Cross-modal body representation based on visual attention by saliency

Mai Hikita, Sawa Fuke, Masaki Ogino, and Minoru Asada

Abstract— **In performing various kinds of tasks, body representation is one of the most fundamental issues for physical agents (humans, primates, and robots). Especially during tool-use by Japanese macaque 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 solved 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. When the robot touches something, the activation of tactile sense triggers the construction process of the visual receptive field for body parts that can be found by visual attention based on saliency map and consequently regarded as the end effector. Simultaneously, proprioceptive information is associated with this visual receptive field to construct the cross-modal body representation. The computer simulation results are comparable to the activities of parietal neurons found in the Japanese macaque monkeys. Various conditions are also investigated so that what kind of information is important to generate the same results as findings in neurophysiology.**

I. INTRODUCTION

Humans can perform various kinds of tasks through interaction with objects, usually without consciousness of their own body representation in their brains. Such representation has been referred to as ”body schema” [2] or ”body image” [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 Japanese macaque 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. [ex., [8], [9], [10], and [11]]. 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. 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 start from latter.

This paper presents a method that constructs cross-modal body representation from vision, touch, and proprioception. When the robot touches something, the activation of tactile sense 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 computer simulation results are comparable to the activities of parietal neurons found in the Japanese macaque monkeys. Various conditions are also investigated so that what kind of information is important to generate the same results as findings in neurophysiology.

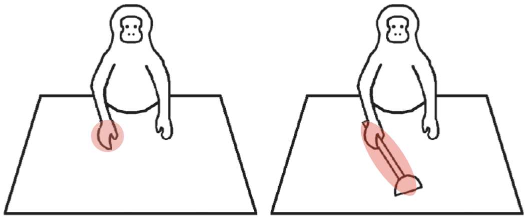
II. NEUROPHYSIOLOGICAL FINDING

Iriki et al. [1] recorded the activities of some neurons in the intraparietal cortex before and after the Japanese macaque monkeys were trained to use a tool. In that training, the monkeys did not imitate experimenter's behavior but learned how to use a tool abruptly after about 2 weeks. The investigated neurons named ”bimodal neurons” respond to both somatosensory stimuli at the hand and visual stimuli. Fig. 1 (a) shows its visual receptive field defined as the region

Mai Hikita, Sawa Fuke, and Minoru Asada are with the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Japan [mai.hikita],[sawa.fuke], [asada]@ams.eng.osaka-u.ac.jp

Masaki Ogino is a researcher of Asada Synergistic Intelligence Project, ERATO, Japan ogino@jeap.org

Minoru Asada is a research director of Asada Synergistic Intelligence Project, ERATO, Japan



(a) Before tool-use

(a) After tool-use

Fig. 1. Changes in bimodal receptive field properties (from [13])

where the neurons are activated with the visual pointer. After the monkeys became able to reach the target (food) with it through the training, bimodal neurons also responded to the visual stimuli near the tool as shown in Fig. 1 (b).

These results suggest that the 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 a viewpoint of cognitive developmental robotics is how a robot can acquire such adaptive representations that is expected to be developed to higher and more abstracted ones.

III. BASIC IDEA FOR A SYNTHETIC MODEL

As the first step towards the 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 Japanese macaque 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 visual salient features with the proprioceptive data.

IV. THE MODEL FOR BODY REPRESENTATION BASED ON VISUAL ATTENTION

Considering the basic idea mentioned above, the synthetic model for body representation is proposed. An overview of the whole system is shown in Fig. 2 where three modules are involved. The arm posture module corresponds to the proprioception, representing various kinds of postures in terms of joint angels that are collected and structured as SOM (self organizing map) [15]. The attention module detects the salient features in the camera image as the candidates for

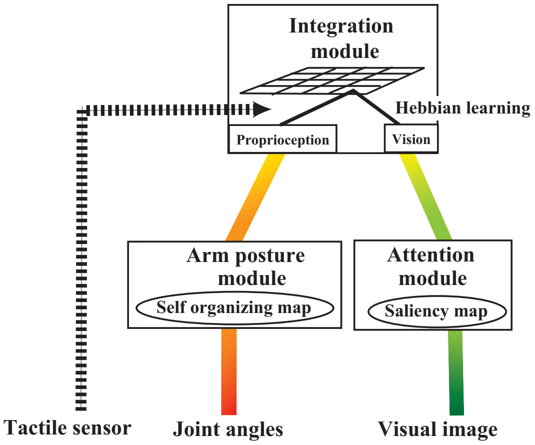
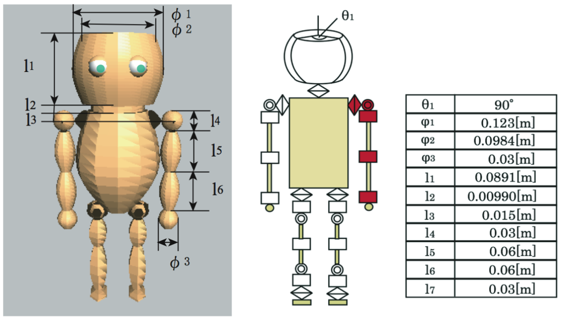


