Binding tactile and visual sensations via unique association by cross-anchoring between double-touching and self-occlusion

Yuichiro Yoshikawa
Graduate School of Engineering,
Osaka University
Yamada-Oka 2-1, Suita, Osaka Japan
yoshikawa@er.ams.eng.osaka-u.ac.jp

Koh Hosoda and Minoru Asada
[1)]Graduate School of Engineering,
[2)]HANDAI Frontier Research Center,
Osaka University
Yamada-Oka 2-1, Suita, Osaka Japan
{hosoda, asada}@ams.eng.osaka-u.ac.jp

Abstract

Binding is one of the most fundamental cognitive functions, how to find the correspondence of sensations between different modalities such as vision and touch. Without *a priori* knowledge on this correspondence, binding is regarded to be a formidable issue for a robot since it often perceives multiple physical phenomena in its different modal sensors, therefore it should correctly match the foci of attention in different modalities that may have multiple correspondences each other. We suppose that learning the multimodal representation of the body should be the first step toward binding since the morphological constraints in self-body-observation would make the binding problem tractable. The multimodal sensations are expected to be constrained in perceiving own body so as to configurate the unique parts of the multiple correspondence reflecting its morphology. In this paper, we propose a method to match the foci of attention in vision and touch through the unique association by cross-anchoring different modalities. Simple experiments show the validity of the proposed method.

1. Introduction

Binding is one of the most fundamental cognitive functions, how to find the correspondence of sensations between different modalities such as vision and touch, both of which are major sources of perception not only for the external world but also for the agent's body itself. The latter is closely related to the body representation which is often given by the designer and fixed but has much influence on the adaptability to the changes in the environment and the robot body itself. Assuming that the designer does not give any explicit knowledge on the body representation, a robot should construct its body representation only from its uninterpreted multimodal sensory data. In this process, *Binding* has a significant role.

Recently, researchers in other fields focus on the binding problem, which concerns the capability to integrate information of different attributes (Treisman, 1999). To propose the model for the binding mechanism of humans based on a constructivist approach, we should start with an assumption that the designer does not give any *a priori* knowledge on what the robot's sensors receive, but the robot can discriminate the different sensor modalities such as vision and touch. Since the previous work in the constructivist approach focused on the binding problem between visual attributes (Tononi et al., 1992, Seth et al., 2003), it has still not been clear how to bind different sensor modalities. Generally, receptive fields for touch and vision are simultaneously stimulated, but often respond to different physical phenomena since the foci of attention in these modalities are often different. In other words, the robot does not always watch its touching region. Therefore, to bind different modalities, the robot should correctly match the foci of attention in different modalities that may have multiple correspondences each other. However, the previous work escaped from this kind of problem by assuming that it can observe only matched sensations in different modalities (ex. (MacDorman et al., 2000)).

We suppose that learning the multimodal representation of body should be the first step toward binding since the morphological constraints in self-body-observation would make the binding problem tractable. The multimodal sensations are expected to be constrained in perceiving own body so as to configurate the unique parts of the multiple correspondence reflecting its morphology. Therefore,

building a robot that can acquire the representation from multimodal sensory data is an interesting issue from a viewpoint of a constructivist approach towards both establishing the design principle for an intelligent robot and understanding the process how humans acquire their body representation. In this study, as an example of the binding problem, we focus how it can learn to watch its body part when it detects the collision on it.

Yoshikawa et al. have proposed the method to learn the multimodal representation of the body surface through *double-touching*, that is touching its body with its own body part (Yoshikawa et al., 2002). It is based on the idea that the tactile sensors which collide with each other also coincide with each other in its vision. In other words, *self-occlusion*, that is the occlusion caused by covering its body with its own body part in its view, always occurs at the double-touching part. They assumed that there is only one self-occlusion at a moment. However, there can be multiple self-occlusions since the body occupy a certain volume in the physical space. For example, when the agent touches its body trunk with its hand, not only the hand but also its arm cover its body trunk from its sight, i.e., multiple self-occlusions occur. Therefore, there still remains the binding problem where it must determine which self-occlusion should be bound to the double-touching and vice versa.

