Subjective Robot Imitation by Finding Invariance

（不変性の発見によるロボットの主観的模倣）

A dissertation submitted to the Department of Adaptive Machine Systems in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Engineering at the OSAKA UNIVERSITY

Yuichiro Yoshikawa

January 2005

Subjective Robot Imitation by Finding Invariance
by
Yuichiro Yoshikawa

Abstract

Recent studies point out the virtues of a relatively new field of research with regard to robotics: imitation. The idea is that a robot can acquire new behaviors by just observing human demonstrations. In addition, building a robot with such a competence might help us understand human intelligence underlying the capability of imitation based on a constructivist approach. In my work I concentrate on "subjective robot imitation", that means that the robot imitates the demonstrator autonomously. The designer analyzes neither the demonstration, the demonstrator, nor the robot itself. Subjective robot imitation faces two main problems: How can demonstrations be identified even if they are seen from different viewpoints, and how can a demonstration be imitated if the embodiment of the demonstrator differs from the robot's own embodiment? In order to tackle these two problems this dissertation addresses three issues by focusing on what types of invariance can be used by the robot.

In the first part of my work, the robot should learn to imitate the demonstration of another identical robot by observing it (1). The opt-geometric constraint between views that originates from the correspondence of body parts is utilized to map the observed demonstration to its corresponding motion only through mappings between its sensorimotor space. Then, how a robot can find its body from its uninterpreted sensory data is addressed. This is a fundamental step for acquiring body representation to construct the mapping between bodies (2). The invariance in self-body-observation is modeled as a statistical distribution of the variance of its sensory data and is utilized to discriminate its body from non-body. Finally, we cope with the issue how a vocalizing robot can acquire human vowels that it cannot duplicate as they are (3). The invariance in the interaction with a human caregiver who imitatively responds to the robot's behavior and the robot's subjective criteria that consider the toil involved in the articulation are utilized to find how to articulate sounds so that the caregiver interprets them as vowels.

Acknowledgement

I am very grateful to Prof. Minoru Asada for his guidance, patience, and encouragements all along this project. He introduced me to the field of cognitive developmental robotics and gave me large amount of freedom to study while he has spent a lot of time for discussion. His critical and valuable comments have helped me shape my thought and complete this dissertation.

I am also very grateful to Prof. Koh Hosoda for his companionable encouragements and a lot of constructive comments throughout this work. I was very lucky to have many chances to talk with him during the trips for conferences. I have learned various kinds of things from him.

I also thank Prof. Yoshiaki Shirai and Prof. Hiroshi Ishiguro for their constructive readings of this dissertation and their valuable advice.

I wish to thank all members of the laboratory. Junpei Koga and Yoshiki Tsuji helped the important experiments in this dissertation. Dr. Minato Takashi, Dr. Yukie Nagai, Mr. Masaki Ogino, and Mr. Tetsushi Ikeda has spent a lot of time for discussion and given me a lot of valuable comments on my work. Dr. Yasutake Takahashi, Dr. Noriaki Mitsunaga, Mr. Yasunori Tada, Mr. Takashi Takuma, Mr. Issei Tsukinoki have helped me to implement real robots. Other members who I do not list up here also have surely supported me in various ways. I am grateful to them for the enjoyable laboratory-life.

Finally, I wish to greatly thank to my parents, Yasuji and Nanako, my sister, Asako, and my grandparents, Masao and Hatsuko, for their constant support and patience through my growth. I am sure that I could not accomplish anything without them.

I gratefully acknowledge financial support from the Japan Society for the Promotion of Science in the form of a grant for JSPS Research Fellowships for Young Scientists.

January, 2005
Yuichiro Yoshikawa

Contents

Chapter 1

Introduction

1.1 Overview

Until recently, a robot has been seen as a hard-headed machine that was programmed to perform a number of repetitive tasks but was specialized to a limited environment such as a factory. The main reason for such a limitation seems caused by the way how its behavior is designed, that is programmed based on the designer's analysis of the constraints in the task. Consequently, the adaptability of the robot was limited to what extent the designer had supposed in advance. Alternatively, providing a robot with a competence of learning by which it adapts to the changes in the task, the environment, and its body by itself, is a promising approach. However, in the case of learning by a robot with many sensors and degrees of freedom, we face with a problem called "curse of dimension", that is, increase in the number of sensory modalities and degrees of freedom causes an exponential explosion of the searching space for learning. *Learning by imitation* is supposed to be a promising approach to overcome the curse and has been one of the most interesting topics in recent robotics (see a survey [1]) because the size of searching space needed to be explored might be drastically reduced by starting from imitation [2].

Learning by imitation without the designer's analysis It is suggested that a robot can structure its own sensory input through the interaction with the environment, and thereby induce regularities that significantly simplify learning [3, 4]. However, it is usually difficult for the designer to realize such a mechanism of the robot that causes reguralities suitable for learning since the analysis of the interaction between the body and the environment (including the demonstrator) is not trivial in general. Therefore, a robot is expected to

find how to induce the reguralities from its own viewpoint since its sensorimotor sequence reflects the reguralities. We believe that, with the competence of imitation from its own viewpoint instead of relying on the designer's analysis of the behavior to be imitated, it can acquire behaviors in its own way by utilizing such reguralities even though they are difficult for the designer to analyze. As a result, even a naïve user about the robot would be able to provide it with behaviors since he/she needs just to show the desired one.

In the previous work on robot imitation, however, the system of imitation was usually constructed based on the designer's analysis of the situations of imitation. For example, the designer introduced the model of behavior by analyzing what kinds of demonstrations shall be shown [5, 6, 7, 8], or the transformation from the observed motion or behavior to the robot's one by analyzing the relationship between bodies of the demonstrator and the robot [9, 10, 11, 12, 13].

