Imitation Learning Based on Visuo-Somatic Mapping

Minoru Asada*, Masaki Ogino, Shigeo Matsuyama, and Jun'ichiro Ooga

Adaptive Machine Systems and *HANDAI FRC, RoboCup Humanoid Robot Project Leader
Graduate School of Engineering
Osaka University, 2-1, Yamada-Oka, Suita, Osaka 565-0871, Japan
{asada, ogino, shigeo, ooga}@er.ams.eng.osaka-u.ac.jp
http://www.er.ams.eng.osaka-u.ac.jp/

Abstract. Imitation learning is a powerful approach to humanoid behavior generation, however, the most existing methods assume the availability of the information on the internal state of a demonstrator such as joint angles, while humans usually cannot directly access to imitate the observed behavior. This paper presents a method of imitation learning based on visuo-somatic mapping from observing the demonstrator's posture to reminding the self posture via mapping from the self motion observation to the self posture for both motion understanding and generation. First, various kinds of posture data of the observer are mapped onto *posture space* by self organizing mapping (hereafter, SOM), and the trajectories in the posture space are mapped onto a *motion segment space* by SOM again for data reduction. Second, optical flows caused by the demonstrator's motions or the self motions are mapped onto a *flow segment space* where parameterized flow data are connected with the corresponding motion segments in the motion segment space. The connection with the self motion is straightforward, and is easily acquired by Hebbian Learning. Then, the connection with the demonstrator's motion is automatic based on the learned connection. Finally, the visuo-somatic mapping is completed when the posture space (the observer: self) and image space (the demonstrator: other) are connected, which means observing the demonstrator's posture associcates the self posture. Experimental results with human motion data are shown and the discussion is given with future issues.

1 Introduction

Recent progress of humanoids such as ASIMO [9] and QRIO [6] has been attracting many people for their human-like behaviors towards symbiotic relationship between humans and robots. However, the current design and implementation of these behaviors mainly owes to the designers' deep knowledge and skills. In order to realize the truly symbiotic relationship, the robots are expected to be much more adaptive and flexible so that they can understand and generate human's various motions.

Imitation learning is one of the most powerful approach to humanoid behavior generation [11]. The most existing methods assume the availability of the information on the internal state of a demonstrator such as joint angles, which humans usually cannot directly access to imitate the observed behavior [3][7]. Some studies have been conducted without this assumption. Asada et al. [1] proposed a method for learning by observation based on the demonstrator's view recovery and adaptive visual servoing. Kuniyoshi et al. [5] showed the view-based imitation based on the similarity in the optical flows and on the association with motor commands.

These studies are intended not simply for efficient behavior generation, rather for understanding how humans learn to imitate the observed behaviors. However, the variety of the behaviors seems limited due to their simple flow matching methods.

This paper presents a method of imitation learning based on visuo-somatic mapping from observing the demonstrator's posture to recalling the self posture. First, various kinds of posture data of the observer are mapped onto a *posture space* by self organizing mapping [4] (hereafter, SOM), and the trajectories in the posture space are mapped onto a *motion segment space* by SOM again for data reduction. Second, optical flows caused by the demonstrator's motions or the self motions are mapped onto a *flow segment space* where parameterized flow data are connected with the corresponding motion segments in the motion segment space. The connection with the self motion is straightforward, and is easily acquired by Hebbian Learning. Then, the connection with the demonstrator's motion is automatic based on the learned connection. Finally, the visuo-somatic mapping is completed when the posture space (the observer: self) and image space (the demonstrator: other) are connected, which means observing the demonstrator's posture associates the self posture. Experimental results with human motion data are shown and the discussion is given with future issues.

2 A System Overview

2.1 Basic assumptions

Here, we assume the followings to realize the visual imitation based on the visuo-somatic mapping:

1. No *a priori* knowledge on the link structure, that is, connections between joints.
2. No *a priori* knowledge on the body part (joint) correspondence between the demonstrator and the observer.
3. Both the demonstrator's and the self motions can be observed in terms of a temporal sequence of joint vectors in image space.
4. The joint angles of the self posture can be observed, but no relationship between the self posture and the visual feature space is given.
5. Currently, we focus on the mirror image imitation. This means the right (left) side of the demonstration corresponding to the left (right) side side of the observation.