Fig. 2. An overview of the proposed model

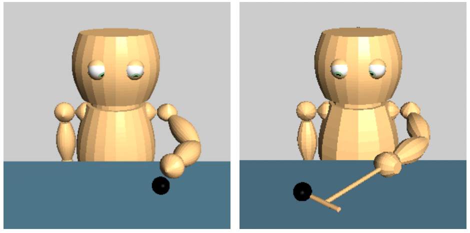
attention point based on saliency map algorithm [14] in every step. The integration module associates the arm posture with the visual attention point by Hebb Learning when the robot detects the tactile sensation by hitting a target with its hand or a tool. This module can be regarded as a model of the neuron in the parietal cortex. Details of each module are given next.

A. Robot Model



θ_1	90°
φ_1	0.123[m]
φ_2	0.0984[m]
φ_3	0.03[m]
l_1	0.0891[m]
l_2	0.00990[m]
l_3	0.015[m]
l_4	0.03[m]
l_5	0.06[m]
l_6	0.06[m]
l_7	0.03[m]

Fig. 3. Robot specifications



(a) With a hand

(b) With a tool

Fig. 4. The robot probes on the table with the hand and a tool

In order to validate the model, computer simulations are conducted with the dynamics simulator. The robot model used in the experiment and its specifications are shown in Fig. 3. Although it has both arm each of which has five degrees of freedom and binocular vision system, we utilize the left arm and the central point between two eyes (monocular vision system) for the simplicity of the experiments. The robot is placed in front of the table, 0.2[m] high and 0.5[m] wide. While learning, it probes with its own left hand or tool on the table as shown in Fig. 4 and gazes at a fixation point (no change) on the table. There is a target at the position selected randomly and it is replaced to the different one when the robot hits it with its own body (including a tool).

B. Arm posture module

The robot moves its hand toward random positions on the table. The five joint angles of the left arm (which are colored red in Fig. 3) are recorded and used for the construction of SOM as training data. The dimension of SOM is two and the number of the units is 8×8. Thus, the representative vector of the i-th unit is

$$\boldsymbol{\Theta_i} = (\theta_1^i, \theta_2^i, ..., \theta_n^i), \tag{1}$$

where n is the number of joint angles, which is five here. This learned map of the arm posture is shown in Fig. 5.

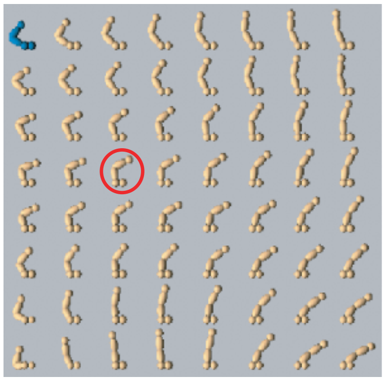


Fig. 5. Arm posture map

C. Visual attention module

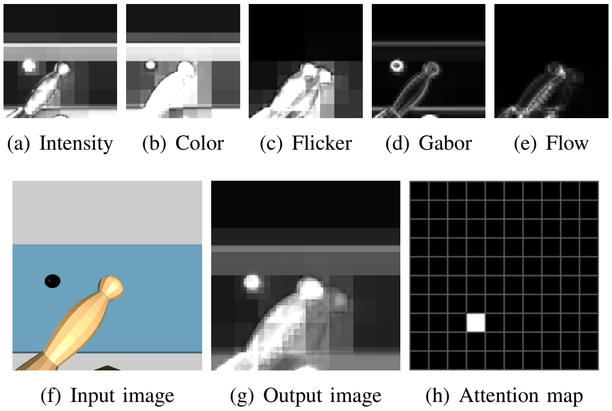
While probing on the table, the robot pays its attention to the point detected in this visual attention module at every step. In order to model the visual attention, the saliency map algorithm is adopted. The saliency map is proposed based on biologically plausible architecture by Itti et al. [14]. The map is constructed by combining several kinds of features in the visual image.