As in the previous work (Yoshikawa et al., 2002), it seems reasonable to utilize the fact that self-occlusion always occurs at the double-touching part. In the rest of this paper, we presents the method to match the foci of attention to its own body in vision and touch by virtue of the morphological constraint. In the proposed method, the mismatched responses in these modalities can be discarded through the process of *unique association* where corresponding pairs of subsets in different attributes are exclusively connected with each other by what we call *cross-anchoring*.

2. Learning mechanism

In the following argument, we suppose that it has a human-like configuration in which it has a trunk with a camera and a end-effector connected through the serial links, that is, the robot consists of its trunk, a camera head and a arm. Furthermore, we assume that the robot has acquired the competences to detect self-occlusion and double-touching.

Problem in the statistical approach Since the robot does not have *a priori* knowledge how to bind, we suppose that it keeps changing the posture both of its arm and its camera head at random to explore for binding. In the exploration, it perceives its posture and the view in the center region of its camera. Fig. 1 illustrates the simplified situations of the robot's exploration for binding.

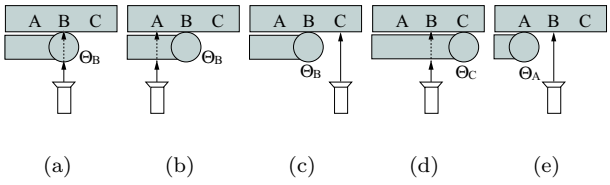


Figure 1: Five simplified situations in the robot exploration: Top rectangle indicates the robot's trunk while the rectangle with a circle in the middle indicates the robot's arm. The arrow from the bottom object shows the its focus of attention in vision.

The fact that the body occupies a certain volume in the physical space remains binding problem formidable. For example, a double-touching posture causes self-occlusions at multiple parts (see Figs. 1 (a) and (b)) while a self-occlusion at a part is caused by several double-touching postures (see Figs. 1 (a) and (d)). In the explorations, the robot sometimes experiences the matched responses in the different modalities which are caused by focusing on the same region, in this case detecting the self-occlusion at the double-touching point (see Fig. 1 (a)). However, such experiences of the matched response is not significantly frequent compared to mismatched responses (see Fig. 1 (b) or (d)) since it explores at random instead of utilizing *a priori* knowledge. In other words, the correctly matched responses are not significantly major in the obtained data. Therefore, it is difficult to associate them by considering all obtained data through the exploration. Then, we need a mechanism to narrow down the influence of the mismatched data on learning while augmenting the influence of the matched one.

Cross-anchoring Hebbian learning rule We can utilize the following two morphological constraints: 1) how many double-touching postures occludes a certain part on the trunk depends on the location of the part to be occluded, and 2) how many parts the robot occludes by a double-touching posture depends on the location of the contact part. These facts indicate that there exist *cue* nodes which have fewer candidates for matched response in other modalities to be bound. Since the desired correspondence between touch and vision can be found by unique association in this case, we can utilize such cue nodes as anchors of the unique association. Therefore, we introduce a learning rule with an anchoring mechanism which can adapt the learning rate according how much the responses simultaneously observed are regarded as unique to each other.

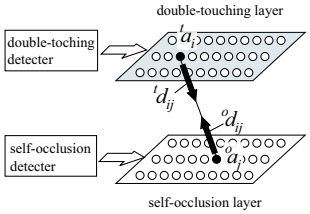


Figure 2: The architecture

The architecture consists of two layers called the double-touching layer and the self-occlusion one (see Fig. 2). In the double-touching layer, there are N_t nodes each of which is responsible for a set of certain posture of the arm $\Theta_i, (i = 1, \cdots, N_t)$ which is assumed to be pre-clustered. When the posture of the arm is $\boldsymbol{\theta} \in \Re^m$, the activation of the i-th node ${}^t a_i(\boldsymbol{\theta})$ is 1 if $\boldsymbol{\theta} \in \Theta_i$, otherwise 0. On the other hands, in the self-occlusion layer, there are N_o nodes each of which is responsible for the self-occlusion in a set of certain posture of the camera head $\Phi_j, (j = 1, \cdots, N_o)$ which is assumed to be pre-clustered. When the posture of the camera head is $\boldsymbol{\phi} \in \Re^n$, the activation of the j-th node ${}^o a_j(\boldsymbol{\phi})$ is 1 if $\boldsymbol{\phi} \in \Phi_j, O$, otherwise 0, where O is the phenomenon of detecting occlusion.

