VIP neuron model: Head-centered cross-modal representation of the peri-personal space around the face

Sawa Fuke*, Masaki Ogino**, Minoru Asada* **
Graduate School of Engineering, Osaka University, JST ERATO Asada Synergistic Intelligence Project***
2-1,Yamadaoka,Suita,Osaka 565-0871,Japan
sawa.fuke@ams.eng.osaka-u.ac.jp,[ogino,asada]@jeap.org

***Abstract*—Since body representation is one of the most fundamental issues for physical agents (humans, primates, and also robots) to adaptively perform various kinds of tasks, a number of learning methods have attempted to make robots acquire their body representation. However, these previous methods have supposed that the reference frame is given and fixed a priori. Therefore, such acquisition has not been dealt.**

This paper presents a model that enables a robot to acquire cross-modal representation of its face based on VIP neurons whose function (found in neuroscience) is not only to code the location of visual stimuli in the head-centered reference frame and but also to connect visual and tactile sensations. Preliminary simulation results are shown and future issues are discussed.

***Index Terms*—body representation, integration of the sensor information, multi-modal sensors**

I. Introduction

Humans can perform various kinds of tasks through interaction with objects, usually unconsciously, but sometimes with consciousness of their own body representation in their brains, based on which humans are supposed to decide which action to take. Such representation has been called the ”body schema” an unconscious neural map in which multi-modal sensory data are unified [**?**] or ”body image” an explicit mental representation of the body and its functions [**?**]. Among studies related to the body representation, the results of Ramachandran [**?**] and Iriki et al. [**?**] suggest that representations in biological systems are flexible and acquired by spatio-temporal integration of different modal sensory data. In neuroscience, VIP neurons which are found in the parietal lobe are activated for both visual stimuli coded in a head-centered reference frame and the actual tactile stimuli of the body (face) [**?**][**?**][**?**]. However, the acquisition of such representation remains unclear.

Unlike conventional methods in robotics where the fixed body representation is given by the designer, in cognitive developmental robotics [**?**], a number of learning methods have attempted to make robots acquire their body representation by aiming not only to understand the process of acquiring body representation in humans but also to apply these models to robots that can develop such representation. Yoshikawa et al. [**?**] proposed a model in which the robot can detect its own body in a camera image based on the invariance in multiple sensory data. In Nabeshima et al.'s model [**?**], a robot learns the properties of its controller following the synchronization of the activation of visual and somatic sensations while a robot is using a tool. Natare et al. [**?**] offered the means for a robot not only to predict the visual information of a hand from arm postures but also to estimate the Jacobian for a reaching task. Furthermore, Stoytchev [**?**] proposed a model that enables a robot to detect its own body on a TV monitor based on the synchronization of the activation of vision and proprioception. Hersch et al. [**?**] proposed an algorithm through which a robot learns joint positions and orientations based on the information of the observing hand's positions represented in both the head-centered and the hand-centered reference frames. Additionally, in most studies, the representation of invisible body parts such as a face or a back cannot be acquired due to a lack of information. Fuke et al. [**?**] proposed a model that acquired the body representation of its invisible face by estimating its hand position from the change of the proprioceptive sensation while touching it.

However, the above studies assumed that camera positions are fixed or that the coordinate system to project the positions to a reference frame (body-centered reference frame) in visual space is given by the designer. They have not discussed how the reference frame is acquired with the hierarchical use of raw visual and proprioceptive sensory data. Such visuospatial representation should be considered together with the body representation since it is acquired by the spatio-temporal integration of different modal sensory data with neck and eyes in the developmental process. An experiment with upside-down glasses is one evidence that agents reacquire the reference frame for body parts localization and body representation interdependently since the normal relationship between motor information (eyeball angles) and visual information is broken. According to Stratton [**?**], the subject became able to interpret the visual space and to perceive the integration of sensations about one week later. This implies that visuospatial representation is flexible and is acquired from experiences similar to learning body representation.