2.2 Imitation System

Fig. 1 shows the proposed system, consisting of two sub-processing systems, *Visual Information processing system* and *Somatic Information processing system*. In these processing sub-systems, row sensory data are mapped onto the corresponding two dimensional Self-Organizing Maps (SOMs) [4]. The images observing the demonstrator's motion are first mapped onto an *image space* which includes the posture image of the demonstrator, and then *flow segment space* in which the changes in posture are represented. The flow segment space is also utilized to represent the self motion, too. On the other hand, the self somatic sensory data are mapped onto a *posture space* and their changes are mapped onto a *motion segment space*. After generation of these maps independently, the flow segment space and the motion segment space are connected based on Hebbian learning.

The connection between the flow segment space and the motion segment space is easily carried out by using Hebbian learning based on the simultaneous activations of segments in both spaces during the self motions. Once this connection is acquired, the connection between the flow segment space for the demonstrator's motion and the motion segment space is automatic based on the learned connection between the flow segment space and the motion segment space. Through these connections, the mapping from the image space of the demonstrator's posture to the self posture space is enabled, that is, visuo-somatic mapping can be obtained.

In the followings, the details of each sub processing system are explained in section 3, and the mapping among them are shown in section 4.

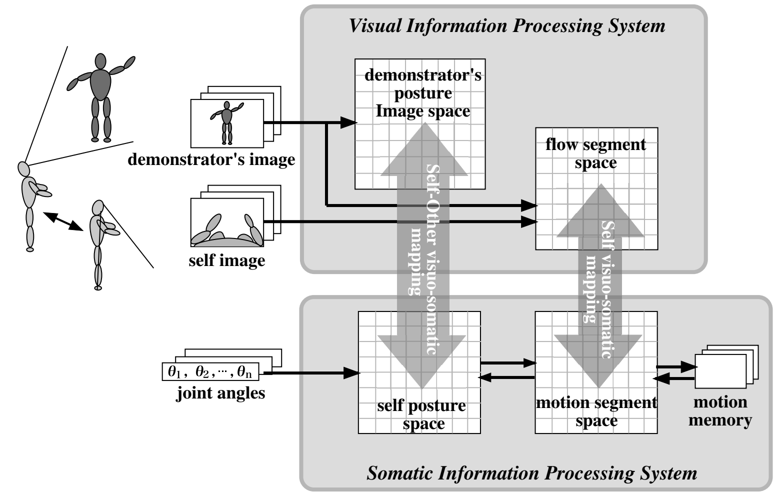
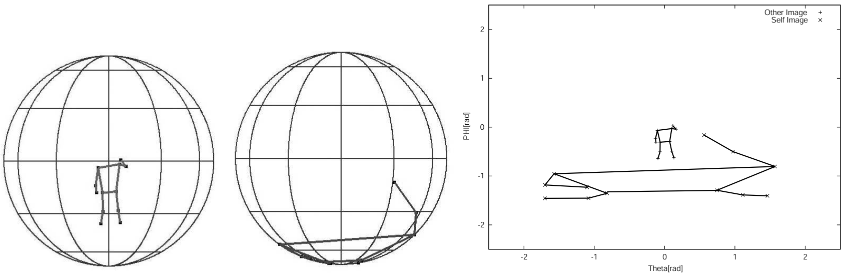

Visual Information Processing System
demonstrator's posture Image space
flow segment space
demonstrator's image
self image
Self-Other visuo-somatic mapping
Self visuo-somatic mapping
$\theta_1, \theta_2, \cdots, \theta_n$
joint angles
self posture space
motion segment space
motion memory
Somatic Information Processing System

Fig. 1. System overview

2.3 Sensory Data

We prepare the sensory data by using human motions acquired by a motion capturing system. The captured data, which are three dimensional data sets in the global coordinate system, are converted to the two dimensional data on a virtual camera images captured by the observer (self). The angles between links in a human model are also calculated to be used as the self posture data. A joint angle vector consisting 16 joint angles is mapped onto the self posture space and the segmented trajectories on the map are mapped on motion segment space. The spherical image projection from the camera position at the observer head is assumed to capture the whole self body image. Fig. 2 shows examples of the self body image (a) and the demonstrator's one (b) on the spheres, and their development onto a plane (c).

