spatial representation as a cluster-like visual and tactile receptive fields of neurons. However, it can be resolved by adjusting the parameter of the activity of units in the spaces such as β . Then, sizes of SOMs have to be examined carefully because there is a possibility that it causes some errors. So far, we have not implemented an algorithm to explore the most appropriate size to cluster the data satisfactorily. However, when we try to scale up this model to larger tactile spaces and other representations in other kinds of reference frames in the future, it might

still be an issue how the sizes are optimized and vast amounts of data should be compressed. We are working on a method in which the robot can determine the size that is adequate enough to cover all input data autonomously and, at the same time, consider other algorithms for compression such as deep belief nets [30].

Finally, in the present study, a robot hand is moved in the small space in front of the face by giving a force only to the hand. Then, the arm postures that provide a hand position are specified. On the other hand, if joint angles are randomly selected, it is difficult to utilize them in order to represent the same position. In case of human infants, before the hand regard behavior starts to be observed, they tend to acquire some kinds of motor primitives based on the physical interaction between body and environment, and the effect of gravity. Thus, it is highly probable that the actual infant hand position correlates with the arm posture during hand regard behavior too. We are going to discuss a relationship between the acquisition of motor representation and visuo-spatial representation in the next stage.

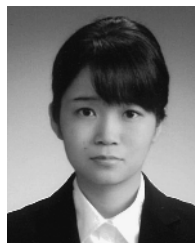
Next, not only to improve the model, but also to understand the human mechanism in more detail, it is also important to approach the problem of how a robot can find the object that it should pay attention to as reference information for the acquisition of the head-centered or other reference frames. For example, in our study, the proprioception of the arm was set to be adopted as reference information by the designer, but if the robot is able to select an object in the surrounding space as reference information autonomously based on an internal attention mechanism, we can also discuss the acquisition of visuo-spatial representation of extrapersonal space. It might be required that the robot can predict the change of the visual information (optical flow) in the image from the ocular motor information and have the memory system at that time.

As the main topic in this paper was first inspired by some findings of humans, finally we try to compare functions of each space in this model to those of some regions that are found in the neurophysiological studies here. First, we pay attention to some findings about lateral intraparietal (LIP) area [31]. Andersen [32] found neurons in the monkey parietal cortex area (LIP area) that combine three kinds of signals: the position of the stimulus on the retina, the positions of the eyes in the orbit, and the neck angles. The LIP area connects to the VIP area [33] and is reported to have both eye-centered and head-centered visual receptive fields [34]. The head movement is not dealt with in our study, but it can be assumed that the eye information space corresponds to the LIP area as shown in Fig. 14. As mentioned in the introduction, VIP area is known as the region that has the “peri-personal” visuo-spatial representation. The “peri-personal space” is defined as the space within reach of the arm in the neurophysiological studies. Actually, it was revealed that the visuospace is represented in different regions in the brain, peri-personal space [24], and extrapersonal space that is out of reach of the arm [35] based on the findings of spatial neglect syndromes. This peri-personal space is extended when the subject uses a tool [36]. Rizzolatti *et al.* [37] also reported that connection of this VIP area and the F4 area (the area of arm representation) in the brain is important for that representation. Thus, we also suppose that arm posture space corresponds to the F4

area (the area of arm representation) and their claim might support our hypothesis in which the arm proprioceptive information contribute to the construction of visuo-spatial representation.

