View-based Imitation with Rotation Invariant Pan-Tilt Stereo Cameras

Yuichiro Yoshikawa*, Yoshiki Tsuji*, Minoru Asada*†, and Koh Hosoda*†

*Dept. of Adaptive Machine Systems, †HANDAI Frontier Research Center,
Graduate School of Engineering, Osaka University.
{yoshikawa,tsuji,asada,hosoda}@er.ams.eng.osaka-u.ac.jp

Abstract

In the previous work, we have developed a method for visual imitation by recovering the demonstrator's view based on the stereo epipolar constraint [1]. The method is applied to the stationary pair of the stereo cameras, therefore, the visual fields to observe the both motions of the demonstrator and the learner are limited. This paper presents a method to extend our previous work by adopting a pair of rotation invariant stereo cameras that has pan and tilt motions without changing the optical center, therefore, the stereo epipolar equation does not change. The spherical projection is used to represent the constraint. The experimental results are shown.

1 Introduction

Imitation is one of the most important capability for an intelligent robot to perform a variety of complicated tasks in the real world because learning by imitation is regarded as a promising way to accelerate the learning of a robot which has different sensory modalities and many degrees of freedoms such as a humanoid robot [2, 3]. Another aspect of the imitation capability is that it is also one of the most interesting cognitive issues to model how we human beings learn to acquire various kinds of behaviors by building real robots capable of imitation learning [4].

Most of the existing robotic approaches assume an observation capability by which the robot knows the demonstrator's internal states such as joint angles since they have focused on how to encode the sequence of them. The assumptions are held by the module of behavior recognition [5], a motion capture system [6, 7], a sensor-suit attached on the demonstrator's body [8], the coordinate transformation [9], and so on. However, these solutions seem unnatural for a real, autonomous robot because they need the calibration process by the designer. Instead, it is an interesting issue how to acquire such a capability from its sensory information by itself.

Asada et al. [1] proposed one of the view-based imitation methods which consists of two parts, the view transformation to recover the demonstrator's view and the adaptive visual servoing [10] to follow the recovered trajectory of the demonstration. However, the learner is not allowed to move its camera when it observes the demonstration or its body, because it utilizes opt-geometric constraint between image planes, called *epipolar geometry* [11]. Therefore, in order to perform imitation, the both motions of the demonstrator and the learner are limited.

This paper presents a method to cope with the case that it needs to move its cameras to capture the both motions of the demonstrator and the learner. In order to begin with easier case of imitation, we assume that the both body structures are the same as well as the previous work [1]. We adopt a pair of the rotation invariant stereo cameras each of which has pan and tilt motions without changing its optical center since epipolar geometry between them is invariant if their optical center are fixed. We call this property rotation invariant. In this case, we can apply the spherical projection to the invariant representation of epipolar geometry for such rotations. According to epipolar geometry, we derive the method how to find the spherical image coordinates in the virtual view, called the view transformation.

The rest of paper is organized as follow: first a mechanism of the view transformation is described, and the method of demonstrator's view recovery is given. In order to confirm the validity of the method, at first, the computer simulation is given. Then, the camera system to hold the assumption of optical center invariance is built and the real robot experiment is shown.

2 View transformation among image spheres

Suppose that the attentional point in the three dimensional space projects onto two image spheres and their coordinates are given. When we add one more image sphere, we propose a method of the view transformation by which we find the coordinate of the matched point in the newly added image sphere in this section. In order to allow the extension of the visible region by rotating the cameras, we assume the rotation invariant motion and adopt spherical images.

2.1 Epipolar geometry

Image sphere is a spherical image surface of which center coincides with the optical center, and *sphere projection* is the projection where the attentional point in the three dimensional space is projected onto the intersection between the image sphere and a line which contains it and the optical center. Here, the camera coordinate system of a spherical projection is the three dimensional coordinate system fixed on the optical center.

When a 3D point is projected on two image spheres, the point, the projected 3D points, and the optical centers are on one plane, called *epipolar plane* (see Fig. 1). The opt-geometric constraint is called *epipolar geometry*, and it is described in the epipolar equation, such as,

$$ {}^{l}\boldsymbol{x}^{T\,lr}\boldsymbol{E}^{r}\boldsymbol{x} = 0, \qquad (1)$$

where ${}^{l}\boldsymbol{x}$ and ${}^{r}\boldsymbol{x}$ are the coordinates of the projected points in the image sphere $[l]$ and $[r]$, respectively. ${}^{lr}\boldsymbol{E} \in \Re^{3\times 3}$ is called *essential matrix* determined by the geometrical relationship between two camera coordinate systems.