Understanding the mechanism of imitation in human brain On the other hand, *mirror neuron* found in the brain of a macaque monkey [14] has motivated a lot of researchers in neuroscience, cognitive science, and so on, which concern the system of imitation because of its interesting activation. It is activated not only in the execution of a particular type of action but also in the observation of the same action performed by other agents. It is reported that the human brain also exhibits the similar phenomena [15, 16, 17, 18]. Interestingly, the "mirror" system in human brain involves the Broca's area that is thought to be a substrate for speech production [19]. Based on these findings, the system of imitation is speculated to play important roles in several aspects of human intelligence such as understanding the other's behavior [20], understanding other's mental state [21], and language acquisition [19]. However, the acquisition process of the system of imitation has not been revealed yet due to the limitation in the methodology to study the development of living system.

Synthetic study is expected to serve an alternative methodology to understand the cognitive developmental process of biological agents [22, 23] since we can refine our understanding through the iteration of evaluating, improving, and reimplementing the hypothesized model. At the same time, experimental data concerning the developmental process of a biological agent are expected to give us some hints to build an intelligent robot. Therefore, as aimed in "cognitive developmental robotics [22]", we might be able to approach to both understanding the cognitive developmental process of human beings and obtaining the design principle of an intelligent robot by building a robot that

reproduces the similar cognitive developmental process as human beings do. Although there have been already constructivist studies on imitation, they usually rely on the designer's analysis of the situation of imitation as mentioned above.

Therefore, building a robot that performs *subjective imitation* is addressed in this dissertation, which is determined as imitation by using only its sensory data instead of relying on the designer's analysis of the situation of imitation. It is considered to not only be one of the fundamental problems to make robot learn by imitation but also be one of the prerequisites to approach to reveal the human developmental process of imitation through synthesizing.

1.2 The problems

We consider subjective robot imitation where a robot observes the demonstration by its own sensors and then imitates it. Unlike the previous studies on robot imitation, we assume that the designer does not give any explicit knowledge on either bodies of itself or the demonstrator, or the relationship between them. To imitate the demonstration from subjective viewpoint can be rephrased to find how to map the observed sensor data to its motor commands. Learning the mapping is not trivial since this mapping could be intrinsically many-to-many due to what we call *different viewpoint problem* and *different body problem*.

Different viewpoint problem The *different viewpoint problem* is caused since we assume that a robot can sense the world only through the observation with the sensors embedded on its body, in other words, the states of the demonstrator and of itself are only partially observable. For example, when it uses visual sensors, it cannot see the state of the demonstrator but can obtain appearances of demonstration that varies depending on its viewpoint. As well as the auditory sensors differently receive the source sound depending on the position from where it hears the sound. It should be able to cope with such many-to-one correspondence. Therefore, the different viewpoint problem can be reduced to a problem to identify different appearances of a demonstration that varies depending on the viewpoint.

In the previous studies, to avoid the different viewpoint problem, the designer usually introduced a mechanism to reconstruct the state. However, both bodies of the demonstrator and itself were needed to be modeled in the

designer's coordinate system. Therefore, they and their relationship must be calibrated by the designer since it cannot access the designer's coordinate systems, in other words, the references to learn about them are not available from its own viewpoint.

How can a robot solve the different viewpoint problem from its own viewpoint? Since only the sensorimotor space can be accessible by a robot, the references to learn the mappings between different sensorimotor sub-spaces is expected to be found by the robot. For example, visual congruence of the bodies both of the demonstrator and the robot or the contingency between the motion and the sensory changes could be utilized as the references. Therefore, we address the issues to map the observed demonstrations to its motor command only through mappings between its sensorimotor space. We need to consider how to construct the mapping based on such references and how to construct the representation of the references themselves.

different body problem When the robot's body is different from the demonstrator's one, *different body problem* occurs, that is, it cannot duplicate the demonstration as it is. For example, in the case of imitation between 2-link robots with different link parameters, an imitator cannot reproduce both trajectories of the end-effector and joint angles of the other. What make the issue more difficult is the difference in body structure. For example when a dog-like robot tries to imitate the human behavior with his/her hand, it is difficult even to define what kinds of aspect in demonstration should be reproduced. To perform imitation in these cases, it needs to abstract observed behavior to some extent. However, abstraction brings arbitrariness in the imitation process. In other words, a demonstration can be matched with different behavior depending on the way to abstract, that is the definition of imitation. Therefore, the different body problem can be reduced as a problem to find a possible definition.

However, to find the definition of imitation is a formidable issue since there does not seem the universal definition of imitation. In this dissertation, therefore, we focus on what kinds of information can be utilized from the robot's own viewpoint as the constraints to define imitation. Although there will remain a formidable problem how to select or combine the constraints, it seems to be considered first.

Since it is situated in the interaction with the demonstrator, a robot might be able to utilize the invariance in the demonstrator's tendency of reaction to its behavior. In other words, if it can induce the imitative response of the demonstrator to its behavior as observed in the interaction between humans

[24, 25, 26, 27] or non-human primates [28, 29], the sensorimotor experience can be directly utilized as references. Utizling subjective criteria reflecting the property of the robot's body such as energy minimum or jerk minimum [9] is also promising not only to reduce the arbitrariness of mapping but also to optimize the acquired behavior under the constraint of its embodiment. Therefore, we address the issues to build a prototype robot that learns to imitate the behavior of demonstrator that has different body structure by utilizing both the invariance in interaction and the subjective criteria.

1.3 Issues to be considered

To solve these problems, the invariance that originates from the correspondence of the body and the interaction with the demonstrator could be utilized. In this dissertation, we address three partial problems towards subjective imitation (see Fig. 1.1). Fig. 1.2 illustrates what kinds of constraints are utilized in each issue. The contents are following:

Chapter 2 To solve the different viewpoint problem, we introduce a paradigm called the demonstrator's view recovery to perform view-based imitation. It is assumed that the learner has the same body structure as the demonstrator. This assumption means that the parts of the bodies of the learner and the demonstrator are visually identical to each other, therefore, only the scale variation is allowed. In this paradigm, at first the learner recovers the demonstration observed in the demonstrator's view to observe the demonstrator's self-body. Then, it reproduces the visually same motion by performing the recovered trajectory of the demonstration in its own view to observe its self-body. In the proposed method, the demonstrator's view recovery is performed by utilizing the opt-geometric constraint, that is epipolar geometry [30] of the views that originates from the correspondence of the body parts.