Here, we propose a learning model in which a robot acquires not only the head-centered reference frame but also the cross-modal representation of the face based on raw sensory data during self body observation. The acquired

cross-modal representation corresponds to the properties of VIP neurons found in neuroscience. The rest of this paper is organized as follows. First, we introduce neurophysiological findings that provide a valuable clue about the acquisition of face representation. Next, the system and the learning algorithm details are described. Then the simulation results are shown, and a discussion and a conclusion are given.

II. What is a VIP neuron?

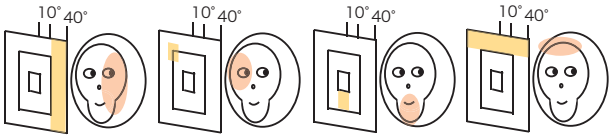


Fig. 1. Visual and somatosensory receptive fields of neurons in VIP. The same VIP neuron is activated when something is shown on the screen's shaded area in front of the monkey and when the face's shaded area is stimulated (modified from Fig. 1 in [**?**]).

The neurons in the adjacent ventral intraparietal area (hereafter called VIP neurons) are known to have bimodal properties, activated when a somatosensory receptive field is stimulated and when a visual stimulus approaches the face regardless where the monkey is looking. Fig. 1 (modified from Fig. 1 in [**?**]) shows examples of the visual and somatosensory receptive fields that the same VIP neurons have. Therefore, the VIP area is supposed to be where the spatial representation with respect to the head-centered reference frame is integrated with tactile sensation. Here, we propose a learning model that enables a robot to acquire the head-centered reference frame and then to integrate the tactile representation in the face with the acquired reference frame as observed in VIP neurons.

III. VIP neuron model

An overview of the proposed model is shown in Fig. 2, where two modules are involved. First, the robot acquires the head-centered reference frame module. It has many sets of eyeball angles and the retinotopic image (camera image) which are represented in the eye information space in Fig. 2. Fig. 3 indicates the two sets of eye angles and position in the camera image for the red object as an example. Before learning, the robot could not figure out that this red object was located at the same position in the head-centered reference frame even though the two sets appear different. To construct a head-centered reference frame, the robot needs an object that can be assumed to be static in the surrounding space as reference information. In our study, the robot associates the eyeball angles and camera image by regarding "the proprioceptive sensation of its own body 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 each module are given next.

To validate the model, computer simulations were conducted with a dynamics simulator that has arms with five degrees of freedom and a binocular vision system as shown in Fig. 4(a). Each eye has two degrees of freedom (pan and tilt directions). 108 ($6 \times 6 \times 3$) green points in Fig. 4(a) are given by the designer as reaching targets and placed at 0.02[m] intervals in the x, y, and z directions. In addition, the blue ball represents the gaze point of the two eyes.

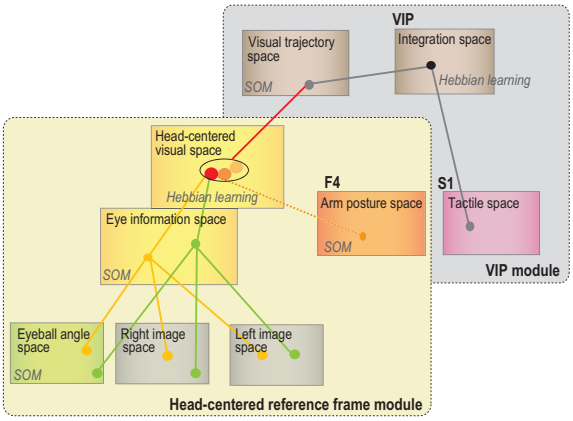


Fig. 2. Overview of the proposed model

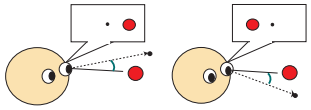
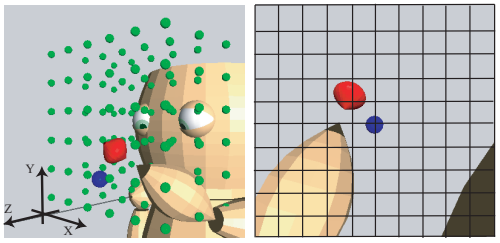