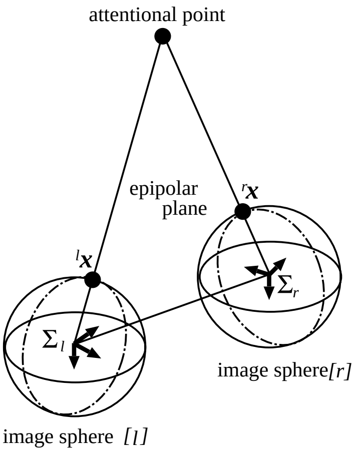

Figure 1: *Epipolar geometry between two image spheres.*

If we have more than eight coordinates of the matched projected points, the essential matrix can be estimated by solving the simultaneous equation of eq. (1) based on the least square method [11].

2.2 View translation based on epipolar geometry

We add one more camera $[L_D]$ ($[R_D]$) observing a point which is also observed in $[l]$ and $[r]$ (see Fig. 2). The problem is how to find the corresponding points in the view $[L_D]$ ($[R_D]$) with ones in the views $[l]$ and $[r]$.

Based on the epipolar geometry, the matched points ${}^{L_D}\boldsymbol{x}$ (${}^{R_D}\boldsymbol{x}$) for ${}^{l}\boldsymbol{x}$ is constrained to lie on the epipolar plane, and simultaneously on the image sphere $[L_D]$ ($[R_D]$). Therefore, it is constrained to the great circle on $[L_D]$ ($[R_D]$) which is the intersection of them. Since it is also constrained on another great circle which is derived from epipolar geometry between $[r]$ and $[L_D]$ ($[R_D]$), it is determined by finding the intersection of two great circles.

We can determine it by solving following simultaneous equations, such as

$$ {}^{l}\boldsymbol{x}^{T\,lL_D}\boldsymbol{E}^{L_D}\boldsymbol{x} = 0, \qquad (2)$$
$$ {}^{r}\boldsymbol{x}^{T\,rL_D}\boldsymbol{E}^{L_D}\boldsymbol{x} = 0, \qquad (3)$$
$$ {}^{L_D}\boldsymbol{x}^{T\,L_D}\boldsymbol{x} = c^2, \qquad (4)$$

where the first and the second ones represent epipolar planes on which ${}^{L_D}\boldsymbol{x}$ is constrained, the last one represents the condition that ${}^{L_D}\boldsymbol{x}$ is on the image sphere, and c is arbitrary positive value which indicates radius of the image sphere. ${}^{R_D}\boldsymbol{x}$ can also be determined in a similar manner.

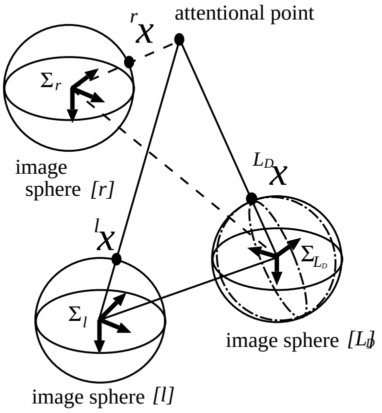

Figure 2: *The mechanism of the view transformation based on epipolar geometry.*

Summing up, ${}^{L_D}\boldsymbol{x}$ (${}^{R_D}\boldsymbol{x}$) is determined by the function, such as

$$ {}^{L_D}\boldsymbol{x} = \boldsymbol{f}({}^{l}\boldsymbol{x}, {}^{r}\boldsymbol{x}, {}^{lL_D}\boldsymbol{E}, {}^{rL_D}\boldsymbol{E}) \qquad (5)$$
$$ ({}^{R_D}\boldsymbol{x} = \boldsymbol{f}({}^{l}\boldsymbol{x}, {}^{r}\boldsymbol{x}, {}^{lR_D}\boldsymbol{E}, {}^{rR_D}\boldsymbol{E})). \qquad (6)$$

It means that the matched point in the unknown image sphere can be determined if the projected points in the known image spheres and the essential matrices of epipolar geometry between them are given.