Adaptive visual servoing method which involves online estimator of the parameters to control the quantity of the sensory features [31], is applied both to estimate the parameter of view recovery and perform the recovered trajectory of the demonstration. Experiments with two identical manipulator show the validity of the proposed method.

Chapter 3 Then, we have another fundamental question how a robot can acquire the representation of its body to consider the correspondence of the body parts. Therefore, we introduce a method to find its body from its sensor

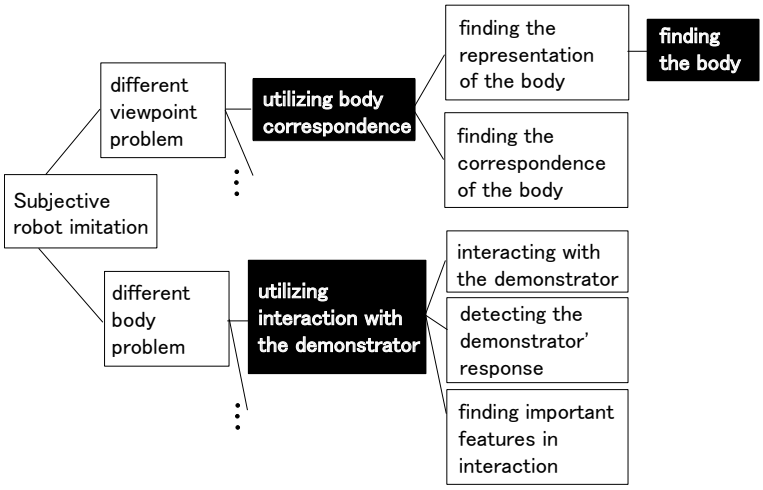


Figure 1.1: A load-map to approach subjective robot imitation where issues in the black boxes are addressed in this dissertation: The correspondence of the bodies and the interaction with the demonstrator are considered to be available to solve the different viewpoint problem and different body problem, respectively. Finding the representation of the body and the correspondence of bodies are assumed to be already acquired to cope with how to utilize body correspondence. How to interact with the demonstrator, detect the demonstrator's response, and find important features in the interaction are also assumed to be solved to focus on how to utilize the interaction.

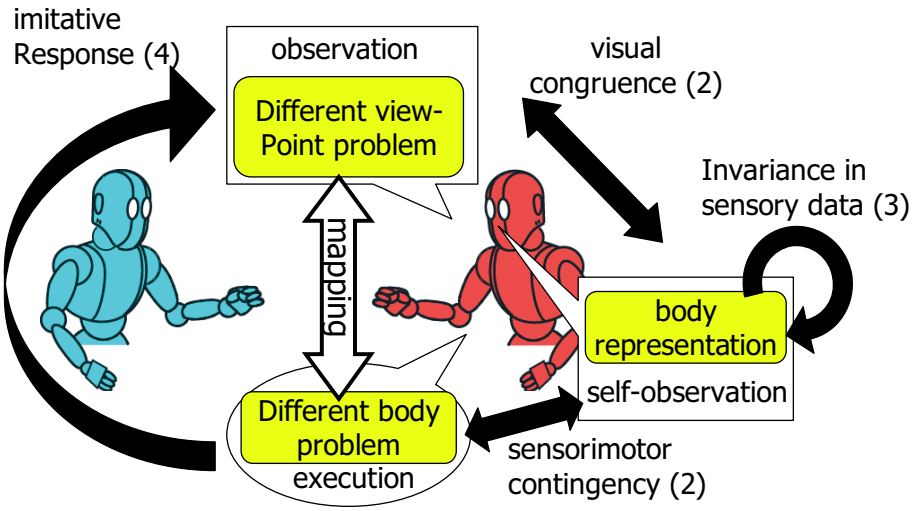


Figure 1.2: The constraints utilized to construct the mapping between the observed sensory data and the robot's motor command in each chapter. The numbers associated with the filled arrows correspond to the chapter in which the constraints are utilized. Visual congruence and sensorimotor contingency, and imitative response are utilized to construct the mapping in chapter 2 and 4, respectively, while the invariance of the sensory data is used to find the robot's body in sensory data in chapter 3.

data without *a priori* knowledge of the relationship between the sensor data and its body since this process seems a first step towards body representation. The basic idea of the proposed method is to utilize the invariance of sensor data in self-body-observation. In other words, the robot regards what it always observes as its body.

The sensory data might perturb for the changes in observing posture and sensory noise both of which depend on the combination of sensory data and the properties of the perceived body surface. Therefore, to discriminate its body from non-body, a robot should complementarily utilize the invariance in multiple sensory data. The proposed method is based on a conjecture about the distribution of the variance of sensations in terms of each observing posture. It can be approximated by a mixture of two Gaussian distributions for observing the body and non-body, respectively. After estimating the distribution by an EM algorithm, it can discriminate its body from non-body by judging which distribution likely causes the variance of sensory data in the current observing posture. Experiments with real robots show the validity of the proposed method.

Chapter 4 To tackle the different body problem, we focus on the developmental process of phonemes in a human infant. Inspired by the observation that infants acquire phonemes common to adults without having the capability to articulate, or having *a priori* knowledge about the relationship between the sensorimotor system and phonemes, a constructivist approach toward building a robot that reproduces a similar developmental process is conducted. Two general issues are addressed: what are the interactive mechanisms involved and what should be the behavior of the caregiver/teacher? Based on findings in developmental psychology, it is conjectured that (a) the caregiver's vocalization in response to infants' cooing reinforces the infant's articulation along the caregiver's phonemic categories, and (b) the caregiver's repetition with adult phonemes helps to specify the correspondence between cooing and the caregiver's phonemes as well as determining the acoustic properties of the phonemes.