In our experiment, the camera image size is 512 × 512, 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
- Flow: normal flow vectors of edge segments

The computation of these features are based on the method by Itti et al. [14]. Flicker and flow are motion saliency.

Figs. 6 (a)-(e) show these visual features for the input image (Fig. 6 (f)) while the robot is probing on the table. Fig.6 (g) indicates the final saliency map.



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

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

Fig. 6. 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^{unit} is calculated as the represented saliency value of the area. Then the attention unit is selected from the points whose S_j^{unit} is over the threshold D. D is selected randomly from 0 to 1. Finally, the attention map is constructed as shown in Fig. 6 (h).

D. Integration module

1) Calculation of the activation level: Before the integration, we define the activation level of arm posture map and attention map. While the robot is probing on the table, the activity level of the i-th unit of the arm posture map is computed with the actual values of the joint angles of the arm, $\boldsymbol{\Theta}$, as described below,

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

where,

$$\begin{aligned} d_i &= \|\boldsymbol{\Theta}_i - \boldsymbol{\Theta}_c\|, \quad (3) \\ c &= \arg \underset{i}{min} \|\boldsymbol{\Theta} - \boldsymbol{\Theta}_i\|. \quad (4) \end{aligned}$$

In this experiment, $\beta = 100$.

In the same manner, the k-th unit of activation level of the attention map is calculated as follows.

$$a_k^{attention} = e^{-\gamma s_k^2}. \tag{5}$$

$$s_k = \|\boldsymbol{x_k} - \boldsymbol{x_c}\|, \tag{6}$$

where $\boldsymbol{x_c}$ is the ID of the activated unit of the attention map.

2) Association: 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

$$w^A_{jk} = \begin{cases} 1 & (\text{if } j = k) \\ 0 & \text{else} \end{cases} . \tag{7}$$

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 connection weights between two maps, w^B_{ik}, are updated as follows,

$$\begin{aligned} \Delta w^B_{ik} &= \epsilon a^{arm}_i a^{integrate}_k, && (8) \\ w^B_{ik}(t+1) &= w^B_{ik}(t) + \Delta w^B_{ik}, && (9) \\ w^B_{ik}(t+1) &\leftarrow \frac{w^B_{ik}(t+1)}{\sum_{k=0}^{N_a} w^B_{ik}(t+1)}, && (10) \end{aligned}$$

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

$$a^{integrate}_k = a^{attention}_k . \tag{11}$$

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

V. EXPERIMENTAL RESULTS

The proposed model described above is applied to the simulation model under some different conditions as shown in Table 1.

A. The effect of the visual features

The cases 1-6 are arranged to examine what information is effective for the construction of the body representation.

1) 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. At the beginning, one unit of the arm posture map is associated with all units of the integration module, therefore the variance at this stage is very large. As the learning process proceeds the connections between them are pruned, and therefore the variance converge to a small value. 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

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

where $\boldsymbol{r}_k$ denotes the position vector of the k-th unit of 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

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

Then, finally the connection-weight-evaluation is performed with

$$R = \frac{\sum_{i=1}^{N_b} \hat{r}^i}{N_b}, \tag{14}$$

where N_b is the number of the units of arm posture map and is 64. The result during learning is shown in Fig. 7. As the learning proceeds, each variance converges and the connection between the units seems to be potentiated. In case 6, when the robot uses a tool, the change of the value is smaller because there are more hitting points that cause the tactile activation of its own hand.