Let the connection weight between the i-th node in the double-touching layer and the j-th node in the self-occlusion layer be w_{ij}. By the cross-anchoring Hebbian learning rule, $w_{i^*j^*}$ is updated as following:

$$\Delta w_{i^*j^*} = \eta({}^t d_{i^*j^*} {}^t a_{i^*} \cdot {}^o d_{i^*j^*} {}^o a_{j^*} - w_{i^*j^*}), \quad (1)$$

where i^* and j^* are the most activated units in the double-touching and the self-occlusion layer, η is a constant learning rate. The dynamic anchoring rates, ${}^t d_{ij}$ and ${}^o d_{ij}$, determine the degrees of anchoring on the j-th node in the self-occlusion layer from the i-th nodes in the double-touching layer and on the i-th node in the double-touching layer from the j-th nodes in the self-occlusion layer, respectively. They are calculated by

$$\begin{aligned} {}^t d_{ij} &= \exp\left(-\frac{\sum_{k,k\neq j} w_{ik}}{{}^t\sigma_a^2}\right), \\ {}^o d_{ij} &= \exp\left(-\frac{\sum_{k,k\neq i} w_{kj}}{{}^o\sigma_a^2}\right), \end{aligned} \quad (2)$$

where ${}^t\sigma_a$ and ${}^o\sigma_a$ are parameters that determine the degree of anchoring. Meanwhile, the remaining connection weights are decreased because they loss the competition;

$$\begin{aligned} w_{ij^*}(t+1) &= w_{ij^*}(t) - \eta_t(1 - {}^t d_{ij^*})\Delta w_{i^*j^*}, \\ w_{i^*j}(t+1) &= w_{i^*j}(t) - \eta_o(1 - {}^o d_{i^*j})\Delta w_{i^*j^*}, \end{aligned} \quad (3)$$

where η_t and η_o are constant coefficients of the competition.

In such an anchoring process, more unique combinations of double-touching and self-occluded are bound earlier since they obtain larger anchoring rates according to eqs. (2). Meanwhile, some of the rest combinations become more unique since the other responses decrease the number of candidates to be bound by losing the responses that are already bound to others. Therefore, the process of binding proceeds step by step. This process is expected to converge since it is considered that there exist cue nodes in each modality due to the constraint in its human-like configuration.

3. Simulation

As preliminary experiments, we tested the cross-anchoring Hebbian learning rule works so that the robot solves the binding problem by using the computer simulation of a robot with a 1-DoF sliding arm, a 1-DoF rotating camera head as in Fig. 1. During the exploration for binding, it moves its arm and camera head at random and detects self-occlusion and double-touching.

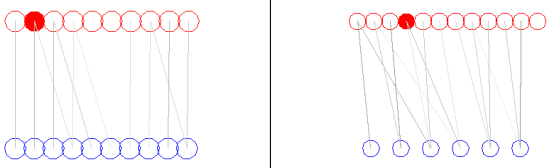
For the reader's understanding, we pre-clustered the posture space both of the arm and the camera head so that the nodes in both layers were matched with each other in one-to-one manner. The robot was trained for binding in 4,000 double-touching trials with the following network parameters: $N_t = N_o = 10$, $\eta = 0.1, \eta_t = \eta_o = 0.5$, and ${}^t\sigma_a = {}^o\sigma_a = 1.0$. Figs. 3 (a):(I) $\sim$ (IV) show the process of learning connection between double-touching layer and self-occlusion one. It can be seen that it starts with multiple connections and finally succeeded in binding since it obtained the correct one-to-one mapping at the 4,000-th step.

Furthermore, we can see that the connections grew up both from the right and left ends to the center. It seems to show the process that cross-anchoring between a pair of nodes seems to make neighbor pairs of nodes more unique to each other and therefore guides cross-anchoring between the neighbor pairs. Such propagation of cross-anchoring starts from the pairs of nodes, either of which is a cue node. Consistently with the analysis of the learning procedure, the left end node in the bottom layer and the right end node in the top layer were cue nodes due to the morphological constraints. In this case, since the camera and the end-effector were connected through a serial link, how to double-touch and how to self-occlude were constrained. For example, the double-touching at the left end of the trunk could guide the self-occlusion only at the same part while the self-occlusion at the right end could be caused by the double-touching only at the same part.