Twelve kinds of motions are captured from the human motion performances. They are combinations of motion, side, and part such as "raise," "wave," and "rotate" as motions, "left," "right," or "both" as sides, and "hand," and "knee" as parts. Also a "walking" motion is added as a whole body motion.



(a) Image of a demonstrator

(b) Image of Self body

(c) The development of the spherical projection

Fig. 2. Image data to be input to the system

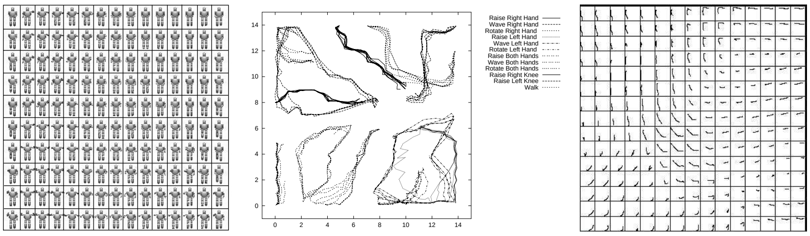
3 Construction of SOMs for behavior recognition and generation

3.1 Posture space and motion segment space

We construct a *posture space* SOM from the *somatic information*, which is a sequence of the vectors consisting of sixteen joint angles between links calculated from the captured motion data. The size of *posture space* SOM is 15 × 15, and it is constructed by 240 [frames] (8 [sec]) per each motion. Fig. 3 (a) shows the resultant SOM, where 15 × 15 postures are representative ones from row data.

Since the posture data are input sequentially, we can visualize how posture data are connected each other in the posture space. Fig. 3 (b) shows such data indicating that the trajectries of motions are roughly segmented and construct the clusters corresponding to performed actions.

These trajectories are divided into small segments, each of which consists of 10 [frames] temporal sequence of posture data, and are clustered into another SOM, *motion segment space*, (Fig. 3 (c)).



(a) Posture space

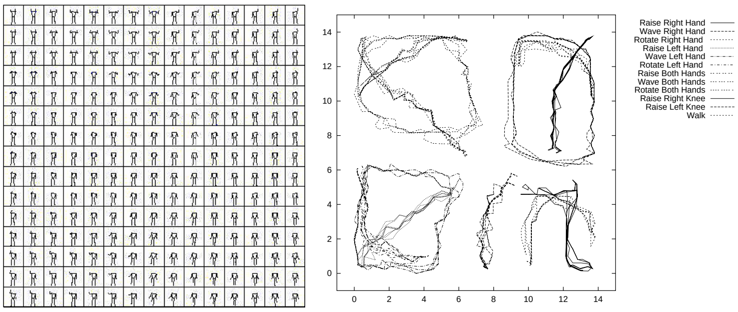
(b) Trajectories of motions on posture space

(c) Motion Segment Space

Fig. 3. The SOMs constructed in the self somatic information processing system

3.2 Demonstrator's posture image space

A demonstrator's posture image space (hereafter, image space in short) consists of the representative image position vectors obtained by self organizing mapping of image positions of joints of the human model. Fig. 4 (a) shows the image space where various postures are clustered into 15 $\times$ 15 representative postures. Similar to the posture space based on the somatic information, we can visualize how posture image data are connected each other in this space. Fig. 4 (b) shows such data which indicate that the trajectories of motions are roughly segmented and construct the clusters corresponding to performed actions.