3 Demonstrator's view recovery [1]

In this section, the method of recovering the demonstrator's view to imitate its motion using the view transformation is given.

In order to avoid a difficult issue on the definition of imitation, we deal with the case that the learner and the demonstrator have the same body structure because such an assumption gives us a simple definition of imitation such as reproducing the same trajectories of the body parts in the three dimensional space. Practically, we assume that the link parameters and the camera one are the same. In such a case, if the learner knows the trajectories of the demonstrator's body parts in the demonstrator's view, it can perform imitation by realizing the same trajectory by its matched body parts.

Fig. 3 shows the relationship between the views of the demonstrator and the learner, where V_O^A indicates the view of agent O observing the motion of acting agent A (A, O: the demonstrator or the learner). The learner needs to recover the demonstrator's views, V_D^D ($[L_D]$ and $[R_D]$) by which the demonstrator is supposed to observe itself during the demonstration, from observed information in the learner's views, V_L^D ($[L_D]$ and $[R_D]$). If essential matrices of epipolar geometry between V_L^D and V_D^D are given, the learner can find the trajectories of the demonstrator's body parts in V_D^D by the view transformation explained in the previous section. After finding them, the learner can perform imitation through the control to follow them by the adaptive visual servoing [10]. However, how to know the essential matrices is still a question.

3.1 Estimation of the essential matrices

The learner needs to estimate the essential matrices, ${}^{lL_D}\boldsymbol{E}$, ${}^{rL_D}\boldsymbol{E}$,${}^{lR_D}\boldsymbol{E}$, ${}^{rR_D}\boldsymbol{E}$, of epipolar geometry between V_L^D and V_D^D. As mentioned in the previous section, more than eight pairs of the corresponding projected points are needed to estimate the essential matrices [11]. However, the learner does not know directly the corresponding point in V_D^D.

Suppose that the demonstrator's initial posture (joint angles) is the same as the learner's one. Since it is assumed that the learner has the same body structure and the same camera parameters as the demonstrator does, the learner's body parts projected on V_L^L is the completely same as the demonstrator's one in V_D^D (see Fig. 4). Therefore, instead of using points in unobservable view V_D^D, the essential matrices can be estimated by using observed points which are considered to correspond ones in V_L^L.

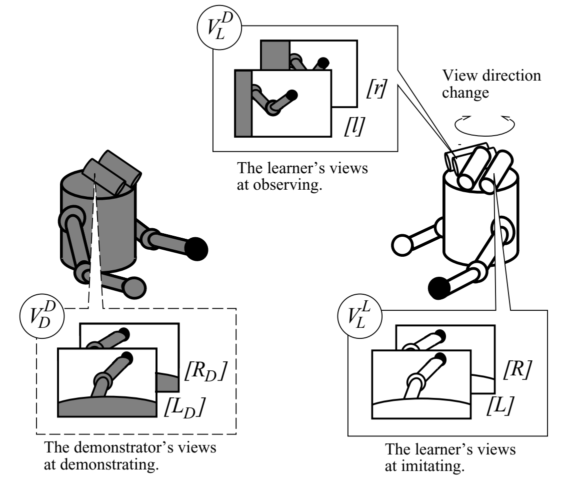
V_L^D
[r]
[l]
View direction change
The learner's views at observing.
V_D^D
$[R_D]$
$[L_D]$
The demonstrator's views at demonstrating.
V_L^L
[R]
[L]
The learner's views at imitating.

Figure 3: *Demonstrator's view recovery.*

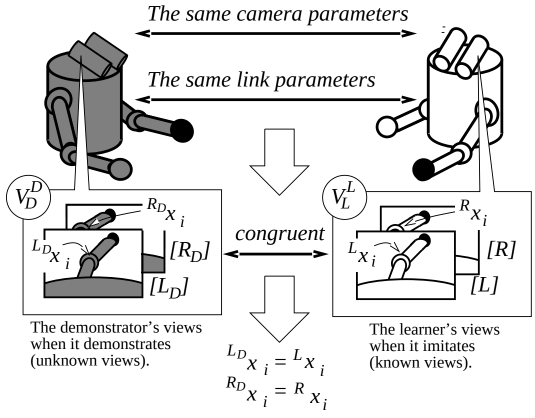
The same camera parameters
The same link parameters
V_D^D
${}^{R_D}x_i$
${}^{L_D}x_i$
$[R_D]$
$[L_D]$
congruent
V_L^L
${}^{R}x_i$
${}^{L}x_i$
[R]
[L]
The demonstrator's views when it demonstrates (unknown views).
The learner's views when it imitates (known views).
${}^{L_D}x_i = {}^{L}x_i$
${}^{R_D}x_i = {}^{R}x_i$