Figs. 3 (b):(I) $\sim$ (IV) show the process of the learning connections in the case that the posture spaces of the camera head and the arm were pre-clustered in different resolutions. In this case, the resolution of the double-touching was twice in the case of self-

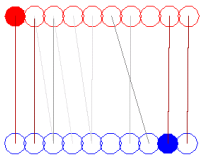
occlusion. The parameters were $N_t = 12$, $N_o = 6$, $\eta = 0.1, \eta_t = 0.5, \eta_o = 0.25$, and ${}^t\sigma_a = {}^o\sigma_a = 1.0$. Since it finally obtains the desired one-to-many mapping, we may conclude that it succeeds in binding despite the different resolutions.

After these processes, when the robot double-touches its body trunk, it can use the acquired mapping to know how to shift the focus of attention in the vision to the double-touching part by propagating the activation of the nodes responsible for the double-touching through the learned connection. Shortly, it can watch its touching part on its body.

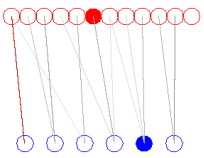


(II) 400 [step]

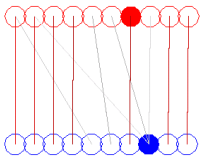
(II) 400 [step]



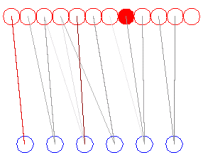
(III) 1000 [step]



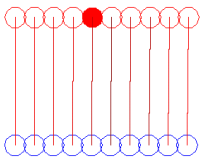
(III) 1000 [step]



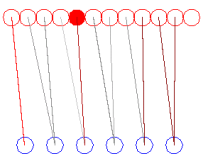
(IV) 2000 [step]



(IV) 2000 [step]



(V) 4000 [step]



(V) 6000 [step]

(a) one-to-one

(b) one-to-many

Figure 3: The process of learning connection between the layers ((I) $\sim$ (V)) with the same resolution (a) and with the different one (b).

4. Conclusion

In this paper, we addressed the issue how to solve the binding problem in different modalities for body representation. We proposed a method called cross-anchoring Hebbian learning rule to perform binding by virtue of the morphological constraints in self-body-observation. In the computer simulations, we showed that the robot can succeed in matching its foci of attention in vision and touch.

There are parameters in the proposed learning rule that determine how much the degree of anchoring is. Since it should be well selected to obtain the unique association, we should put a mechanism to adapt it when the system fail to bind. Topographical constraint caused by the the receptive fields with continuity that reflects the physical continuity could be a criteria for the adaptation. Furthermore, the robot needs the competence of binding in the case where it learns multimodal representation of the external objects. Although we concentrated on the binding problem concerning the self body in this paper, extending the proposed method for the binding problem involving tactile sensations of being touched by others is one of our future work.

Acknowledgment The Advanced and Innovational Research program in Life Sciences of the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government and a Research Fellowship for Young Scientists from Japan Society for the Promotion of Science supported this research.

References

MacDorman, K. F., Tatani, K., Miyazaki, Y., and Koeda, M. (2000). Proto-sysmbol emergence. In *Proc. of the Intl. Conf. on Intelligent Robot and Systems*, pages 1619–1625.

Seth, A., McKinstry, J., Edelman, G., and Krichmar, J. (2003). Visual binding, reentry, and neuronal synchrony in a physically situated brain-based device. In *Proc. of the 3rd Intl. Workshop on Epigenetic Robotics*, pages 177–178.

Tononi, G., Sporns, O., and Edelman, G. (1992). Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. *Cerebral Cortex*, 2:310–335.

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. *Neuron*, 24:105–110.

Yoshikawa, Y., Kawanishi, H., Asada, M., and Hosoda, K. (2002). Body scheme acquisition by cross modal map learning among tactile, image, and proprioceptive spaces. In *Proc. of Intl. Workshop on Epigenetic Robotics*, pages 181–184.