The robot consists of an artificial articulatory system with a 5-DoF mechanical system deforming a silicon vocal tract connected to an artificial larynx, an extractor of formants, and a learning mechanism with self-organizing auditory and articulatory layers. Starting off with random vocalizations, the system uses the caregiver's repetitive utterances to bootstrap its learning. In order to resolve the arbitrariness in determining proper articulations, the torque to deform the tract and its resultant deformation are minimized. The experimental

results, discussion, and future issues are given.

As listed above, what are tackled in this dissertation are partial problems of imitation. However, they involve indispensable, fundamental issues to build a robot that performs imitation from subjective viewpoint. we believe that the implications from the work could be elements of the design principle of an intelligent robot or understanding human intelligence. Therefore, in the final part of this dissertation, we would like to argue the further extensions and the integration of these elements.

Chapter 2

Imitation based on the demonstrator's View recovery

2.1 Introduction

In recent robotics, the capability of imitation has been focused as an important element of an intelligent robot. Leaning by imitation seems an promising approach to solve so-called "curse of dimension" problem where increase in the number of sensory modalities and degrees of freedom causes an exponential explosion of the searching space for learning [1]. In the fields of studies concerning human intelligence involving neuroscience, psychology, and so on, the system underlying the competence to imitate is said to be a basis for several aspects of human intelligence such as action execution and recognition [20], estimating other's mental state [21], and even language acquisition [19]. However, the developmental process of the system of imitation has not been evident yet. Therefore, building a robot that performs imitation is an interesting issue from both viewpoints of obtaining the design principle of intelligent robot and understanding intelligence of humans [22, 32, 33].

The research of imitation has started from the studies on teaching by doing in the industrial applications. Instead of time-consuming path planning, the users could provide the desired motion by letting their robots memorize the demonstrated one through teaching pendant. Although there still remain important issues on imitation (ex. motion segmentation [34, 35]) after the motion is directly given by the designer, teaching by showing should be addressed since we consider a robot that learns from subjective viewpoint.

Kuniyoshi et al. [5] and Ikeuchi and Suehiro [6] built robotic systems which learn assembly tasks by watching a human demonstration. Based on the mod-

els of specific behaviors given by the designer in advance, it can recognize assembly tasks demonstrated by the experimenter and then perform corresponding behaviors. Apart from static behavior like assembly task, imitation of dynamic movement such as dancing [7] or learning by imitation of dynamic tasks such as air hokey and marble maze [8] have been also addressed based on given primitives of behaviors. However, it seems difficult for the designer to prepare the universal model or primitive that can be applied for general behaviors. Since we assume that it does not have any explicit models of behaviors to be imitated, reproducing the motion seems the first step to imitate the behaviors.

In the previous studies on imitating the motion without the explicit models of behaviors given by the designer, it was not necessarily considered how can a real robot reproduce the demonstration from its own viewpoint since they focused on how it articulates or memorizes the demonstration. Therefore, it was usually assumed that it can obtain the motor command or the joint angle to reproduce the demonstration by using a motion capture system [13, 10, 11], or attaching a sensor-suit on the demonstrator [12]. In other words, the designer must have paid much effort to let it know the demonstrator's internal state. Namely, the designer must have modeled the body of the learner and the demonstrator to transform the joint angles from the observed data. To release the designer from such a burden, how to perform imitation by a view-based mechanism should be addressed.

As one of view-based studies, Miyamoto et al. [9] have proposed a method based on 3-D geometric reconstruction. In their approach, the observed trajectory of the demonstration in the image space is transformed into the three dimensional coordinate system. However, the designer needs to calibrate the transformations between the robot centered coordinate system and the world coordinate one. Furthermore, the parameters calibrated by the designer prone to be error sensitive and less adaptive to the changes in the robot's body.

In other type of the view-based imitation mechanism [36, 37, 38], imitation are emerged by maintaining the visual feature that indicates geometrical relationship between the demonstrator and the learner. For example, locomotion behavior is imitated by maintaining the distance between them [36, 37], and head movement of the demonstrator is imitated by maintaining the angle between the facial normal of the demonstrator and the learner's view directions [38]. But, in these approaches, it can imitate limited types of behaviors because it seems difficult to choose the suitable features to be maintained.

Kuniyoshi et al. [39] proposed the mechanism that models the neonatal imitation [40]. In their model, the agent first learns to associate its motor

commands with an optical flow which is caused by its own pre-defined motions by using a Hopfield type neural network with non-monotonic activation function. Then, when it observes an optical flow caused by demonstration, it reflectively generates the motor commands associated with the perceived optical flow. Based on the similar idea, Andry et al. [41] or Ogino et al. [42] have addressed how to reproduce the mirrored motion. Note that the systems work only when they face with the demonstrator in a mirroring manner and the mirror image of the demonstration can be still considered to be the same as the original one. It is not clear how to extend the mechanism from the limitation of mirroring imitation.

In this chapter, to perform various demonstrations in a view-based manner, we will introduce a mechanism based on a paradigm, what we call *demonstrator's view recovery*. In the paradigm, the robot that imitates the demonstration (hereafter, *the learner*) recovers the demonstration observed in the demonstrator's view to observe the demonstrator's self-body, then it reproduces the visually same motion by performing the recovered trajectory of the demonstration in its own view to observe its self-body.

(1) To perform the demonstrator's view recovery, we propose a basic mechanism based on "epipolar geometry [30]" between views. Here, we assume that the learner's body structure is the same as the demonstrator's one, where only scale variation is allowed, and that the initial postures of both are the same, instead of assuming the designer's calibration. Note that the assumption of initial posture will be released in the latter part of this chapter. First, it recovers the observed trajectory of the demonstration in its own view, and then control to follow the recovered trajectory by an adaptive visual servoing method [31] that does not need the designer's calibration or three dimensional reconstruction. As a result, it reproduces the demonstration from its own viewpoint. The transformation is called "view transformation". To know the parameters of view transformation, it estimates a fundamental matrix of epipolar geometry.