Fig. 3. Two sets of eyeball angles and positions in the camera image for red objects placed at same position in head-centered reference frame

A. Head-centered reference frame module



(a) The robot (b) Image of right eye

Fig. 4. Simulation model

1) Arm posture space: The robot randomly moves its hand toward one of the green points. The five joint angles of the left arm are recorded and used to construct a Self Organizing Map (SOM) [**?**] as training data. The SOM's size is 10×10 and

the learned map is shown in Fig. 5. The number of learning steps is 500.

After learning, in each step, the Euclidean distance between the representative vector of the i-th unit, $\mathbf{\Theta_i} = (\theta_1^i, ..., \theta_n^i)$, and the actual arm joint angles, $\mathbf{\Theta} = (\theta_1, ..., \theta_n)$ $(n = 5)$, is calculated. Then, using the winner unit (here the c_{arm}-th unit) with the smallest Euclidean distance, activity α_i^{arm} of the arm posture space is computed as described below:

$$\alpha_i^{arm} = e^{-\beta(d_i^{arm})^2}, \tag{1}$$

$$d_i^{arm} = |\mathbf{\Theta}_i - \mathbf{\Theta}_{c_{arm}}|, \tag{2}$$

$$c_{arm} = \arg\min_i |\mathbf{\Theta} - \mathbf{\Theta}_i|. \tag{3}$$

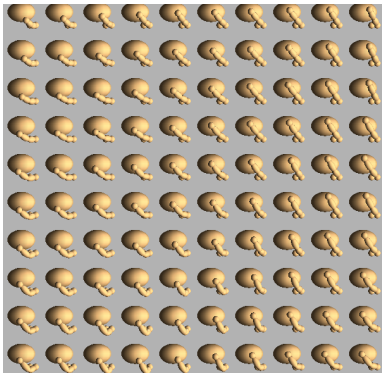


Fig. 5. Arm posture space

2) Eyeball angle space: To collect the sets of the eyeball angles and the location of the visual stimuli in the camera image, the robot records the eyeball angles (pan-tilt angles of each eye) while moving its gaze point around the hand and simultaneously recording the arm joint angles. The data are used to construct SOM as training data and the size is 15×15 as shown in Fig. 6. The number of learning steps is 1000. In the same manner as the arm posture space, the winner unit whose ID is c_{eye} is computed as follows:

$$c_{eye} = \arg\min_i |\mathbf{\Phi} - \mathbf{\Phi}_i|, \tag{4}$$

where the representative vector is

$$\mathbf{\Phi_i} = (\phi^i_{right-pan}, \phi^i_{right-tilt}, \phi^i_{left-pan}, \phi^i_{left-tilt}), \tag{5}$$

and the vector of the actual eyeball angles is

$$\mathbf{\Phi} = (\phi_{right-pan}, \phi_{right-tilt}, \phi_{left-pan}, \phi_{left-tilt}). \tag{6}$$

3) Image space: The robot simultaneously detects its hand position in the camera reference frame. In our experiment, the hand itself is colored red so that the robot can easily detect its position. The right (left) image space is divided into 10×10 units as shown in Fig. 4 (b). The winner unit whose ID is $c_{rightimage}(c_{leftimage})$ is the one in which the center of the hand area is included.

Fig. 6. Eyeball angle space

4) Eye information space: In the next step, eye information space is prepared to combine the activating patterns in the three spaces of the eyeball angle and the right and the left image spaces. SOM is constructed by utilizing the IDs of the winner units in these spaces, $C = (c_{eye}, c_{rightimage}, c_{leftimage})$, as the representative vector. The size is 20×20 and the number of learning steps is 1000. The winner unit whose ID is $c_{eyeinfo}$ and the activity $\alpha_i^{eyeinfo}$ of the eye information space are defined in the same manner as Eqs. (1)-(3).

