Proceedings of the 2nd International Symposium
on Adaptive Motion of Animals and Machines,
Kyoto, March.4-8, 2003

Does the invariance in multi-modalities represent the body scheme? - a case study with vision and proprioception -

Yuichiro Yoshikawa[*], **Koh Hosoda**[*†] **and Minoru Asada**[*†]

[*] Dept. of Adaptive Machine Systems, Graduate School of Engineering, Osaka University.

[†] HANDAI Frontier Research Center, Osaka University.

{yoshikawa, hosoda, asada}@er.ams.eng.osaka-u.ac.jp

Abstract

Adaptability to the changes in the environment and the robot body itself fundamentally depends on the robot body representation, which is usually given by the designer and therefore fixed in many cases. In order for the robot to adapt its body representation to the changes, the robot should have acquired its own body representation by itself. This paper argues how the robot can construct such representation, that is, body scheme or body image from the uninterpreted row sensory information. Supposing that the invariance in multi sensory data represents the body, a cross modal map is proposed as the structure which learns the invariance. A preliminary experiment to learn to represent the body surfaces of the robot by the cross modal mapping between vision and proprioception is performed and future issues are discussed.

1. Introduction

In the existing methods, the designer usually defines the representations of the robot body in the Cartesian coordinate system and needs to calibrate the relationship between the sensorimotor system and the defined coordinate system. Therefore, it seems difficult for the robot to adapt itself to the changes in the environment and its own body. In order for the robot to have such a capability, it seems a promising approach to provide the robot with a mechanism of acquiring representations of its body in its sensorimotor space instead of the pre-defined body representation by the designer.

Although it is a formidable problem, biological agents seem to acquire their body representation without any difficulty. There are studies about the body representations of biological agents, called *body scheme* or *body image* [1, 2, 3, 4]. Although the structure and acquiring process of them have not been revealed yet, suggestions from these studies could be helpful to construct the body representation for robots since biological agents seem in the same situation without any explicit knowledge about their own body representation. As Asada et al. advocated [5], building a robot which acquires the body representation may also enable to provide a constructive model of acquiring process of body scheme/image in human being, and understanding how it works may lead a new design principle of robots at the same time.

How to find out the body representation in the receptive field without any interpretation by the designer is one of the most fundamental problems of acquiring the body scheme. Asada et al. suggested which the robot body or static environment can be defined in a way that notes the changes in the image plane that can be directly correlated with the self-induced motor command [6] . However, discrimination between the robot body and static environment was not dealt. Fitzpatrick and Metta also proposed a similar method to localize its arm position in the vision by utilizing the correlation between optic flow and its motor commands [7]. Although they claimed that the robot found its arm without any knowledge about visual appearance, there seemed a tacit assumption that the designer needed to give a prior knowledge about the properties of its DOFs responsible for the camera motion in order to avoid difficulty of discriminating between the robot body and static environment. These studies implied that the method based on the correlation with the motion needed some prior knowledge to find its body from the correlation.

Instead of finding its body representation from the correlation with its motion, we suggest that the body can be defined by the invariance in the multi-modal sensory data caused by the fact that the sensors are embedded in the rigid robot body while the motion plays a roll of leading experiences to perceive the invariance. When the robot body is captured in some areas of the receptive fields, a kind of relation among them is invariant with the environmental changes since the body structure usually does not change in a certain period. On the other hand, when the captured areas are not the robot body, the relations among the receptive fields

of the multi modalities depend on the environmental changes. Therefore, the robot can find its body by judging whether the multi-modal relation is invariant or not.

According to this idea, Yoshikawa et al. proposed a cross modal map which learns to represent the invariance of the cooccurrence of the multi sensor modalities as the synaptic connections of fully-connected network of the sensor nodes [8]. However, they assumed that the visual patterns were segmented by the designer. In this paper, we begin with the problem how to find out the body surface in the receptive field of vision. We introduce a cross modal map by which the robot learns the invariance in the multi modalities. Based on it, the robot learns to judge whether the fixating area is its body or not.

The rest of this paper is organized as follows. First, we introduce the cross modal map between the visual and the proprioceptive modalities, and describe a learning process of it. Then we show the preliminary experiment using the upper-torso humanoid robot, and discuss our future work.

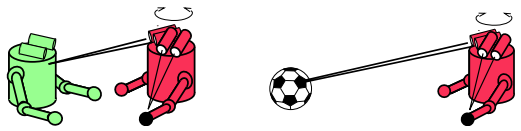
2. Cross modal map learning

In this section, we describe our basic idea to find the body of the robot and introduce the general structure called cross modal map which learns to represents the robot body. Then, we implement a cross modal map between the sensors of postural configuration and the disparity in the stereo vision in order to find the representation of the body surfaces.

