Visual attention by saliency leads cross-modal body representation

Mai Hikita*, Sawa Fuke*, Masaki Ogino†, Takashi Minato† and Minoru Asada*†
*Graduate School of Engineering, Osaka University,
2-1 Yamadaoka, Suita, Osaka, Japan
Email: [mai.hikita, sawa.fuke, asada]@ams.eng.osaka-u.ac.jp
†JST ERATO Asada Synergistic Intelligence Project
2-1 Yamadaoka, Suita, Osaka, Japan
Email: [ogino, minato, asada]@jeap.org

***Abstract*—One of the most fundamental issues for physical agents (humans, primates, and robots) in performing various kinds of tasks is body representation. Especially during tool-use by monkeys, neurophysiological evidence shows that the representation can be dynamically reconstructed by spatio-temporal integration of different sensor modalities so that it can be adaptive to environmental changes [1]. However, to construct such a representation, an issue to be resolved is how to associate which information among various sensory data. This paper presents a method that constructs cross-modal body representation from vision, touch, and proprioception. Tactile sensation, when the robot touches something, triggers the construction process of the visual receptive field for body parts that can be found by visual attention based on a saliency map and consequently regarded as the end effector. Simultaneously, proprioceptive information is associated with this visual receptive field to achieve the cross-modal body representation. The proposed model is applied to a real robot and results comparable to the activities of parietal neurons observed in monkeys are shown.**

I. INTRODUCTION

Humans can perform various kinds of tasks through interaction with objects, usually without consciousness of their own body representation in their brains, based on which humans are supposed to decide what action to take. Such representation has been referred to as "body schema," an unconscious neural map in which multi-modal sensory data are unified [2] or "body image," an explicit mental representation of the body and its functions [3] and brain and medical scientists have investigated the property of such representations. [ex., [4], [5], and [6]]. Among these studies, Iriki et al. [1] focused on the neural activity of the intraparietal cortex in the monkey brain before and after the monkey learned how to manipulate a rake. They showed that the bimodal neurons responding to both somatosensory and visual stimulation of the hand were also activated by the visual stimulation near the rake as if the hand had extended after training. This neurophysiological evidence suggests that the body representations in the biological systems are flexible and acquired by spatio-temporal integration of the different sensory data. However, when and how such representation is acquired is left unsolved.

The conventional approaches in robotics specify the body representation with exact values of parameters such as link structure and internal/external sensing parameters. Therefore, this sort of representation is not as adaptive as that of the biological systems. In cognitive developmental robotics [7], a number of models for adaptive body representation have been studied in order not only to understand the acquisition process of body representation in humans but also to apply these models to robots that can develop such representation. Asada et al. [8] proposed that a robot finds its own body in the visual image based on the change of sensation that correlates with the motor commands. Yoshikawa et al. [9] proposed a model in which a robot can detect its own body in the camera image based on the invariance in multiple sensory data. In Nabeshima et al.'s model [10], the robot learns the properties of the robot controller following the synchronization of the activation of visual and somatic sensations while the robot is using a tool. Furthermore, Stoytchev [11] proposed the model that enables a robot to detect its own body in a TV monitor based on the synchronization of the activation of vision and proprioception.

In these studies, the synchronization of activations among different sensing modalities is a central issue in finding the body parts, and then cross-modal integration of such data is applied to construct the body representation. These representations are necessary to perform various kinds of tasks. However, it does not seem sufficient, especially in the use of tools, when the location of the end-effector and its movements are key components for the task. In order to focus on these components, a visual attention mechanism in the biological systems seems to work well. Generally, four processes are fundamental to attention: working memory, top-down sensitivity control, competitive selection, and automatic bottom-up filtering for salient stimuli [12]. The first three are related to voluntary control of attention (top-down) while the last one is bottom-up. The former supposes that the agent has already experienced many tasks and therefore, has acquired more abstracted representation than in the case of the latter. As the first step towards the adaptive body representation, we may start from the latter.

This paper presents a method that constructs cross-modal body representation from vision, touch, and proprioception. Tactile sensation when a robot touches something triggers the construction process of the visual receptive field for

body parts that can be found by visual attention based on a saliency map and consequently regarded as the end-effector. Simultaneously, proprioceptive information is associated with this visual receptive field to achieve the cross-modal body representation. The proposed model is applied to a real robot and results comparable to the activities of parietal neurons observed in monkeys are shown. Future issues are discussed as well.