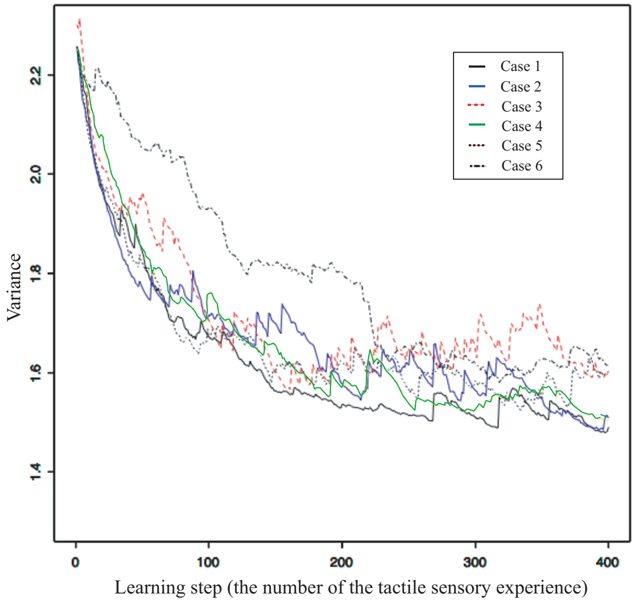


Fig. 7. The variance as Hebbian learning proceeds

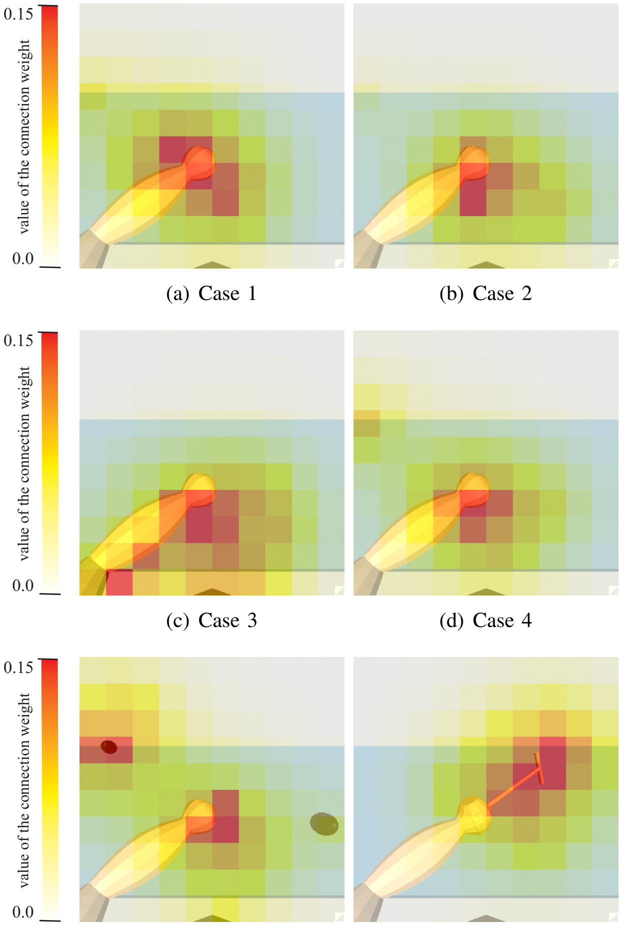
2) The connection weights: We examine the connection weights between one unit of the arm posture map and every unit of integration map. In Figs. 8, we visualize the level of connection weights by color that connect from one unit of the arm posture map (specified by a red circle in Fig. 5) under each condition, superimposed on the robot camera view. 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 these figures.

Fig. 8 (a) shows case 1, the standard condition. This indicates that the connection weights converge most strongly at the area around the hand. This result is comparable to the visual receptive field of the bimodal neurons in the Japanese macaque monkeys shown in Fig. 1 (a).

Fig. 8 (b) shows case 2 where the salient level around the target is different. Since the saliency of the target is lower in case 2 than that in case 1, the frequency how often the robot pays its attention to the target around the hand is lower in case 2. Thus the connection weights do not spread as much as case 1. This result may indicate that the reason why

TABLE I
EXPERIMENTAL CONDITION

Case	End effector	Color of the target	number of the target	Saliency	Timing of the integration
1	Hand	Black	1	all elements	Detection of the tactile sensation
2	Hand	Same with a table	1	all elements	Detection of the tactile sensation
3	Hand	Black	1	only motion	Detection of the tactile sensation
4	Hand	Black	1	without motion	Detection of the tactile sensation
5	Hand	Black	3	all elements	Detection of the tactile sensation
6	Tool	Black	1	all elements	Detection of the tactile sensation
7	Tool	Black	1	all elements	No relationship with tactile sensation



(a) Case 1 (b) Case 2

(c) Case 3 (d) Case 4

(e) Case 5 (f) Case 6

Fig. 8. The levels of weights

the actual receptive field spreads around the hand might be because of using salient target when the body representation is acquired. Figs. 8 (c) and (d) show cases 3 (only motion saliency) and 4 (without motion saliency) where the motion saliency is treated differently. The former indicates that the units of the arm posture map connect not only the units of the hand but also the ones around the arm. The result suggests that it is difficult to detect the appropriate visual information of its own end effector only with the temporal change in the visual image. Fig. 8 (e) shows case 5 where the number of the target is three. The locations of two of three targets (shown in (e)) are fixed and the remaining one target is randomly located. In this case, the wrong connection happens because the system also detects the targets as well as the end effector. Figs. 8 (a) and (f) indicate the difference in tool use. When the robot uses a tool, the connection distribution is spread out over the tool area. These results may correspond to those of the experiments with the Japanese macaque monkeys as shown in Fig. 1 (b).