5) Head-centered visual space: Finally, in the head-centered visual space, the robot learns the association of these combinations to code the same location in the head-centered reference frame based on Hebbian learning. This association is triggered by the same proprioceptive sensation. The units of the head-centered visual space connect to the units of the arm posture space in an one-to-one correspondence. Then activity α_i^{space} of the head-centered visual space is

$$\alpha_i^{space} = \alpha_i^{arm}. \tag{7}$$

In the same way, the robot moves its hand toward the green points and its gaze point around the hand while learning. The connection weight between the i-th unit in the head-centered visual space and the j-th unit in the eye information space, w_{ij}^{space}, is updated based on Eqs. (8)-(10):

$$\bar{w}_{ij}^{space}(t+1) = \frac{w_{ij}^{space}(t+1)}{\sum_{i=0}^{N} w_{ij}^{space}(t+1)}, \tag{8}$$

where

$$w_{ij}^{space}(t+1) = w_{ij}^{space}(t) + \Delta w_{ij}^{space}, \tag{9}$$

$$\Delta w_{ij}^{space} = \epsilon \alpha_i^{space} \alpha_j^{eyeinfo}. \tag{10}$$

N is 100, the number of units of the head-centered visual space. After learning this association, the robot records the $c_{act-space}$-th unit that is most strongly connected to the $c_{eyeinfo}$-th unit.

B. VIP module

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

1) Visual trajectory space: First, the robot repeatedly moves its hand toward the random positions on the surface of its face from the front. In this case, the gaze point is moved the same as before. At that time, the robot computes $c_{act-space}$ by using the input data of the eyeball angles and the positions in the camera reference frame. Then the trajectory of three steps $(c_{act-space}(t-2), c_{act-space}(t-1), c_{act-space}(t))$ is achieved and used as the representative vector to construct SOM (visual trajectory space). t is the time when the hand gets within 0.02[m] of the face. The size is 10×10. After acquiring SOM, activity α_i^{traj} of the visual trajectory space is calculated.

2) Integration (VIP) space: There are 12×12 tactile sensor units on the face's surface in the simulator. These sensor units correspond to units in tactile space. If the robot perceives tactile stimuli within period t_{const} after t, the ID of the activated c_{tac}-th unit in the tactile space is recorded. Additionally, the activity of the tactile space is calculated based on c_{tac}:

$$\alpha_i^{tac} = e^{-\zeta (d_i^{tac})^2}, \tag{11}$$

$$d_i^{tac} = |i - c_{tac}|. \tag{12}$$

Also in this case, the tactile space units are connected to those in the integration (VIP) space an one-to-one correspondence. Activity α_i^{vip} in the latter space is

$$\alpha_i^{vip} = \alpha_i^{tac}. \tag{13}$$

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 Eqs. (14)-(16):

$$\bar{w_{ij}}^{vip}(t+1) = \frac{w_{ij}^{vip}(t+1)}{\sum_{i=0}^{N_2} w_{ij}^{vip}(t+1)}, \tag{14}$$

where

$$w_{ij}^{vip}(t+1) = w_{ij}^{vip}(t) + \Delta w_{ij}^{vip}, \tag{15}$$

$$\Delta w_{ij}^{vip} = \epsilon \alpha_i^{traj} \alpha_j^{vip}. \tag{16}$$

N_2 is 100, the number of the units of visual trajectory space. Finally, by calculating the $c_{act-vip}$-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

Our proposed method 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 adopted. 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 is.

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

$$\bar{\boldsymbol{r}}^i = \frac{\sum_{j=1}^{N_3} w_{ij}^{space} \boldsymbol{r}_j}{\sum_{j=1}^{N_3} w_{ij}^{space}}, \tag{17}$$

where $\boldsymbol{r}_j$ 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

$$(\hat{r}^i)^2 = \frac{\sum_{j=1}^{N_3} w_{ij}^{space} \|\boldsymbol{r}_j - \bar{\boldsymbol{r}}^i\|^2}{\sum_{j=1}^{N_3} w_{ij}^{space}}, \tag{18}$$

where N_3 is 400, the number of the units of eye information space. The result of 6000 steps during learning is shown in Fig. 7. As learning proceeds, variance converges and the connection between the units seems potentiated.