2.1. A basic idea

Multi modal sensors of the robot are related with each other since they are embedded in its body although a part of the relations depends on the environment. For example, when it fixates one object in the environment, the view changes depending on the environmental changes. However, when it fixates its body, the view is independent of the environment (see Fig. 1). Our basic idea in order to find a representation of the body is learning the invariance of the relation among the multi modal sensors in a certain period. That is, what is always observed is its body.

As a structure to find the invariance in the multi modalities, we introduce a full-connected network called cross modal map. A cross modal map consists of various sensor nodes which are hardwired to real sensor units and have prototype vectors with specific dimension (see Fig. 2). When real sensors output an



(a) (b)

Figure 1: The invariance/variance of the relationship among the disparity and the postural configuration in different environments ((a) and (b)).

sensor vector $\boldsymbol{x}$, the hardwired sensor node i with a prototype vector $\boldsymbol{x}_i$ outputs an activation a_i given by the equation,

$$a_i = \exp(||\boldsymbol{x} - \boldsymbol{x}_i||/\sigma^2), \quad (1)$$

where $||\boldsymbol{x} - \boldsymbol{x}_i||$ is the distance of the vectors, σ is a scalar constant. The synaptic weight w_{ij} between node i and j is updated according to the following equation,

$$\tau \dot{w}_{ij} = -w_{ij} + c a_i a_j, \quad (2)$$

where τ is a time constant of learning, c is a learning rate. Based on the updating law (eq. (2)), w_{ij} is converged such as

$$w_{ij} = cE\{a_i a_j\}, \quad (3)$$

where $E\{a_i a_j\}$ is the average of $a_i a_j$ [9] . Actually, we use the discretizing version of the updating low (eq. (2)) such as,

$$w_{ij}(t+1) = w_{ij}(t) + \frac{1}{\tau}(a_i(t)a_j(t) - c w_{ij}(t)), \quad (4)$$

where t denotes the time stamp.

Based on this learning low, only the synaptic weights between the nodes which are simultaneously activated in a certain period are increased. Therefore, the connections which have large synaptic weight represent the body.

2.2. An implementation of a cross modal map

Suppose that a robot has m DOFs and stereo cameras, and that the center of the left camera is an fixation point. Let the disparity of the fixation point be d and the postural configuration of it be $\theta \in \Re^m$. When the posture of the learner is $\boldsymbol{\theta}$, if the fixation point is on the body, d is constant even if the environment changes, else d changes for different ones (see Fig. 1). According to this idea, the learner can find its body where the relation between d and $\boldsymbol{\theta}$ is invariant.

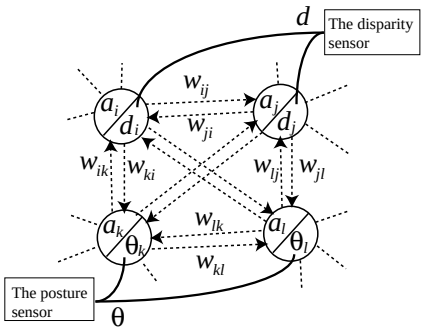
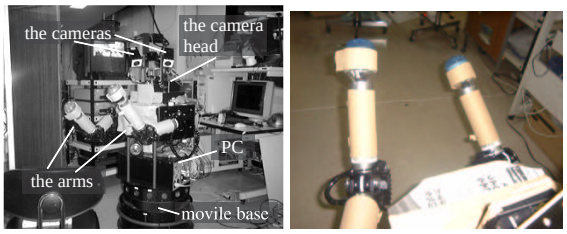


Figure 2: Elements of the cross modal map, where i, j, k,and l define ID of nodes, $d, \boldsymbol{\theta}$ denote the type of hardwired sensors, and w denotes the synaptic weights. Here, i and j are disparity nodes while k and l are posture nodes.



(a) An overview (b) An egocentric view

Figure 3: An overview and an egocentric view of the upper-torso humanoid robot and the environment of learning.

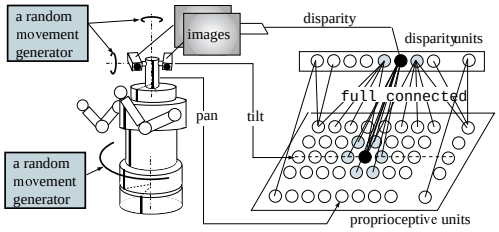


Figure 4: An overview of the experimental system.

A cross modal map consists of the two types of sensor nodes (see Fig. 2). A "disparity node" $i(j)$ hardwired to the sensor of the disparity of the fixation point has a prototype vectors $d_i(d_j)$, as well as a "posture node" $k(l)$ hardwired to the sensors of the postural configuration has a prototype vector $\boldsymbol{\theta}_k, (\boldsymbol{\theta}_l)$.

By learning through the experiences to fixate various objects and moving around, only the synaptic weights between the nodes which are activated when the fixation point is on its body are increased. Therefore, it can judge whether the fixation point is its body or not by checking whether the synaptic weight between the postural node corresponding to the current postural configuration and a disparity node is sufficiently large.