Figure 4: *The key idea of the method of the parameter estimation.*

When we define ${}^{L}\boldsymbol{x}_i$ and ${}^{R}\boldsymbol{x}_i$ as the i-th projected point on the learner's body in V_L^L, as well as ${}^{L_D}\boldsymbol{x}_i$ and ${}^{R_D}\boldsymbol{x}_i$ as the corresponding one on the demonstrator's body in V_D^D, they satisfy the following equations, such as

$$\begin{aligned} p^{L_D}\boldsymbol{x}_i &= {}^{L}\boldsymbol{x}_i, \\ {}^{R_D}\boldsymbol{x}_i &= {}^{R}\boldsymbol{x}_i. \end{aligned} \qquad (7)$$

It means that the learner can use ${}^{L}\boldsymbol{x}_i$ and ${}^{R}\boldsymbol{x}_i$ as alternativeness of ${}^{L_D}\boldsymbol{x}_i$ and ${}^{R_D}\boldsymbol{x}_i$ to estimate the essential matrices.

In order to release the assumption that both postures are the same, the method to estimate the essential matrices by the control of its joint angles to minimize the error of view transformation using currently estimated essential matrices has been proposed [12] although it is a method when the learner uses the no-spherical image planes.

4 Experiment using the computer simulation

In order to confirm the validity of the view transformation, experimental result in the computer simulation is shown in this section.

We create the four image spheres, $[l]$,$[r]$,$[L_D]$, and $[R_D]$, in the computer. Setting 64 points as the attentional point in the three dimensional space, we calculate the projected points of them in each image sphere. The residuals in the estimation of the essential matrices are shown (see Tab. 1).

Table 1: *The residuals in the estimation of the essential matrices in the computer simulation.*

essential matrices	residual of estimation
${}^{lL_D}\boldsymbol{E}$	1.8×10^{-19}
${}^{rL_D}\boldsymbol{E}$	2.6×10^{-19}
${}^{lR_D}\boldsymbol{E}$	4.1×10^{-19}
${}^{rR_D}\boldsymbol{E}$	5.8×10^{-19}

The matched points in $[L_D]$ and $[R_D]$ with ones in $[l]$ and $[r]$ is recovered by the view transformation using the estimated essential matrices (see Fig. 5). Fig. 5 shows the recovered points (dots) as well as the true ones (cubes) in the image sphere $[R_D]$. Since they almost coincide with each other, the matched points can be recovered by the view transformation.

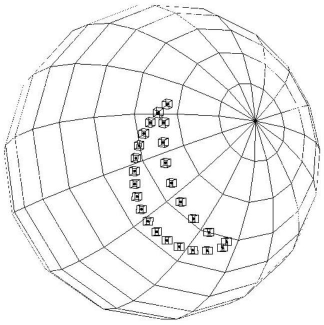

Figure 5: *The result of the view transformation.*

5 Experiment using the real robot

In order to confirm whether the demonstrator's view can be recovered by the proposed method, the experimental result using the real robot is shown in this section. In the experiment, two identical manipulators are supposed to be the learner's and the demonstrator's body (see Fig. 6).

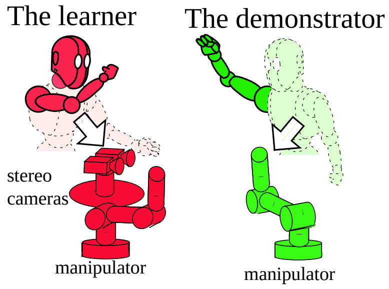

Figure 6: *The learner and the demonstrator have identical manipulators as their bodies.*

5.1 Rotation invariant camera head

We use a pair of stereo camera heads each of which consists of three helical gears to realize camera rotation in which the optical center is invariant (see Fig. 7). First, the torques of the two motors are transmitted to the upper helical gears which have the same radius by the belts. Then, the lower helical gear as the stage of the CCD camera is driven. The rotation center of the stage is the intersection of the rotational axes of three helical gears. It is designed so as to coincide the optical center of CCD camera with the rotation center. The rotation angles (*pan* and *tilt*) are calculated by following function, such as

$$\begin{aligned} pan &= r/R \cdot (\theta_1 - \theta_2), && (8) \\ tilt &= \theta_1 + \theta_2, && (9) \end{aligned}$$

where θ_1 and θ_2 are the angles of the upper helical gears, and r and R are radius of upper ones and lower one, respectively.