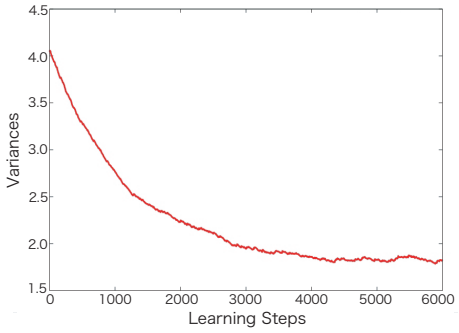


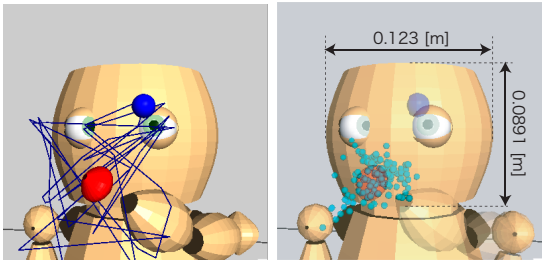
Fig. 7. Variances of the weights during the Hebbian learning of the association between the eye information and the head-centered visual spaces

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 eyeball angles and positions in the camera image are different. As indicated in Fig. 8(a), the robot places its hand at the fixed point and moves its gazing point for 300 steps as plotted with blue lines. In each step, the robot calculates $c_{eyeinfo}$ using the perceived sensation of $c_{eye}, c_{rightimage}, c_{leftimage}$ and determines $c_{act-space}$ in the

head-centered visual space. Moreover, by assigning the representative vector of $\boldsymbol{\Theta}_{\boldsymbol{c_{act-arm}}} = (\theta_1^{c_{act-arm}}, ..., \theta_n^{c_{act-arm}})$ of the $c_{act-arm}$-th unit in the arm posture space that is interlinked to $c_{act-space}$ to Eq. (19), the position of the hand $\boldsymbol{X}_{\boldsymbol{c_{act-arm}}} = (x_{c_{act-arm}}, y_{c_{act-arm}}, z_{c_{act-arm}})$ in the global reference frame is calculated. The x, y, and z directions are shown in Fig. 4(a).

$$\boldsymbol{X}_{\boldsymbol{c_{act-arm}}} = f(\boldsymbol{\Theta}_{\boldsymbol{c_{act-arm}}}), \tag{19}$$

where f is a transform function that is given to examine the learning results. In each step, the moving average of $\boldsymbol{X}_{\boldsymbol{c_{act-arm}}}$ in the last four steps is computed and indicated as the light blue point in Fig. 8(b) and Fig. 9. The robot can approximately recall the arm posture that resembles the actual one regardless of the eyeball angles and the positions in the camera image. In addition, the histogram of difference (error) between $\boldsymbol{X}_{\boldsymbol{c_{act-arm}}}$ and the positions of the actual hand is shown in Fig. 9. The average values of 300 errors for the three directions are $0.01034[m]$(x axis), $0.01057[m]$(y), and $0.01289[m]$(z) and the mean error of the perspective direction is bigger than the others. One reason may be that the number of the units in the eye information space is insufficient to cover a large amount of training data.



(a) Movement of gazing point (b) Estimated position

Fig. 8. Estimated hand position while the robot randomly moves its gaze point around the hand: blue lines show trajectory of 300 gaze points (blue ball) and light blue points shows estimated hand positions.

$\boldsymbol{X}_{\boldsymbol{c_{act-arm}}}$ and the actual hand positions while the robot moves its hand toward the green points in the order are shown in Fig. 10, where the errors in z direction are bigger than the others.

B. VIP module

To check the Hebbian learning maturation in the integration (VIP) space, the averaged variance of the weights of the connection between the one unit of the integration (VIP) space and all units of the visual trajectory is also computed in the same manner shown in the last section. The variances of 2000 steps during learning are shown in Fig. 11. As learning proceeds, the connection between the units is estimated to be potentiated.