(a) *Demonstrator's posture image space*

(b) Trajectories of motions on demonstrators posture image space

Fig. 4. Demonstrator's posture image space

3.3 Flow segment space construction

The problem here is how to associate the observed flow caused by the demonstrator's motion with the self motion. If the flows by the demonstrator's motions are similar to the flows by the self motions, the desired association seems easy to find because the connection between the observed self motion and the self motion segment can be easily found based on the simultaneous activations during the self motion. However, it would not be so due to the viewpoint difference. Then, as the common features of the flow segments, we chose the direction and the relative position of the flow segments. The directions of flow segments by the demonstrator (other) and the observer (self) are very similar to each other although there are slight differences in the directional changes between them. The relative positions are quantized into four regions (top left, top right, bottom left, and bottom right) by setting the centroid of posture image vectors as the origin. These four regions are called attention areas. The positions of joints are similar to each other between the demonstrator's and the observer's in spite of large shape difference.

By using the quantized directions and the normalized magnitudes of the flows, and the attention area, the flow segment space is constructed. A data structure for the flow segment space is shown in Fig. 5 (a).

Flow segment space The directions of the flows are segmented when the sign of horizontal or vertical element of flow vector is inverted. In each segment, the directions are averaged. Suppose the time when n-th flow inversion happens is T_n, then the averaged flow direction is given by

$$\S_i^{\phi}(t) = \frac{1}{T_{n+1} - T_n} \int_{T_n}^{T_{n+1}} \phi_i^F(s) ds \quad (T_n < t < T_{n+1}), \tag{1}$$

where $\phi_i^F(t)$ indicates the flow direction of body segment i at time t. The averaged direction data are sorted by their length of the flow vectors. And the i-th data is assigned to the i-th layer in *flow segment space*. In each layer, the unit that has the nearest direction to the input data is activated.

Attention area space Although the positions of flow vectors in the robot's view are quite different between the self and the demonstrator, the relative positions among them (upper right, upper left, lower right and lower left) are roughly maintainted well. Using this feature, the *attentional area space* describes what part of the self image includes first flow vectors.

Let N_f the number of observed points of the self and the demonstrator's body and the regions around the center of observed points R_1, R_2, R_3 and R_4, then the total flow speed included in each region R_i is given by

$$F_j(t) = \int_{i=1}^{N_f} p_i(t) ||v_i(t)||, \tag{2}$$

$$p_i(t) = \begin{cases} 1\ if & u_i(n) \in R_j \\ 0 & else \end{cases} \quad (j = 1, \cdots 4), \tag{3}$$

where $v_i(t)$ is the observed flow vector, and $u_i(t)$ is the position vector of observed point i at time t. Note that i does not correspond to the labeled point of the body. We define the relative total strength of flow among regions as

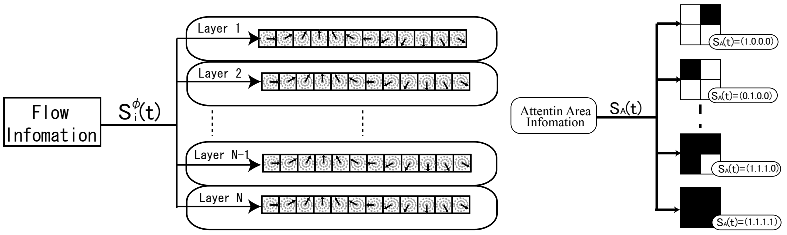
$$A_j(t) = \frac{F_j(t)}{\sum_{n=1}^{4} F_n(t)} \quad (j = 1, \cdots 4). \tag{4}$$

The input vector to *attentional area space*, $S_A(t)$, consists of binarized $A_j(t)$,

$$S_A(t) = (A_1^S(t), A_2^S(t), A_3^S(t), A_4^S(t)) \tag{5}$$

$$A_j^S(t) = \begin{cases} 1, if & A_j(t) \geq 0.20 \\ 0, & else \end{cases} \tag{6}$$

Attention area space consists of all the combinations of activated areas, $2^4 = 16$.