Although it is designed to have the rotation invariance of the optical center, it is not guaranteed because the CCD camera is attached to the stage by hand.

5.2 Experimental setup

The detail of the experimental setup is shown in Fig. 8. We use two identical 7 DOFs manipulators (PA10, MHI) as the bodies of the learner and the demonstrator, and two sets of pan-tilt camera heads mentioned above as the learner's. In this experiment, we use

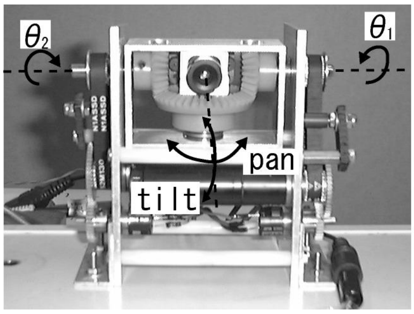

Figure 7: *The pan-tilt camera.*

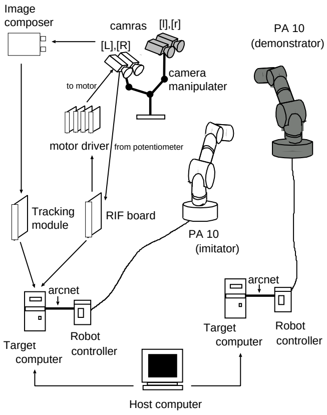

Figure 8: *Experimental setup*

the three axes of the manipulator and the remaining axes are fixed.

The image streams from two CCD cameras are combined to one (image size: 640[pixel] × 480[pixel]) by compressing each images into the half along the vertical axis (640[pixel] × 240[pixel]) in the field multiplexer, and then sent to a tracking module equipped with a high-speed correlation processor based on SAD (Sum of Absolute Difference) manufactured by Fujitsu. Before starting an experiment, we specify target images to be tracked by the module. The CPU (VxWorks, WindRiver) calculates and outputs control signals to the each controller of the manipulator and the camera heads.

5.3 Projection to image sphere

In order to utilize the image sphere, the learner should have a mapping from the image plane to it. Since the optical center coincides with the rotation center, the angles (*pan* and *tilt*) of the camera stage correspond to the angles of the polar coordinate system. Therefore, it can acquire the mapping through the experience of focusing the attentional point in the image plane, that is rotating the camera stage so as to capture it at the center of the image plane.

We gave 100 feature points in the three dimensional space and let the learner gaze them. Using the data in the focusing experiences, a feed-forward neural network which has one hidden layer with three units learns the mapping from the coordinate of the projected features in the image plane to the angles of the camera stage by the backpropagation method. The learning curve of the mapping is shown in Fig. 9, where it is confirmed that the learner acquire the mapping since the squared error is sufficiently small.

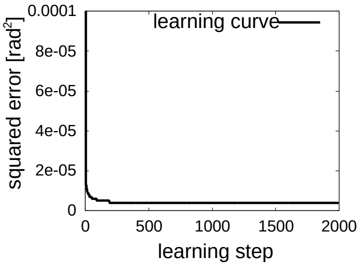

Figure 9: *Learning curve: squared error of the mapping from image plane to image sphere.*

5.4 Recovering the demonstration observed in the demonstrator's view

In order to confirm the validity of the method to imitate, the demonstrator shows the motion in which a triangle is drawn by its end-effector, and the learner recovers the trajectory observed by the demonstrator in the learner's view. Given 79 points in the three dimensional space, the residuals in the estimation of the essential matrices are shown (see Tab. 2).

Table 2: *The residuals in the estimation of the essential matrices in the real robot experiment.*

essential matrices	residual of estimation
${}^{lL_D}\boldsymbol{E}$	1.1×10^{-3}
${}^{rL_D}\boldsymbol{E}$	8.2×10^{-4}
${}^{lR_D}\boldsymbol{E}$	1.0×10^{-3}
${}^{rR_D}\boldsymbol{E}$	5.8×10^{-4}

Then, the result of recovering the trajectory of the demonstrator's end-effector is shown in Fig. 10, where the recovered points (dots) as well as the true ones (cubes) in the image sphere $[R_D]$ are shown. Since they almost coincide with each other, the matched points are found by the view transformation, that is the demonstrator's view is successfully recovered.