Additionally, we visualize the level of each weight by color connected to the c_{traj}-th unit in the visual trajectory space

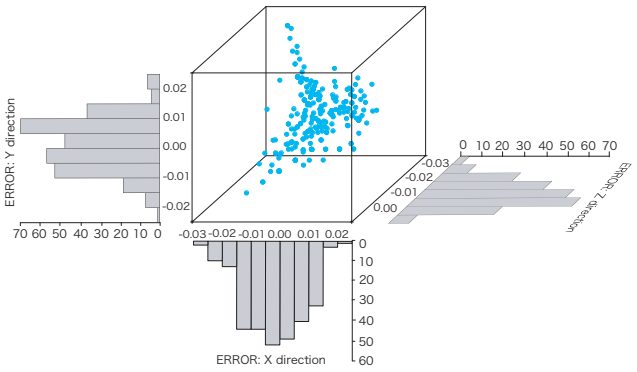


Fig. 9. Histogram of differences between actual and estimated hand positions for Fig. 8

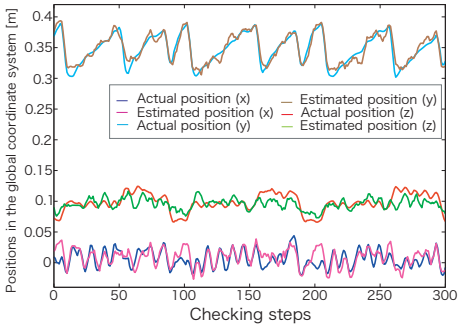


Fig. 10. Difference between the actual and estimated hand positions (while the robot is moving its hand)

while the robot is moving its hand toward the face from the front. As a result, the robot can roughly estimate the tactile units that are going to be activated regardless of the position of the gaze point.

On the other hand, the histogram of the Euclidean distances of $c_{act-vip}$ and c_{tac} for 200 steps after learning is shown in Fig. 13. These errors probably happened happen because the training data of the visual trajectory space, $c_{act-space}(t -$

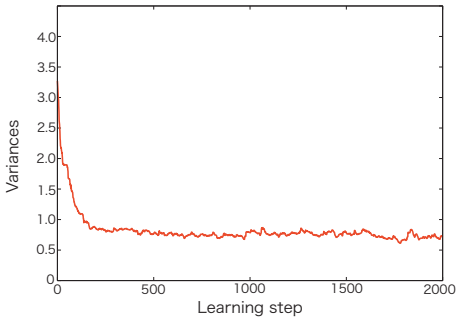
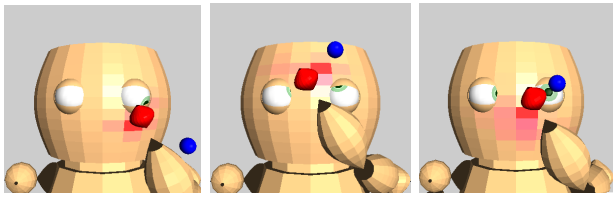


Fig. 11. Variances of weights during Hebbian learning of the association between visual trajectory and integration spaces



(a) Result1 (b) Result2 (c) Result3

Fig. 12. Values of weights connected to activated unit of visual trajectory space

2), $c_{act-space}(t-1)$, and $c_{act-space}(t)$, were influenced by the errors of the head-centered visual space. Another reason is suggested that the robot sometimes loses sight of its hand by moving it outside of the field of view while recording the trajectory.