(2) Then we introduce a method to perform imitation based on the demonstrator's view recovery even if the initial postures are different from each other. In the proposed method, at first, the learner estimates the parameters of a view transformation by regarding both postures as the same even though they are actually different. Therefore, the estimated parameters causes errors in the view transformation. Then, it controls its posture to minimize the errors, and estimates them again at the new posture. Iterating these procedures, the parameters which cause no transforming errors are estimated. It is approximately ensured that it can minimize the error by the control based on the gradient

method when the demonstrator's posture is not so different from the learner's one.

The rest of this chapter is organized as follow: after defining imitation in this study, the basic mechanism of demonstrator's view recovery based on epipolar geometry and the experiments are given. Then, the method to cope with the case that the demonstrator's initial posture is different from the learner's one is proposed and examined. Discussion and summary are given in the final part of this chapter.

2.2 A definition of Imitation

We assume that the learner has the same body structure and stereo cameras as the demonstrator. This assumption enables us to avoid the different body problem. For example, when they have different link parameters, the both motions which are resultant from the same joint angle trajectories may have different meanings [43]. In such a case and a further case that even degrees of freedom is different, we must discuss what kind of aspect in imitation should be supposed to be similar. However, there does not seem a clear definition which is applicable to all the cases of imitation. Instead of stopping at the different body problem, we assumes such similarity of bodies in order to study the mechanism of imitation in a simple case. This assumption allows a simple definition, such as

> *generating a motion so that the trajectory of own body observed in own view is congruent with one of the demonstrator's body observed in the demonstrator's view when the demonstrator is supposed to observe itself in the same posture as the learner.*

Denote a image plane of camera p as $[p]$ and a view of an agent A that observes an agent O as V_A^O. In this case, A and O can be L or D that indicates the learner or the demonstrator, respectively (see Fig. 2.1). Since the agents possess two cameras, V_A^O indicates two image planes. Suppose that $[L_D]$ and $[R_D]$ denotes the demonstrator's left and right views to observe itself, respectively, and $V_D^D \triangleq \{[L_D], [R_D]\}$ denotes both of them. In the same way, $V_L^L \triangleq \{[L], [R]\}$ denotes the learner's views to observe itself and $V_L^D \triangleq \{[l], [r]\}$ denotes the learner's views to observe the demonstrator. If the learner knows V_D^D, it can imitate by performing the trajectory of the demonstration observed in V_D^D as the desired one in V_L^L.

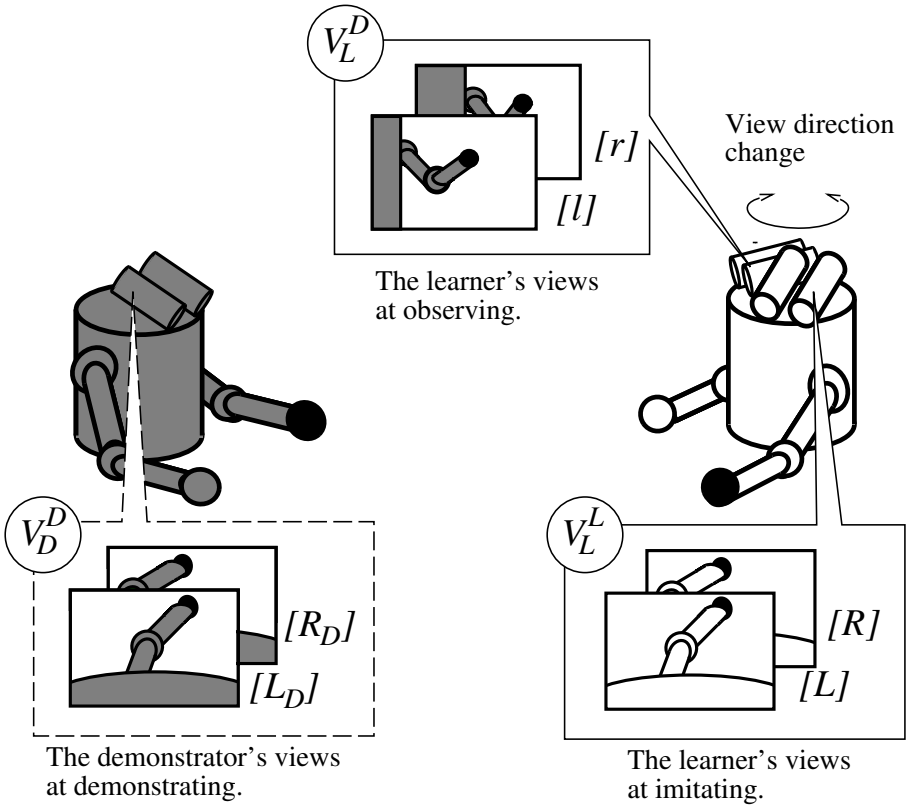


Figure 2.1: An overview of imitation. The learner's view at observing the demonstration $V_L^D = \{[l], [r]\}$ and at imitating $V_L^L = \{[L], [R]\}$, and the demonstrator's view $V_D^D = \{[L_D], [R_D]\}$ at demonstrating. The body structure and the stereo cameras of the learner are the same as ones of the demonstrator.

Since it cannot directly access to V_D^D, it needs to recover the demonstration supposed to be observed in V_D^D from the observation in V_L^D. We call the transformation to recover unknown views form known views "view transformation", and the recovery of the demonstrator's unknown views "demonstrator's view recovery".

2.3 Demonstrator's View Recovery based on Epi-polar Geometry

The basic idea of the demonstrator's view recovery is to utilize epipolar geometry between the learner's views ($[l]$ and $[r]$) and the demonstrator's views ($[L_D]$ and $[R_D]$) both of which are to observe the demonstration. It is assumed that the learner's body structure and initial posture are the same as the demonstrator's one. By virtue of this assumption, it can estimate the parameters of epipolar equation, and thereby construct the view transformation between V_L^D and V_L^L. After recovering the trajectory of demonstration using the constructed view transformation, the learner performs imitation by reproducing the recovered one with adaptive visual servoing [31]. In the rest of this section, the detail mechanism of imitation based on the demonstrator's view recovery is introduced.