II. NEUROPHYSIOLOGICAL FINDING

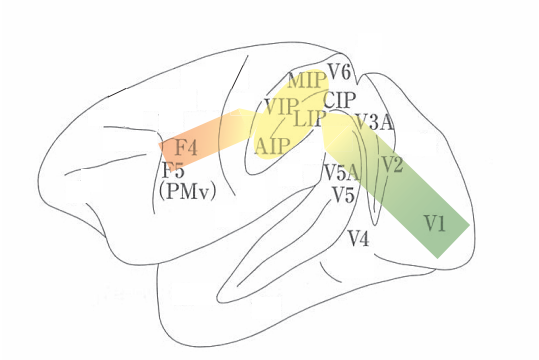
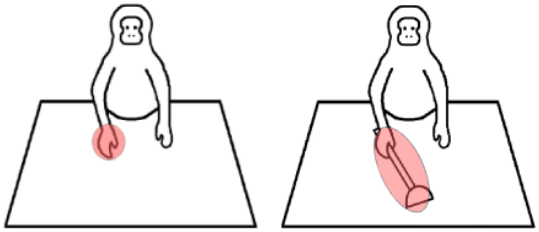


Fig. 1. Parietal cortex in the brain of a monkey (from[13])

Fig. 1 shows the parietal cortex and related areas in the monkey brain. The parietal cortex (the yellow region) is reported as being the area where the multi-modal sensations such as proprioception, vision and auditory are integrated [14][15]. As mentioned in the previous section, Iriki et al. [16] recorded the activities of some neurons in that area before and after Japanese macaques were trained to use a tool. In that training, the monkeys did not imitate the experimenter's behavior but learned how to use the tool abruptly after about 2 weeks. The investigated neurons named "bimodal neurons" responded to both somatosensory stimuli at the hand and visual stimuli. Fig. 2 (a) shows its visual receptive field defined as the region where the neurons are activated with the visual pointer. After the monkeys became able to reach the target (food) with the tool through the training, the bimodal neurons also responded to the visual stimuli near the tool as shown in Fig. 2 (b).

Furthermore, they also investigated the activities of these neurons while the monkeys were using the tool through a TV-monitor [17] with several conditions to check how adaptive the visual receptive field is. It is easily and dynamically modified according to the change of its visual sensation and surprisingly the monkey's intention of using the tool as well. These results suggest that monkeys can dynamically change their visual receptive field as if the tool became a part of their own bodies during tool use. Through these experiences, the monkeys are expected to have a category of "tools" and to regard the end effectors (hands) as tools vice versa. A big mystery from the viewpoint of cognitive developmental robotics is how a robot can acquire such adaptive representation expected to develop higher and more abstracted representation.



(a) Before tool-use (b) After tool-use

Fig. 2. Changes in bimodal visual receptive field(from[17])

III. BASIC IDEA FOR A SYNTHETIC MODEL

As the first step toward solving this big mystery, let us consider the following points

1) We suppose that the visual attention is a key issue to achieving the visual receptive field not simply because the visual attention mechanism can focus on the salient features (bottom-up flow) but also because such a field can be activated when attention is directed to it in some way (top-down flow), like the activation of the monkey's visual receptive field by the visual pointer.
2) Considering the acquisition of tool category in the future, it is important to start with finding the end effector not supposing it can be easily found by its salient visual feature such as color [10] but expecting that it is constructed through the learning process.
3) To acquire the cross-modal body representation, tactile sensation is essential. In our model, we utilize the tactile sensation to trigger the process, that is, when it touches something, a robot associates the salient visual features with the proprioceptive data.

IV. THE MODEL FOR BODY REPRESENTATION BASED ON VISUAL ATTENTION

Considering the basic idea mentioned above, a synthetic model for body representation is proposed. An overview of the whole system is shown in Fig. 3 where three modules are involved. The arm posture module corresponds to the proprioception, representing various kinds of postures in terms of joint angles that are collected and structured as an SOM (self-organizing map). The attention module detects the salient features in the camera image as the candidates for attention point based on a saliency map algorithm [18] in every step. The integration module associates the arm posture with the visual attention point by Hebb Learning when the robot detects the tactile sensation by hitting a target with its hand or a tool. This module can be regarded as a model of the neuron in the parietal cortex. Details of each module are given next.