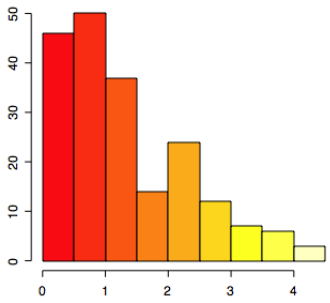


Fig. 13. Histogram of differences between the actual activated and estimated units of tactile space

V. Conclusion and Discussion

The robot acquired the visual-spatial perception by which the surrounding space is roughly encoded in a head-centered reference frame. In addition, it can integrate the visual stimuli coded in this reference frame and tactile stimuli on the face, and can acquire the representation whose function is similar to the one of VIP neurons by using SOM and Hebbian learning hierarchically. In brain science, it has been reported that the VIP-F4 (the area of arm representation) circuit in the brain is important for spatial perception [**?**] around the body. Neurophysical findings exist in which space is differentially represented depending on whether the area is in reach of the hand (peripersonal space) or out of reach of the hand (extrapersonal space) [**?**]. These findings support the possibility that the arm's proprioceptive sensation contributes to the acquisition of the head-centered reference frame.

However, since the estimated positions in the head-centered reference frame are not accurate enough, we are going to continue to improve the model. For example, not only by using the proprioceptive sensation, the visual targets might be useful as a reference point for the acquisition of the head-centered reference frame if the robot can expect the change of the visual information (optical flow) in the image from the eyeball motor information. Moreover, by using ax similar hierarchical construction,in the future we are planning to make the robot learn the spatial representation in a body-centered or global reference frame.

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 (edited by P. D. Helena)" John Benjamins Publishing Company, 2005.

[3] V.S. Ramachandran, and S. Blakeslee, "Phantoms in the Brain: Probing the Mysteries of the Human Mind" William Mollow, 1998.

[4] A. Iriki, M. Tanaka, S. Obayashi and Y.Iwamura, "Self-images in the video monitor coded by monkey intraparietal neurons" Neuroscience Research, Vol. 40, 163-173, 2001.

[5] J. R. Duhamel, C. L. Colby, and M. E. Goldberg, "Ventral intraparietal area of the macaque: Congruent visual and somatic response properties" Journal of Neurophysiology, Vol. 79, pp. 126-136, 1998.

[6] M. S. A. Graziano and D. F. Cooke, "Parieto-frontal interactions, personal space, and defensive behavior" Neuropsychologia, Vol. 44, pp. 845-859, 2006.

[7] M. I. sereno and R. Huang "A human parietal face area contains aligne d head-centered visual and tactile maps" Nature Neuroscience, Vol. 9, pp.1337-1343, 2006.

[8] 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 System, Vol. 37, pp. 185-193, 2001.

[9] Y. Yoshikawa, "Subjective robot imitation by finding invariance" Ph. D thesis, Osaka University, 2005.

[10] C. Nabeshima, M. Lungarella, and Y. Kuniyoshi, "Body schema adaptation for robotic tool use", Advanced Robotics, Vol. 20, 10, pp. 1105-1126. 2006.

[11] L. Natale, F. Orabona, G. Metta, and G. Sandini, "Sensorimotor coordination in a "baby" robot: learning about objects through grasping" Progress in Brain Research, From Action to Cognition, Vol. 164, pp. 403-424, 2007.

[12] A. Stoytchev, "Toward Video-Guided Robot Behaviors" Proceedings of the 7th International Conference on Epigenetic Robotics. pp. 165-172. 2007.

[13] M. Hersch, E. Sauser and A. Billard, "Online learning of the body schema" International Journal of Humanoid Robotics, 2008.

[14] S. Fuke, M. Ogino, and M. Asada "Body Image Constructed From Motor and Tactile Images With Visual Information" International Journal of Humanoid Robotics (IJHR) Vol. 4, 2, pp.347-364, 2007.

[15] G. M. Stratton "Vision without inversion of the retinal image" Psychological review, Vol. 4, pp. 463-481. 1897.

[16] T. Kohonen, "Self-organizing maps" Springer-Verlag Verlin Heidelverg, 1995.

[17] G. Rizzolatti and M. Matteli, "Two different streams form the dorsal visual system: anatomy and functions" Experimental Brain Research, Vol. 153, 2, pp. 146-157, 2003.

[18] A. Berti and F. Frassinetti, "When far becomes near:Remapping of space by tool use" Journal of Cognitive Neuroscience, Vol. 12, pp. 415-420, 2000.