Although it is almost close, it has some errors because of the residuals in the estimation of the essential matrices. It may be caused by the fact that the rotation invariance is not guaranteed in the current camera head. We plan to provide the camera head with the mechanism by which we can adjust the position of the CCD camera and expect to show the result with less error at the conference.

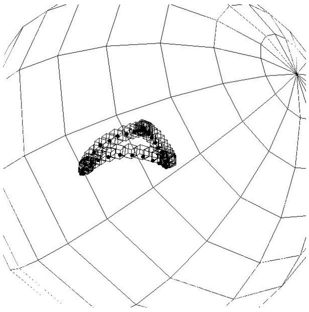

Figure 10: *The trajectory recovered by the view transformation and the true trajectory measured by the designer in advance.*

6 Conclusion

This paper proposed the extension of the method to imitate from observation based on demonstrator's view recovery. Assuming that the optical centers of the learner's stereo cameras are rotation invariant, the demonstrator's view is recovered by the view transformation based on epipolar geometry with the learner's one. Therefore, it allows the learner to move its cameras to extend its visible regions unlike the previous method.

The validity of the proposed method is confirmed by the experimental results in the computer simulation and the real robot experiment. In order to reduce the recovering errors caused by the residual in the estimation, we plan to provide the camera head with a mechanism to adjust.

Although we assume that the demonstrator's posture is the same as the learner to estimate the parameters of epipolar geometry, it should cope with the situation when they are different. Combining the idea to estimate in such a situation [12] is one of our future work.

Acknowledgments

This research was 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] M. Asada, Y. Yoshikawa, and K. Hosoda. Learning by obsevation without three-dimensinal reconstruction. In *Proceedings of the 6th International Conference on Intelligent Autonomous Systems*, pp. 555–560, 2000.

[2] J. Demiris and G. Hayes. Imitative learning mechanisms in robots and humans. In *Proceedings of the 5th European Workshop on Learning Robots*, pp. 9–16, 1996.

[3] S. Schaal. Is imitation learning the route to humanoid robots? *Trends in Cognitive Science*, Vol. 3, No. 6, pp. 233–242, 1999.

[4] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotic and Autonomous System*, Vol. 37, pp. 185–193, 2001.

[5] Y. Kuniyoshi, M. Inaba, and H. Inoue. Learning by watching: Extracting reusable task knowledge from visual observation of human performance. *IEEE Trans. on R & A*, Vol. 10, No. 6, pp. 799–821, 1994.

[6] T. Inamura, Y. Nakamura, H. Ezaki, and I. Toshima. Imitation and primitive symbol acuisition of humanoids by the integrated mimesis loop. In *Proceedings of the 2001 IEEE International Conference on Robotics & Automation*, pp. 4208–4213, 2001.

[7] A. Billard and M. Mataric. Learning human arm movements by imitation: Evaluation of a biologically-inspired connectionist architecture. In *First IEEE-RAS International Conference on Humanoid Robotics*, 2000.

[8] A. J. Ijpeert, J. Nakanishi, and S. Schaal. Trajectory formation for imitation with nonlinear dynamical systems. In *Proceedings of the 2001 IEEE/RSJ International Conference on Intellignent Robots and Systems*, pp. 752–757, 2001.

[9] H. Miyamoto, S. Schaal, F. Gandolfo, H. Gomi, Y. Koike, R. Osu, E. Nakano, Y. Wada, and M. Kawato. A kendama learning robot based on bi-directional theory. *Neural Networks*, Vol. 9, pp. 1281–1302, 1996.

[10] K. Hosoda and M. Asada. Versatile visual servoing without knowledge of true jacobian. In *Proceedings of IEEE International Conference on Robotics & Automation*, pp. 186–193, 1994.

[11] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene of from two projections. *Nature*, Vol. 293, pp. 133–135, 1981.

[12] Y. Yoshikawa, M. Asada, and K. Hosoda. View-based imitation learning by conflict resolution with epipolar geometry. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1416–1427, 2001.