A. Robot Model

In order to validate the model, the proposed system is applied to a humanoid robot, CB^2 [19] (Fig. 4). This robot was developed by the JST ERATO Asada Project as a research platform for cognitive developmental robotics. It has soft skin and flexible joints (51 pneumatic actuators). Under the

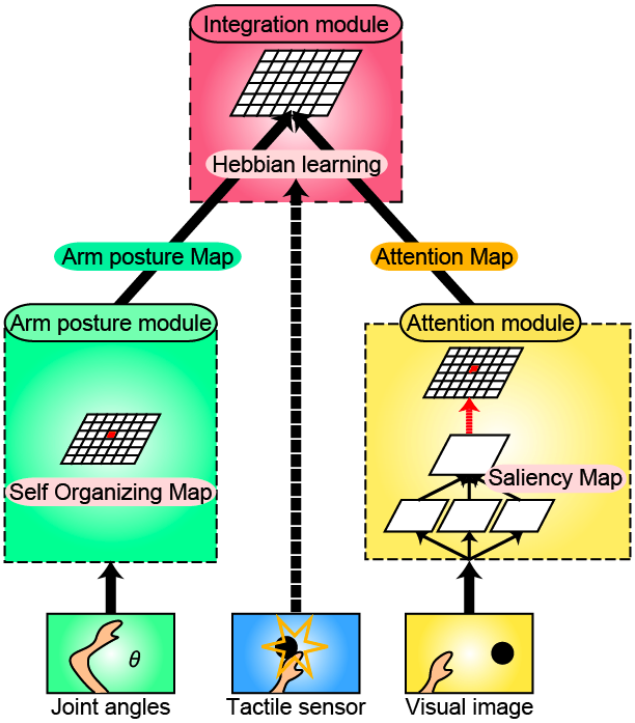


Fig. 3. Overview of the proposed system

soft skin, it has about 200 tactile sensors to achieve tactile perception. It is about 130 [cm] tall and weights about 33 [kg].

Fig. 4. Child-robot with Biomimetic Body for cognitive developmental robotics: CB^2

B. Arm posture module

The robot moves its hand toward random positions on the table. The six joint angles of the left arm are recorded and used for the construction of an SOM as training data. SOM is a kind of artificial neural network [20]. It describes a mapping from a higher dimensional input space to a lower dimensional (typically two dimensional) map space. In this experiment, the dimension of the SOM is two and the number of units is 8×8. Thus, the representative vector of the i-th unit is

$$\mathbf{\Theta}_i = (\theta_1^i, \theta_2^i, ..., \theta_n^i), \tag{1}$$

where n is the number of joint angles, which is six here.

While the robot is probing on the table, the activity level of the i-th unit in the arm posture map is computed with the actual values of the joint angles of the arm, $\mathbf{\Theta}$, as described below

$$a_i^{arm} = e^{-\beta d_i^2} \tag{2}$$

where,

$$d_i = \|\mathbf{\Theta}_i - \mathbf{\Theta}_c\|, \tag{3}$$

$$c = \arg\min_i \|\mathbf{\Theta} - \mathbf{\Theta}_i\|. \tag{4}$$

In this experiment, β is 100.

C. Visual attention module

While probing on the table, the robot pays its attention to the point detected in this visual attention module at every step. In order to model the visual attention, the saliency map algorithm is adopted. The saliency map is proposed based on biologically plausible architecture by Itti et al. [18]. The map is constructed by combining several kinds of features in the visual image.

In our experiment, the camera image size is 320×240, and a pyramid structure is adopted for quick localization of image features. The scale parameter $\sigma = [0, 1, ..., 8]$. The following visual features are used to construct the saliency map.

- Intensity: summation of rgb components
- Color: rgb components and other color features from rgb components
- Flicker: simple temporal difference (subtraction between consecutive frames)
- Gabor: the value of Gabor filters with four directions
- Motion: normal flow vectors of edge segments

The computation of these features is based on the method by Itti et al. [18].

Figs. 5 (a)-(e) show these visual features for the input image (Fig. 5 (f)) when the robot touches and pushes the object on the table. Fig. 5 (g) indicates the final saliency map that shows the moving hand and pushed object are evaluated as highly salient.