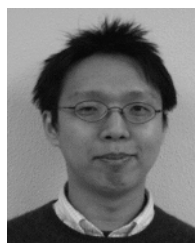
REFERENCES

- [1] H. Head and G. Holmes, "Sensory disturbances from cerebral lesions," *Brain*, vol. 34, pp. 102–254, 1911/1912.
- [2] S. I. Maxim, *Body Image and Body Schema*, P. D. Helena, Ed. Amsterdam, The Netherlands: John Benjamins, 2005.
- [3] V. S. Ramachandran and S. Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human Mind*. New York: William Mollow, 1998, vol. 2.
- [4] A. Iriki, M. Tanaka, S. Obayashi, and Y. Iwamura, "Self-images in the video monitor coded by monkey intraparietal neurons," *Neurosci. Res.*, vol. 40, pp. 163–173, 2001.
- [5] T. Jellema, G. Maassen, and D. I. Perrett, "Single cell integration and aminate form, motion, and location in the superior temporal cortex of the macaque monkey," *Cerebr. Cortex*, vol. 14, pp. 781–790, 2004.
- [6] J. R. Duhamel, C. L. Colby, and M. E. Goldberg, "Ventral intraparietal area of the macaque: Congruent visual and somatic response properties," *J. Neurophysiol.*, vol. 79, pp. 126–136, 1998.
- [7] M. S. A. Graziano and D. F. Cooke, "Parieto-frontal interactions, personal space, and defensive behavior," *J. Neuropsychol.*, vol. 44, pp. 845–859, 2006.
- [8] M. I. Sereno and R. Huang, "A human parietal face area contains aligned head-centered visual and tactile maps," *Nature Neurosci.*, vol. 9, pp. 1337–1343, 2006.
- [9] G. M. Stratton, "Vision without inversion of the retinal image," *Psychol. Rev.*, vol. 4, pp. 463–481, 1897.
- [10] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 12–34, 2009.
- [11] Y. Yoshikawa, "Subjective robot imitation by finding invariance," Ph.D. dissertation, Osaka University, Osaka, Japan, 2005.
- [12] C. Nabeshima, M. Lungarella, and Y. Kuniyoshi, "Body schema adaptation for robotic tool use," *Adv. Robot.*, vol. 20, pp. 1105–1126, 2006.
- [13] A. Stoytchev, "Toward video-guided robot behaviors," in *Proc. 7th Int. Conf. Epigenetic Robot.*, 2007, pp. 165–172.
- [14] A. Pitti, H. Mori, S. Kouzuma, and Y. Kuniyoshi, "Contingency perception and agency measure in visuo-motor spiking neural networks," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 86–97, 2009.
- [15] L. Natale, F. Orabona, G. Metta, and G. Sandini, "Sensorimotor coordination in a "baby" robot: Learning about objects through grasping," *Robot. Auton. Syst.*, vol. 37, pp. 185–193, 2001.
- [16] M. Hersch, E. Sauser, and A. Billard, "Online learning of the body schema," *Int. J. Human. Robot.*, vol. 5, no. 2, 2008.
- [17] S. Fuke, M. Ogino, and M. Asada, "Body image constructed from motor and tactile images with visual information," *Int. J. Human. Robot.*, vol. 4, no. 3, pp. 347–364, 2007.
- [18] A. Pouget, S. Deneve, and J. R. Duhamel, "A computational perspective on the neural basis of multisensory spatial representations," *Nature Rev. Neurosci.*, vol. 3, pp. 741–747, 2002.
- [19] T. Kohonen, *Self-Organizing Maps*. Berlin, Heidelberg, Germany: Springer-Verlag, 1995.
- [20] T. N. Affalo and M. S. A. Graziano, "Possible origins of the complex topographic organization of motor cortex: Reduction of a multidimensional space onto a two-dimensional array," *J. Neurosci.*, vol. 26, no. 23, pp. 6288–6297, 2006.
- [21] R. O. Gilmore and M. H. Johnson, "Body-centered representations for visually-guided action emerge during early infancy," *Cognition*, vol. 65, no. 1, pp. 1–9, 1997.
- [22] J. Piaget, *Naissance de l'intelligence chez l'enfant baoule de Cote d'Ivoire*. Paris, France: Delachaux et Niestle, 1936.
- [23] G. Rizzolatti, C. Sinigaglia, and F. Anderson, *Mirrors in the Brain—How Our Minds Share Actions and Emotions*. Oxford, U.K.: Oxford University Press, 2007.
- [24] P. W. Halligan and J. C. Marshall, "Left neglect for near but not for far space in man," *Nature*, vol. 352, pp. 673–674, 1991.
- [25] R. Featherstone, "The calculation of robot dynamics using articulated-body inertias," *Int. J. Robot. Res.*, vol. 2, no. 1, pp. 13–30, 1983.
- [26] D. O. Hebb, *The Organization of Behavior*. New York: Wiley, 1949.
- [27] Von Der Malsburg, "Self-organization of orientation sensitive cells in the striate cortex," *Kybernetik*, vol. 14, pp. 85–100, 1973.
- [28] M. Hikita, S. Fuke, M. Ogino, T. Minato, and M. Asada, "Visual attention by saliency leads cross-modal body representation," in *Proc. 7th Int. Conf. Develop. Learning*, 2008, pp. 157–162.
- [29] P. Foldiak, "Learning invariance from transformation sequences," *Neural Comput.*, vol. 3, no. 2, pp. 194–200, 1991.
- [30] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [31] D. J. Freedman and J. A. Assad, "Experience-dependent representation of visual categories in parietal cortex," *Nature*, vol. 443, pp. 85–88, 1988.
- [32] R. A. Andersen, "Encoding of intention and spatial location in the posterior parietal cortex," *Cerebr. Cortex*, vol. 5, pp. 457–469, 1995.
- [33] G. J. Bratt, R. A. Andersen, and J. R. Stoner, "Visual receptive field organization and cortico-cortical connections of the lateral intraparietal area (area lip) in the macaque," *J. Comparative Neurol.*, vol. 299, pp. 421–445, 1990.
- [34] O. A. Mullette-Gillman, Y. E. Cohen, and J. M. Groh, "Eye-centered, head-centered, and complex coding of visual and auditory targets in the intraparietal sulcus," *J. Neurophysiol.*, vol. 94, pp. 2331–2352, 2005.
- [35] A. Cowey, M. Small, and S. Ellis, "Left visuo-spatial neglect can be worse in far than in near space," *Neuropsychol.*, vol. 32, pp. 1069–1066, 1994.
- [36] A. Berti and F. Frassinetti, "When far becomes near: Remapping of space by tool use," *J. Cogn. Neurosci.*, vol. 12, pp. 415–420, 2000.
- [37] G. Rizzolatti and M. Matelli, "Two different streams form the dorsal visual system: Anatomy and functions," *Exper. Brain Res.*, vol. 153, pp. 146–157, 2003.