3) 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 the Japanese macaque monkeys. After the learning of the connection weights, it is expected that the robot can decide 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 Japanese macaque monkey are pointed at with a laser pointer. In the same manner, the light points are presented in the various points in front of the robot in the dark. In this inspection phase, the activation level of the unit of the integration module, $a_k^{integrate}$, is calculated as follows:

$$\begin{aligned} a_k^{integrate} &= (\sum_j w_{jk}^A a_j^{attention})(\sum_i w_{ik}^B a_i^{arm}) && (15) \\ &= a_k^{attention}(\sum_i w_{ik}^B a_i^{arm}). && (16) \end{aligned}$$

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, $\sum_i w_{ik}^B a_i^{arm}$.

Fig. 9 shows the experimental result. 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 the Japanese macaque monkeys.

B. The effect of the tactile sensation

We use tactile sensation to trigger learning. The case 7 is arranged to examine the effect of using tactile sensation as the trigger of the integration. In this case, the arm posture module and integration one are associated regardless of the tactile sensation. The learned connection weights are shown in Fig. 10. As shown in Fig. 10 (b), without tactile sensation, the relationship between visual and proprioceptive information is acquired, but the connection weights are not extended

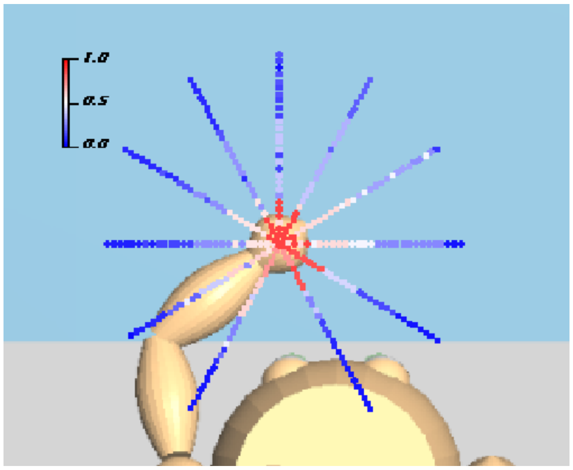
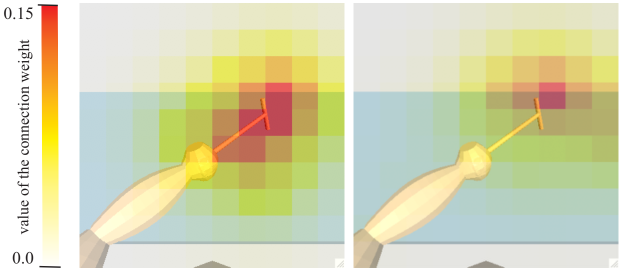


Fig. 9. Activation level of bimodal neuron depending on the laser pointer positions



(a) Case 6

(b) Case 7

Fig. 10. The levels of weights:cases 6 and 7

to the tool. This result conflicts with the experiments with the Japanese macaque monkeys. Thus tactile sensation would be necessary to acquire the body representation.

VI. Discussion

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. 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 of 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 Fig. 8 can be stated as the visual receptive field of the neurons found in the Japanese macaque monkeys in Fig. 1.

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 be accomplished correctly. To solve this, a top-down mechanism (such as affordance) would be useful. Here the visual field of the robot must be fixed in our model. This will be solved if the robot learns the association of eyeball angles and visual information [16].

In addition to the above issues, other future works 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, Y. Iwamura, "Coding of modified body schema during tool use by macaque postcentrak neurons", 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 (edited by P. D. Helena)" John Benjamins Publishing Company, 2005

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

[5] Y. Iwamura, "Hierarchical somatosensory processing" Curr Opin Neurobiol. 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. 636-648, 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, pp. 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 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] A. Stoytchev, "Toward Video-Guided Robot Behaviors", Proceedings of the 7th International Conference on Epigenetic Robotics. pp. 165-172. 2007.

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

[13] A. Maravita and A. Iriki, "Tools for the body (schema)", Trends in Cognitive Sciences, vol. 8, 2, pp. 79-96. 2004.

[14] L. Itti, N. Dhavale and F. Pighin, "Realistic avatar eye and head animation using a neurobiological model of visual attention", B. Bosacchi, D. B. Fogel, and J. C. Bezdek, editors, Proceedings of SPIE 48th Annual International Symposium on Optical Science and Technology, vol. 5200, pp. 64-78. 2003.

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

[16] Sawa Fuke, Masaki Ogino, and Minoru Asada, "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.