(a) Intensity (b) Color (c) Flicker (d) Gabor (e) Motion

(f) Input image (g) Output image (h) Attention map

Fig. 5. Visual attention module

The saliency map is divided into 10×10 units. The center coordinate of the j-th unit is defined as x_j and the sum

of salient feature values in the unit S_j is calculated as the represented saliency value of the area. Then the attention unit is selected from the points whose S_j is over the threshold D. D is selected randomly from 0 to 1. Finally, the attention map is constructed as shown in Fig.5 (h).

The k-th unit of activation level of the attention map is calculated as follows

$$a_k^{attention} = e^{-\gamma s_k^2}, \quad (5)$$

$$s_k = \|\boldsymbol{x}_k - \boldsymbol{x}_c\|, \quad (6)$$

where $\boldsymbol{x}_c$ is the ID of the activated unit on the attention map.

D. Integration module

The integration module has 10×10 units, each of which receives the signal from the units of the attention map and the arm posture map via the connection weight matrices, w^A and w^B, respectively. w^A is fixed so that the activation level of the attention map is directly conveyed to the corresponding unit of the integration module

$$w_{jk}^A = \begin{cases} 1 & (\text{if } j = k) \\ 0 & \text{else} \end{cases}. \quad (7)$$

On the other hand, the arm posture module and integration module are associated based on the Hebbian learning algorithm when the robot detects the tactile activation with its own hand. The weight between a unit of the arm posture map and a unit of the integration map increases if the activation levels of units are high and vice versa.The connection weights between two maps, w_{ik}^B, are updated as follows,

$$\Delta w_{ik}^B = \epsilon a_i^{arm} a_k^{integrate}, \quad (8)$$

$$w_{ik}^B(t+1) = w_{ik}^B(t) + \Delta w_{ik}^B, \quad (9)$$

$$w_{ik}^B(t+1) \leftarrow \frac{w_{ik}^B(t+1)}{\sum_{k=0}^{N_a} w_{ik}^B(t+1)}, \quad (10)$$

where N_a is the number of units of the attention map. In this experiment, N_a is 100 and ϵ is 0.05. $a_k^{integrate}$ is the activation level of the unit of the integration module. In learning phase,

$$a_k^{integrate} = a_k^{attention}. \quad (11)$$

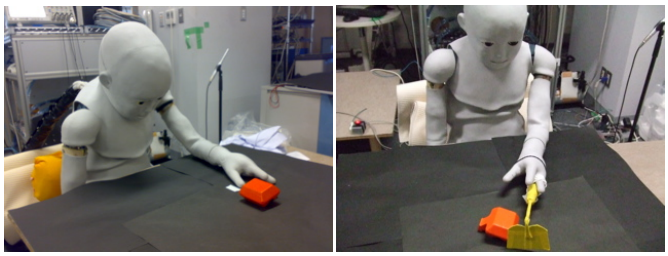
The initial values of w_{ik} are 0.5s, which means that one posture is associated with all units of the integration module at the same level.

V. EXPERIMENTAL RESULT

A. Setting of the experiment

The proposed model was applied to a real robot, CB^2, with an experimental environment similar to that with macaque monkey by Iriki et al. [1] as shown in Fig. 2.

The robot is placed in front of a black table. The robot probes with its own left hand or a tool on the table as shown in Fig. 6. The target object colored orange is positioned randomly every time the robot touches the object with its own body (including a tool). During learning, the posture of the robot is fixed except for the probing hand and the eyes of the robot (from which the camera images are captured). Although the robot has many tactile sensors throughout is body, the density of the tactile sensors in the hand is not enough to detect the touched sense of the object. Thus, the trigger for the learning is given by the experimenter when she observes that the robot hand touches the object. For the tool use case, the learning trigger is given when the experimenter observes the tool touching the object, expecting that the tactile sensing is affected by the contact between the tool and the object.