Sawa Fuke received the B.E. degree from the University of Electro-Communications, in Tokyo, Japan, in 2005 and M.E. degree from Osaka University in Osaka, Japan in 2007. She is currently working towards the Ph.D. degree in the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University and a JSPS Research Fellow.

Her research interests are body-image, spatial perception using visual information, and sensor fusion.



Masaki Ogino received the B.S., M.S., and Ph.D. degrees from Osaka University, Osaka, Japan, in 1996, 1998, and 2005, respectively.

From 2002 to 2006, he was a Research Associate in the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University. From 2006 to 2009, He was a researcher in "ASADA Synergistic Intelligence Project" of ERATO (Exploratory Research for Advanced Technology by Japan Science and Technology Agency). He is currently an Assistant Professor at the Department of

Adaptive Machine Systems, Graduate school of Engineering, Osaka University. His research interests are humanoid robot control, biped walking, and cognitive issues involving humanoid robots.



Minoru Asada (F'05) received the B.E., M.E., and Ph.D. degrees in control engineering from Osaka University, Osaka, Japan, in 1977, 1979, and 1982, respectively.

From 1982 to 1988, he was a Research Associate of Control Engineering, Osaka University, Toyonaka, Osaka, Japan. In April 1989, he became an Associate Professor of Mechanical Engineering for Computer-Controlled Machinery, Osaka University, Suita, Osaka, Japan. In April 1995, he became a Professor of the same department. Since April 1997, he has been a Professor of the Department of Adaptive Machine Systems at the same university. From August 1986 to October 1987, he was a Visiting Researcher at the Center for Automation Research, University of Maryland, College Park. Since 2002, he has been the President of the International RoboCup Federation.

Dr. Asada was elected a Fellow of the IEEE for contributions to robot learning and applications in 2005. In the same year, he was also elected to the position of Research Director of the "ASADA Synergistic Intelligence Project" of ERATO (Exploratory Research for Advanced Technology by Japan Science and Technology Agency).