(a) Flow direction

(b) Attention area

Fig. 5. Flow segment space

4 Mapping between visual and somatic field

4.1 Self Visual-Somatic sensation Mapping

The simultaneous activations of the units in the flow segment space and the self posture space during self motion make it possible to find correspondence between the units in those spaces.The connection coefficients between the units in each space are learned based on Hebbian learning. All the connection coefficients are initialized to 0s, and during the self motion the coefficient, w^{AB}, which is the connection coefficient betwen the i-th unit in space A and the j-th unit in space B, is updated during self motion, as follows,

$$w_{ij}^{AB}(t+1) = w_{ij}^{AB}(t) + \varepsilon(y_i^A(t) y_j^B(t) - y_i^A(t)^2 w_{ij}^{AB}(t)). \tag{7}$$

At the same time, the time sequences of the activated units in motion segment space during various motions are memorized as the motion modules in *motion memory*.

4.2 Recognition of other person's motion

After acquisition of self visual-somatic mapping, the input of image data observing a demonstrator's motion activates the units in *motion segment space* through *visual information processing system* via connections between them. Let the activation level of the i-th unit in flow direction $y_i^F(t)$ and that of the j-th unit in attention area $y_j^A(t)$, the activation level of the k-th unit in motion segment space, $y_k^M(t)$ is given by

$$y_k^M(t) = \sum_{i=1}^{N_F} w_{ik}^{FM} y_i^F + \sum_{j=1}^{N_A} w_{jk}^{AM} y_j^A \tag{8}$$

The quantization in the flow segment space is coarse and the mapping between the flow segment space and the motion segment space is not one-to-one mapping. The motion of a demonstrator activates multiple units in the motion segment space at a time, which makes it difficult to identify the corresponding motion module. So, we compare the temporal sequences of activated units of observed motion with those of memorized motion modules in the motion segment space. To do that, we define the evaluation function, E_m, which indicates the similarity of the time sequence of acvtiated units of an observed motion to that of the m-th memorized motion as follows,

$$E_m = \max_{s_t} \int_0^T \sum_{i=1}^{N_M} y_i(t) m_i(t_s + t) dt. \tag{9}$$

Thus, the observed motion is recognized as the same as the motion module that maximizes E_m.

4.3 Mapping between a self posture image space and a demonstrator's posture image

Recalling the self motion from the observation of a demonstrator's motion makes it possible to correlate the demonstrator's posture image space in visual information processing system with the self posture space in somatic sensation information processing system. When observing a demonstrator's motion, the unit in image space and the unit in posture space activate simultaneously. So we can use Hebbian learning again between these two maps.

Fig. 6 shows the recalled posture (the rightmost figure) from the observed image (the leftmost figure). The two maps in the middle of the figures describe the activated units in image space(left) and posture space (right) after Hebbian learning.

5 Discussion

In this paper we proposed a learning system for imitation based on visuo-somatic mapping. This system excludes the pre-designed model of the demonstrator as much

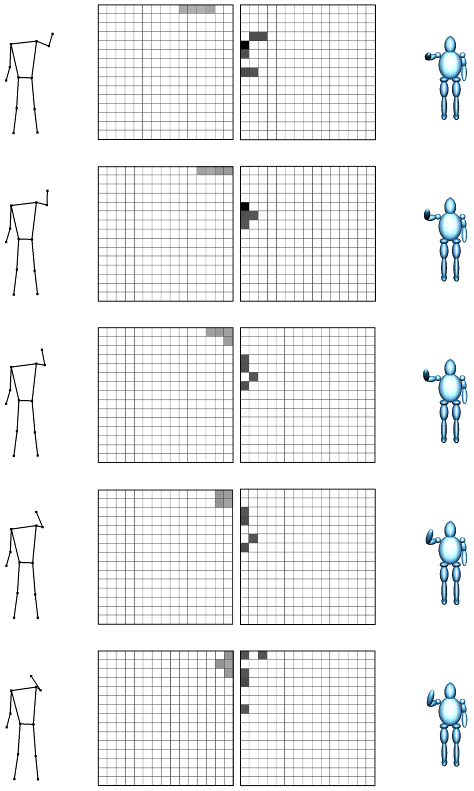


Fig. 6. Recalling somatic sensation by Self-Other visuo-somatic sensation mapping (The left two figures show demonstrator's image and its mapping on a demonstrator's posture image space. The right two figures show the correponding representative vectors in the self posture space and the corresponding self postures.