(a) With hand (b) With tool

Fig. 6. Robot probes on the table with the hand and a tool

B. Evaluation of the learning process

In order to evaluate the learning maturation of Hebbian learning in the integration module, the averaged variance of the weights between one unit of the arm posture map and all units of the integration map is calculated. The averaged position on the integration map, $\bar{\boldsymbol{r}}^i$, connected from the i-th unit of the arm posture map is calculated as

$$\bar{\boldsymbol{r}}^i = \frac{\sum_{k=1}^{N_a} w_{ik} \boldsymbol{r}_k}{\sum_{k=1}^{N_a} w_{ik}}, \quad (12)$$

where $\boldsymbol{r}_k$ denotes the position vector of the k-th unit on the integration map. Furthermore, the variance of the connection weights from the i-th unit of the arm posture map, $\hat{r}^i$, is calculated as

$$(\hat{r}^i)^2 = \frac{\sum_{k=1}^{N_a} w_{ik} \|\boldsymbol{r}_k - \bar{\boldsymbol{r}}_k^i\|^2}{\sum_{k=1}^{N_a} w_{ik}}. \quad (13)$$

The time courses of the variance for one unit of the arm posture map with and without the tool are shown in Fig. 7. As the learning proceeds, each variance becomes smaller and the connection weights between the units converge. The variance curve for the hand with the tool goes up and down a little bit after 30 steps. This may be because the posture of the robot (and so the camera position) slightly changes from the initial position.

C. The connection weights

We examine the connection weights between one unit of the arm posture map and every unit of the integration map. In Figs. 8, the level of connection weights that connect from one unit of the arm posture map visualized by color under each condition is superimposed on the robot camera view. The redder the color, the higher the connection weight is. The

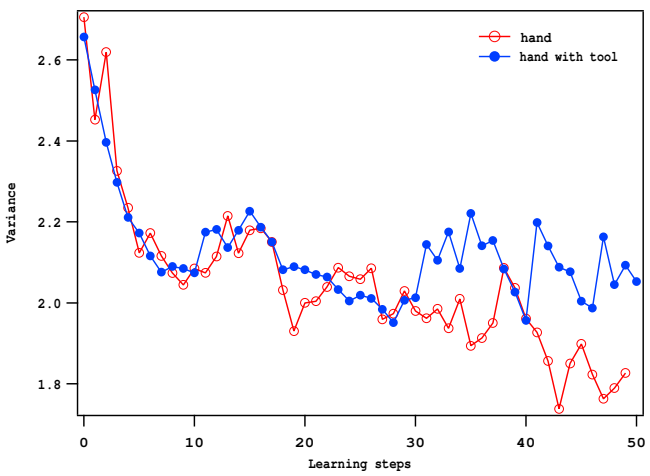
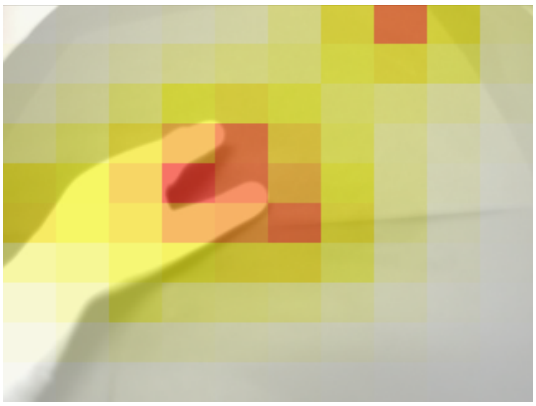


Fig. 7. Variances of connection weights as Hebbian learning proceeds

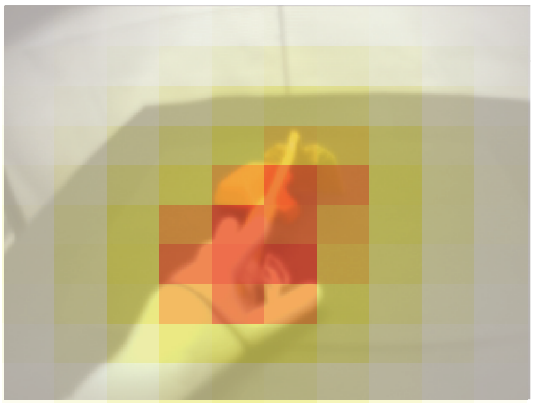
imposed camera image is captured when the robot assumes the posture corresponding to the unit of the arm posture map. Because the attention module is directly connected to the integration module in this model as expressed in Eq. 7 the relation between the arm posture map and the attention map can be directly compared in this figure. Fig. 8 (a) shows that the connection weights converge most strongly at the area around the end effector (the hand of the robot). This implies that the hand area is the most salient for this robot when it touches the object. On the other hand, contrary to the expectation, the upper-right area is also weighted high. When the robot touches the object, the robot posture sometimes changes slightly. This causes the motion flow in the camera image and makes the saliency of the corner of the table same level as that of the hand and the object. Fig. 8 (b) shows that the connection weights are extended to the tool area. These results are comparable to those of the experiments with macaque monkeys as shown in Fig. 2. In Iriki's experiments, the neurons expanded their receptive fields after the training. On the other hand, we deal with the hand-use case and the tool-use case separately. However the same results in computer simulations were obtained from the case the robot used tool after the hand-use case.