2.3.1 The mechanism of view transformation based on epipolar geometry

Epipolar geometry

Suppose that two cameras watch the same point in the world. The point, its projections points on the both image planes, and the centers of both cameras are on the same plane (see Fig. 2.2). The plane is called epipolar plane and such geometry between two cameras is called epipolar geometry. According to epipolar geometry, each projected point is constrained onto a line which is the intersection of epipolar plane and the image plane, and therefore is called epipolar line.

Epipolar geometry between $[p]$ and $[q]$ for the i-th point is described as following epipolar equation:

$$ {}^{p}\tilde{\boldsymbol{m}}_i^{T\,pq}\boldsymbol{F}\,{}^{q}\tilde{\boldsymbol{m}}_i = 0, \tag{2.1} $$

where ${}^{p}\tilde{\boldsymbol{m}}_i = [{}^{p}\boldsymbol{m}_i^T, 1]^T \in \Re^3$ and ${}^{q}\tilde{\boldsymbol{m}}_i = [{}^{q}\boldsymbol{m}_i^T, 1]^T \in \Re^3$ denote homogeneous coordinates of the image coordinates of the i-th projected points, and ${}^{pq}\boldsymbol{F} \in$

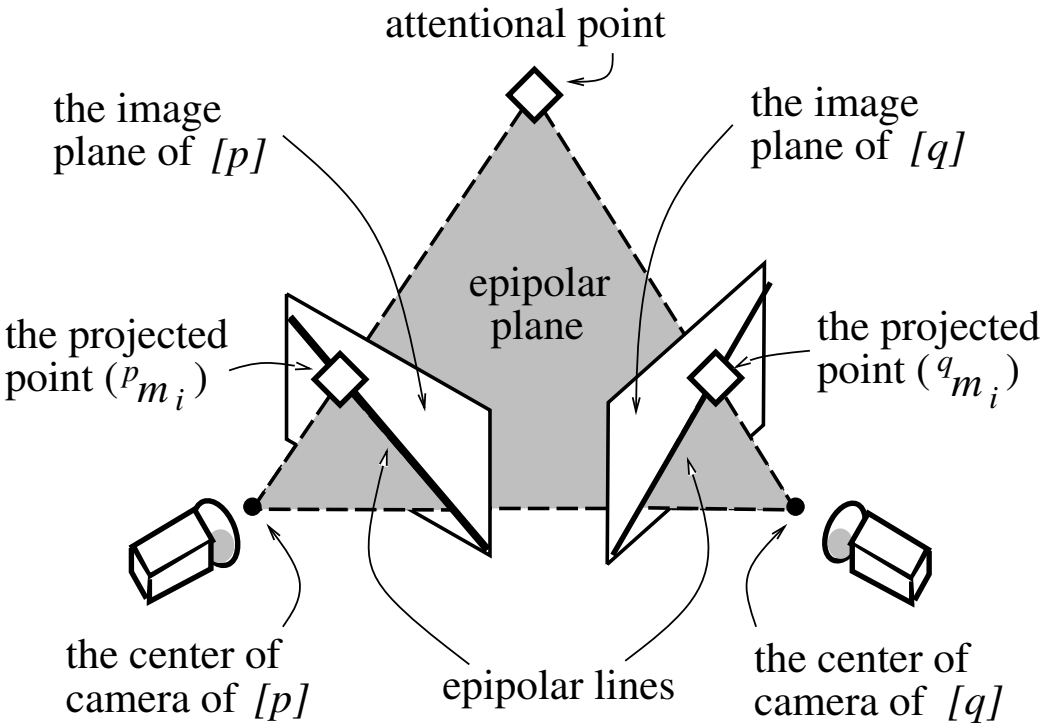


Figure 2.2: An over view of epipolar geometry: when an attentional point is projected on the image planes ([p] and [q]), the attentional point, the projected points ($^{p}\boldsymbol{m}_i$ and $^{q}\boldsymbol{m}_i$), and the center of cameras share the same plane (*epipolar plane*). In other words, $^{p}\boldsymbol{m}_i$ or $^{q}\boldsymbol{m}_i$ is constrained onto a line (*epipolar line*) which is the intersection of the epipolar plane and the image plane [p] or [q], respectively.

$\Re^{3\times3}$ is called fundamental matrix which depends on the internal parameters of two cameras and the relative position and posture between them.

A fundamental matrix has 8 degrees of freedom when the camera is modeled by a perspective camera, or 4 degrees of freedom when it is modeled by an affine camera. In general, therefore, a minimum of 8 pairs of matched projected points $({}^{p}\tilde{\boldsymbol{m}}_i,\ {}^{q}\tilde{\boldsymbol{m}}_i)$ are required to uniquely determine the fundamental matrix of perspective cameras, as well as 4 pairs are required for affine cameras. They can be determined by the minimizing residual of the epipolar equations [30].

When the cameras are modeled by an affine one, we have another method to estimate a fundamental matrix. In the case of affine camera, the fundamental matrix is simplified such as

$$ {}^{pq}\boldsymbol{F} = \begin{bmatrix} 0 & 0 & {}^{pq}f_{13} \\ 0 & 0 & {}^{pq}f_{23} \\ {}^{pq}f_{31} & {}^{pq}f_{32} & {}^{pq}f_{33} \end{bmatrix}. \tag{2.2} $$

Expanding eq. (2.1), we obtain

$$ {}^{pq}\boldsymbol{x}_i{}^{pq}\boldsymbol{f} + {}^{pq}f_{33} = 0 \tag{2.3} $$

where

$$ {}^{pq}\boldsymbol{x}_i = \begin{bmatrix} {}^{p}\boldsymbol{m}_i \\ {}^{q}\boldsymbol{m}_i \end{bmatrix} \in \Re^4, \quad {}^{pq}\boldsymbol{f} = \begin{bmatrix} {}^{pq}f_{13} \\ {}^{pq}f_{23} \\ {}^{pq}f_{31} \\ {}^{pq}f_{32} \end{bmatrix} \in \Re^4. \tag{2.4} $$