as possible. The demonstrator's model is made through demonstrator's images in a demonstrator's posture image space. The model of self is not pre-designed, either. It is constructed by self-organizing the self motion information in self posture space and motion segment space. The primitive visual features are related to the representative vectors in motion segment space during self motion. This connection induces the self motion when observing demonstrator's motions and further mapping between demonstrator's posture image and self posture space is made. After constructing the visuo-somatic mapping, this system can directly activate the self posture corresponding the observed demonstrator's image.

Although initial aims to construct the visuo-somatic mapping through learning are accomplished in this system, it has many problems for practical use as an imitation system. First this system assumes that an observer always stands face to face with a demonstrator, and this sytem does not have concept about the translation or rotation to the ground of the demonstrator. An observer can recognize only jestures of the demonstrator. Second, the resultant visuo-somatic mapping is not so accurate as to make new motion modules only from observation of demonstrator's motion, because the sequence of the activated postures is not smooth.

For the first problem, we are now extending our model so that it can describe the transition and rotation of a demonstrator relative to the ground. The second problem can be solved by using velocity information acquired by another pathway. Acquiring new motions which are not experienced through observation is the next challenge for us.

Acknowledgments This study was performed through the Advanced and Innovational Research program in Life Sciences from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government.

References

1. M. Asada, Y. Yoshikawa, and K. Hosoda, "Learning by observation without three-dimensional reconstruction," In *Proc. of the 6th Int. Conf. on Intelligent Autonomous Systems (IAS-6),* pp. 555–560, 2000.
2. A. Billard, M. Mataric: "Learning human arm movements by imitation: Evaluation of a biologically inspired connectionist architecture", *Robotics and Autonomous Systems*, 37, pp.145–160, 2001.
3. T. Inamura, I. Toshima, and Y. Nakamura. "Acquisition and embodiment of motion elements in closed mimesis loop," In *Proc. of IEEE Int. Conf. on Robotics and Automation*, pp. 1539–1544, 2002.
4. T. Kohonen: *The Self Organization and Associative Memory,* Springer-Verlag, 1989.
5. Y. Kuniyoshi, Y. Yorozu, M. Inaba, and H. Inoue, "From visuo-motor self learning to early imitation – A neural architecture for humanoid learning," In *Proc. of the 2003 IEEE Int. Conf. on Robotics and Automation,* pp. 3132–3139, 2003.
6. Y. Kuroki, T. Ishida, and J. Yamaguchi, A Small Biped Entertainment Robot, In *Proc. of IEEE-RAS Int. Conf. on Humanoid Robot*, pp. 181-186, 2001.
7. A. Nakazawa, S. Nakaoka, K. Yokoi, T. Harada, and K. Ikeuchi: "Imitating human dance motion through motion structure analysis", In *Proc. of 2002 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pp. 2539–2544, 2002.
8. E. Oja, J. Karhunen: "On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix", *J. Math. Anal. and Appl.*, 106, pp. 69–84, 1985.
9. M. Hirose, Y. Haikawa, T. Takenaka, and K. Hirai, Development of Humanoid Robot ASIMO, In *Proc. Int. Conf. on Intelligent Robots and Systems*, 2001.
10. R. Pfefer and C. Scheier.: *Understanding Intelligence*, MIT Press, Cambrridge, Massachusetts, 1999.
11. S. Schaal., "Is imitation learning the route to humanoid robots?," *Trends in Cognitive Science,* pp. 233–242, 1999.