D. The activation of the bimodal neuron

The connection weights shown in Fig. 8 are regarded as the visual receptive field of the neurons of the parietal cortex observed in macaque monkeys. After the learning of the connection weights, it is expected that the robot can evaluate whether the attended point belongs to (or is near) its body or not by the activation level of the unit of the integration module. In order to show this, we conduct an experiment similar to that of Iriki et al. [1]. They investigate the visual receptive field by recording the activation level of the parietal neuron when various positions in front of the monkey are pointed at with a laser pointer. In the same manner, the light of the laser pointer is presented at in the various points in front of the robot in a



(a) With hand



(b) With tool

Fig. 8. The level of the connection weights

dark room. This is very effective for controlling the attention point of the robot because the bright light of the laser in the dark is extremely salient in the robot view.

In this inspection phase, the activation level of the unit of the integration module, $a_k^{integrate}$, is calculated as follows:

$$a_k^{integrate} = (\sum_j w_{jk}^A a_j^{attention})(\sum_i w_{ik}^B a_i^{arm}) \quad (14)$$

$$= a_k^{attention}(\sum_i w_{ik}^B a_i^{arm}). \quad (15)$$

This equation can be interpreted as meaning that the integration unit compares the current attention point, $a_j^{attention}$, with the remembered posture of the body, $\sum_i w_{ik}^B a_i^{arm}$. The activation level is conveyed to the experimenter by the alert sound that is played when the activation level of the unit exceeds some threshold.

Fig. 9 shows the experimental result. The red points indicate the laser point positions pointed at by the experimenter, and the graph shows the alert sound level corresponding to the activation level of the unit of the integration module. This result shows that, when the robot attends to the area that belongs to or near its body, the integration unit is activated in the same manner observed in the experiments with macaque monkeys.

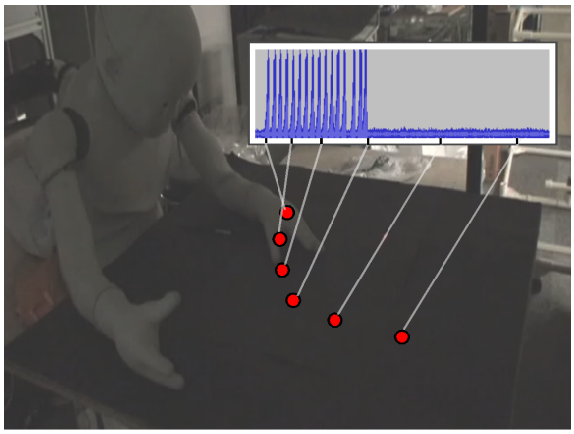


Fig. 9. Activation level of bimodal neuron depending on the laser pointer positions

VI. Discussion and Future Issues

The proposed model enabled the robot to acquire its body part (supposed to be its end effector) representation by associating the proprioceptive and visual information (both its hand and a tool) using the visual attention model based on a saliency map. We use tactile sensation to trigger learning. Without tactile sensation, the relationship between visual and proprioceptive information might be acquired, but the learning would be quite inefficient and inaccurate because tactile sensation makes the possibilities of wrong association much less and the boundary of the visual and proprioceptive information clearer. The changes in tactile information is one of the key to construct the body representations. When the robot pays its attention to a point on the table after learning, it can judge whether the attention point is near its body or not by finding the difference in position associated with the posture. On the other hand, the robot activates the units on the arm posture map when the robot detects the visual stimulus of a red point on the table as shown in Fig. 9. This may imply that the acquired representation is one of the foundations of "body image" that is recalled by its conscious behavior (attention). We also suggest that the representation in Figs. 8 can be stated as the visual receptive field of the neuron in the parietal cortex area in Fig. 2.