These elements of the fundamental matrix are determined by minimizing the sum of distances of each vector ${}^{pq}\boldsymbol{x}_i$ to the hyper-plane in the 4-dimensional space that is represented by eq. (2.3) [44]. Such a vector is determined as an eigenvector associated with the minimal eigenvalue of ${}^{pq}\boldsymbol{W}$, where

$$ {}^{pq}\boldsymbol{W} = \sum_{i=1}^{N} \left({}^{pq}\boldsymbol{x}_i - {}^{pq}\boldsymbol{x}_0 \right) \left({}^{pq}\boldsymbol{x}_i - {}^{pq}\boldsymbol{x}_0 \right)^T, \tag{2.5} $$

where

$$ {}^{pq}\boldsymbol{x}_0 = \frac{1}{N} \sum_{j=1}^{N} {}^{pq}\boldsymbol{x}_j \tag{2.6} $$

then, the rest element ${}^{pq}f_{33}$ is determined by

$$ {}^{pq}f_{33} = -{}^{pq}\boldsymbol{x}_0^T{}^{pq}\boldsymbol{f}. \tag{2.7} $$

A mechanism of view transformation

Suppose that there are three views in which two of them ($[l]$ and $[r]$) are known and the rest ($[L_D]$) is unknown. In this section, the mechanism to determine the image coordinate in $[L_D]$ of an attentional point from the image coordinates in $[l]$ and $[r]$ of the corresponding point is introduced under the assumption that fundamental matrices between these views has been already known. In other words, the problem is to know ${}^{L_D}\boldsymbol{m}_i$ from ${}^{l}\boldsymbol{m}_i$ and ${}^{r}\boldsymbol{m}_i$ by using ${}^{lL_D}\boldsymbol{F}$ and ${}^{rL_D}\boldsymbol{F}$.

Corresponding pairs, $\{{}^{l}\boldsymbol{m}_i, {}^{L_D}\boldsymbol{m}_i\}$, and $\{{}^{r}\boldsymbol{m}_i, {}^{L_D}\boldsymbol{m}_i\}$ are satisfied with an epipolar equation, respectively, such as

$$ {}^{l}\tilde{\boldsymbol{m}}^T \, {}^{lL_D}\boldsymbol{F} \, {}^{L_D}\tilde{\boldsymbol{m}} = 0, \tag{2.8} $$

$$ {}^{r}\tilde{\boldsymbol{m}}^T \, {}^{rL_D}\boldsymbol{F} \, {}^{L_D}\tilde{\boldsymbol{m}} = 0. \tag{2.9} $$

Expanding these equation, we obtain

$$ \boldsymbol{A}({}^{l}\boldsymbol{m}_i, {}^{r}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}) \cdot {}^{L_D}\boldsymbol{m}_i = \boldsymbol{d}_i({}^{l}\boldsymbol{m}_i, {}^{r}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}), \tag{2.10} $$

where

$$ \begin{aligned} \boldsymbol{A}({}^{l}\boldsymbol{m}_i, {}^{r}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}) &= \begin{bmatrix} a({}^{l}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}) & b({}^{l}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}) \\ a({}^{r}\boldsymbol{m}_i; {}^{rL_D}\boldsymbol{F}) & b({}^{r}\boldsymbol{m}_i; {}^{rL_D}\boldsymbol{F}) \end{bmatrix}, \\ \boldsymbol{d}_i({}^{l}\boldsymbol{m}_i, {}^{r}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}) &= \begin{bmatrix} -c({}^{l}\boldsymbol{m}_i; {}^{lL_D}\boldsymbol{F}) \\ -c({}^{r}\boldsymbol{m}_i; {}^{rL_D}\boldsymbol{F}) \end{bmatrix}, \end{aligned} \tag{2.11} $$

and where the functions a, b, and c are given by

$$ \begin{bmatrix} a({}^{p}\boldsymbol{m}; {}^{pq}\boldsymbol{F}) \\ b({}^{p}\boldsymbol{m}; {}^{pq}\boldsymbol{F}) \\ c({}^{p}\boldsymbol{m}; {}^{pq}\boldsymbol{F}) \end{bmatrix} = ({}^{p}\tilde{\boldsymbol{m}}^T \, {}^{pq}\boldsymbol{F})^T = \begin{bmatrix} {}^{pq}f_{11}{}^{l}x + {}^{pq}f_{21}{}^{l}y + {}^{pq}f_{31} \\ {}^{pq}f_{21}{}^{l}x + {}^{pq}f_{22}{}^{l}y + {}^{pq}f_{32} \\ {}^{pq}f_{31}{}^{l}x + {}^{pq}f_{32}{}^{l}y + {}^{pq}f_{33} \end{bmatrix}. \tag{2.12} $$

Therefore, if $\boldsymbol{A}$ is full rank, ${}^{L_D}\boldsymbol{m}_i$ can be determined by solving the simultaneous equations (2.10) such as,

$$ {}^{L_D}\boldsymbol{m}_i(t) = \boldsymbol{A}^{-1}({}^{l}\boldsymbol{m}_i(t), {}^{l}\boldsymbol{m}_i(t); {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}) \cdot \boldsymbol{d}({}^{l}\boldsymbol{m}_i(t), {}^{l}\boldsymbol{m}_i(t); {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}) \tag{2.13} $$

Since each row in eq. (2.10) represents an epipolar line, solving the simultaneous equations is equivalent to determine the intersection point between them (see Fig. 2.3). When $\boldsymbol{A}$ is not full rank, we cannot avoid numerical sensitivity in calculating eq. (2.13), in which epipolar lines are parallel.