3. Experiment

We use upper-torso humanoid robot (see Fig. 3) for preliminary experiments. Based on the proposed method, it learns a cross modal map that represents its body surface through experiences to fixate the various objects and move around in the environment.

3.1. Experimental setup

The robot has two cameras (SONY, CCB-EX37), stereo-camera head which rotates in the pan/tilt/roll axes, a couple of 4-DOF arms, and a PC (Pentium II 400MHz) to control them. They are on a mobile base (Nomad150) which has facilities to moving around. Fig. 3b shows a view of it. The disparity of the fixation point is calculated in every frame when the both camera captures the same object.

3.2. Learning process and a result

As mentioned in the section 2, the robot will learn the cross modal map through the experiences of fixating various objects and moving around in the environment independent of its DOFs. We shows a section of the acquired cross modal map in which the postural configuration of the arms is a certain values (see Fig. 5), since the acquired one which consists of the high dimensional posture nodes is not comprehensible for the experimenter. Fig. 5 shows which disparity node have the largest connection with which posture nodes as a function of the disparity with respect to angles of the camera head. The range of the disparity ($d = -128 \sim 128$)) is divided into 15 prototype vectors of the disparity nodes, and the range of the angles ($pan = -45 \sim 45[deg], tilt = 10 \sim 70$) is divided into 20×15 vectors which is the elements of the posture nodes. In the learning process, the random control signals are sent to the camera head to fixate the various points and the mobile base to move around for about six minutes. The weights are updated 5875 times.

Fig. 5 is similar to one of the robot body which is observed in the real view of the robot (see Fig. 3b). The fixation areas of which disparity nodes have strong connections (large weights) to the posture nodes were parts of the robot body. Therefore, the acquired cross

modal map represents the body surface of the learner.

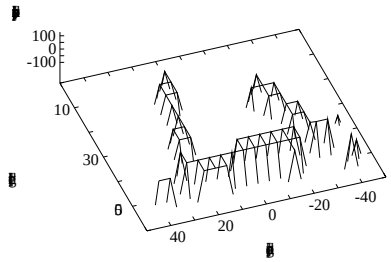


Figure 5: Activations of the disparity nodes of the acquired map.

4. Discussion and Conclusion

In this paper, we introduce the cross modal map which learns the invariance in the multi modal sensory data without any explicit knowledge, according to our consideration that such invariance represents the body. The acquired cross modal map can be used for judging whether the fixation point is its body or not. Although we cope with the cross modal map between the visual and the proprioceptive modalities, other modalities, for example tactile, could be used for learning the body representation.

As mentioned in the introduction, we aim at building a constructive model of the body scheme/image in biological agents. Although we do not have sufficient one in the current stage, we conjecture that the representation of what is always observed in a certain period is one of the elements which constitutes the body scheme/image. In the neuroscience study [2], the experimenter trained macaque monkeys to use tools and found bimodal neurons which seemed to represent image of the hand into which the tool was incorporated as its extension. This discovery does not conflict with our conjecture because the tools is always observed in the same way during their use. However, the activations of the bimodal neurons were depend on the intention of monkeys to use them. It may means that we should improve the structure of the cross modal map to be modifiable for tasks.

Although the acquired representation is only able to be used for judging whether the fixation point is its body or not, how to describe the task in the sensory-motor space, that is in the cross modal map, is one of the future work. In addition, we also need to cope with the problem how the robot can acquire some representation of its body part without the designer's explicit knowledge. For these problems, it seems necessary to integrate the multi sensor modalities including tactile and the task performing and evaluating system.

Acknowledgment This research was partly supported by the Japan Science and Technology Corporation, in Research for the the Core Research for the Evolutional Science and Technology Program (CREST) titled Robot Brain Project in the research area "Creating a brain."

References

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

[2] A. Iriki, M. Tanaka, and Y. Iwamura. Coding of modified body schema during tool use by macaque postcentral neurons. *Neuroreport*, Vol. 7, pp. 2325–2330, 1996.

[3] 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.

[4] M. S. A. Graziano, D. F. Cooke, and C. S. R. Taylor. Coding the location of the arm by signt. *Science*, Vol. 290, No. 5498, pp. 1782–1786, 2000.

[5] 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.

[6] 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.

[7] P. M. Fitzpatrick and G. Metta. Toward manipulation-driven vision. In *Proc. of the 2002 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*, pp. 43–48, 2002.

[8] Y. Yoshikawa, H. Kawanish, M. Asada, and K. Hosoda. Body scheme acquisition by cross map learning among tactile, image, and proprioceptive spaces. In *Proc. of the 2nd Intl. Workshop on Epigenetic Robotics*, pp. 181–184, 2002.

[9] S. Amari. Neural theory of association and concept-formation. *Biological Cybernetics*, Vol. 26, pp. 175–185, 1977.