In our model, the attention mechanism has an important role in acquiring the body representation as a bottom-up mechanism. Although we have not examined the recalling process exactly, another attention mechanism to trigger the recalling process should be considered as a top-down mechanism. For example, when the robot encounters a novel tool, we would like it to be able to shift its attention to the image features that are similar to those of the end effectors of its body parts, so that the robot can manipulate the tool. This can be a sort of so-called affordance. If the visual background is complex or changing, the integration would not accomplished correctly. To solve this, a top-down mechanism (such as affordance) would be useful. And the visual field of the robot must be fixed in our model. This will be solved if the robot acquires the coordinate system [21].

In addition to the above issues, other future work remain, such as introduction of a reward system for learning how to use a tool, formalization of affordance consistent with the proposed model, and temporal contingency to more adaptive body detection in the visual images.

References

[1] A. Iriki, M. Tanaka, and Y. iwamura, "Coding of modified body schema during tool use by macaque postcentral neurones," *Neuroreport*, vol. 7, pp. 2325–2330, 1996.

[2] H. Head and G. Holmes, "Sensory disturbances from cerebral lesions," *Brain*, vol. 34, pp. 102–254, 1911/1912.

[3] S. I. Maxim, *Body Image and Body Schema*. John Benjamins Publishing Company, 2005.

[4] V. S. Ramachandran and S. Blakeslee, *Phantoms in the Brain: Probing the Mysteries of the Human mind*. New York: William Mollow, 1998.

[5] Y. Iwamura, "Hierarchical somatosensory processing," *Current Opinion in Neurobiology*, vol. 8, pp. 522–528, 1998.

[6] M. V. Peelen and P. E. Downing, "The neural basis of visual body perception," *Nature Reviews Neuroscience*, vol. 8, pp. 638–648, 2007.

[7] 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, no. 185–193, 2001.

[8] M. Asada, E. Uchibe, and K. Hosoda, "Cooperative behavior acquisition for mobile robots in dynamically changing real worlds via vision-based reinforcement learning and development," *Artificial Intelligence*, vol. 110, pp. 275–292, 1999.

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

[10] C. Nabeshima, M. Lungarella, and Y. Kuniyoshi, "Body schema adaptation for robotic tool use," *Advanced Robotics*, vol. 20, no. 10, pp. 1105–1126, 2006.

[11] A. Stoytchev, "Toward video-guided robot behaviors," in *Proceedings of the 7th International Conference on Epigenetic Robotics*, 2007, pp. 165–172.

[12] E. I. Knudsen, "Fundamental components of attention," *Annual Review of Neuroscience*, vol. 30, pp. 57–78, 2007.

[13] H. Sakata, M. Taira, M. Kusunoki, and A. Murata, "The parietal association cortex in depth perception and visual control on hand action," *TRENDS in Neurosciences*, vol. 20, pp. 350–357, 1997.

[14] M. S. Graziano and D. F. Cooke, "Parieto-frontal interactions, personal space, and defensive behavior," *Neuropsychologia*, vol. 44, pp. 845–859, 2006.

[15] M. I. Sereno and R. Huang, "A human parietal face area contains aligne d head-centered visual and tactile maps," *Nature Neuroscience*, vol. 9, no. 10, pp. 1337–1343, 2006.

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

[17] A. Maravita and A. Iriki, "Tools for the body (schema)," *Trends in Cognitive Sciences*, vol. 8, no. 2, pp. 79–96, 2004.

[18] L. Itti and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention," in *Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology*, vol. 5200, 2003, pp. 64–78.

[19] T. Minato, Y. Yoshikawa, T. Noda, S. Ikemoto, H. Ishiguro, and M. Asada, "Cb2: A child robot with biomimetic body for cognitive developmental robotics," in *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, 2007, p. CDROM.

[20] T. Kohonen, *Self-organizing maps*. Verlin Heidelverg: Springer-Verlag, 1995.

[21] S. Fuke, "Vip neuron model: Head-centered cross-modal representation of the peri-personal space around the face," *In Proceedings of the IEEE 7th International Conference on Development and Learning*, 2008.