However, we have two method to determine the intersection point more stably. One is by adding more known views in which the matched image

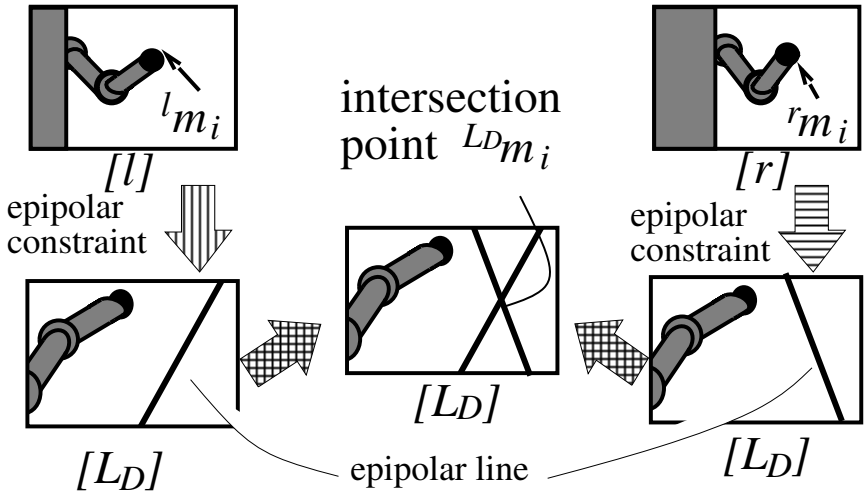


Figure 2.3: The method to find a matched point ($^{L_D}\boldsymbol{m}_i$) in [L_D] with ones ($^{l}\boldsymbol{m}_i$ and $^{r}\boldsymbol{m}_i$) in [l] and [r]. Solving simultaneous equations (2.10) is equivalent to determine an intersection point of epipolar lines.

coordinates and a fundamental matrices between [L_D] and them. By adding such a known view, additional epipolar equation is introduced as a row of simultaneous equation. If we use the affine camera model, another method is available by adding one more unknown view [R_D] in which only a fundamental matrix $^{L_D R_D}\boldsymbol{F}$ is known. When we try to find the intersection point in [L_D], i.e. $^{L_D}\tilde{\boldsymbol{m}}_i$, by solving the eq. (2.10) and one in [R_D], i.e, $^{R_D}\tilde{\boldsymbol{m}}_i$ by solving

$$^{l}\tilde{\boldsymbol{m}}_i^T \, {}^{lR_D}\boldsymbol{F} \, {}^{R_D}\tilde{\boldsymbol{m}}_i = 0, \tag{2.14}$$

$$^{r}\tilde{\boldsymbol{m}}_i^T \, {}^{rR_D}\boldsymbol{F} \, {}^{R_D}\tilde{\boldsymbol{m}}_i = 0, \tag{2.15}$$

we can use an additional epipolar equation between [L_D] and [R_D];

$$^{L_D}\tilde{\boldsymbol{m}}_i^T \, {}^{L_D R_D}\boldsymbol{F} \, {}^{R_D}\tilde{\boldsymbol{m}}_i = 0. \tag{2.16}$$

Therefore, we can determine unknown $^{L_D}\boldsymbol{m}_i$ and $^{R_D}\boldsymbol{m}_i$ by solving the simultaneous equations which consist of eq. (2.8), (2.9), (2.14), (2.15), and (2.16), such as

$$\begin{bmatrix} {}^{L_D}\boldsymbol{m}_i \\ {}^{R_D}\boldsymbol{m}_i \end{bmatrix} = \boldsymbol{A}'({}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}, {}^{lR_D}\boldsymbol{F}, {}^{rR_D}\boldsymbol{F}, {}^{L_D R_D}\boldsymbol{F})^{+} \cdot \boldsymbol{B}'({}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}, {}^{lR_D}\boldsymbol{F}, {}^{rR_D}\boldsymbol{F}, {}^{L_D R_D}\boldsymbol{F}) \cdot \boldsymbol{d}_i'({}^{l}\boldsymbol{m}_i(t), {}^{l}\boldsymbol{m}_i(t)), \tag{2.17}$$

where

$$\boldsymbol{A}'({}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}, {}^{lR_D}\boldsymbol{F}, {}^{rR_D}\boldsymbol{F}, {}^{L_DR_D}\boldsymbol{F}) = \begin{bmatrix} {}^{lL_D}f_{31} & {}^{lL_D}f_{32} & 0 & 0 \\ {}^{rL_D}f_{31} & {}^{rL_D}f_{32} & 0 & 0 \\ 0 & 0 & {}^{lR_D}f_{31} & {}^{lR_D}f_{32} \\ 0 & 0 & {}^{rR_D}f_{31} & {}^{rR_D}f_{32} \\ {}^{L_DR_D}f_{31} & {}^{L_DR_D}f_{32} & {}^{L_DR_D}f_{13} & {}^{L_DR_D}f_{23} \end{bmatrix},$$

$$\boldsymbol{B}'({}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}, {}^{lR_D}\boldsymbol{F}, {}^{rR_D}\boldsymbol{F}, {}^{L_DR_D}\boldsymbol{F}) = \begin{bmatrix} {}^{lL_D}f_{13} & {}^{lL_D}f_{23} & 0 & 0 & {}^{lL_D}f_{33} \\ 0 & 0 & {}^{rL_D}f_{13} & {}^{rL_D}f_{23} & {}^{rL_D}f_{33} \\ {}^{lR_D}f_{13} & {}^{lR_D}f_{23} & 0 & 0 & {}^{lR_D}f_{33} \\ 0 & 0 & {}^{rR_D}f_{13} & {}^{rR_D}f_{23} & {}^{rR_D}f_{33} \\ 0 & 0 & 0 & 0 & {}^{L_DR_D}f_{33} \end{bmatrix},$$

$$\boldsymbol{d}'_i({}^{l}\boldsymbol{m}_i(t), {}^{l}\boldsymbol{m}_i(t)) = \begin{bmatrix} {}^{l}\boldsymbol{m}_i(t) \\ {}^{r}\boldsymbol{m}_i(t) \\ 1 \end{bmatrix}, \tag{2.18}$$

and $\boldsymbol{A}'^{+}$ denotes a pseudo-inverse matrix of $\boldsymbol{A}'$.

2.3.2 Demonstrator's view recovery

To perform the view transformation from known views V_L^D (=$\{[l], [r]\}$) to unknown views V_D^D (=$\{[L_D], [R_D]\}$), the learner needs to estimate fundamental matrices, ${}^{lL_D}\boldsymbol{F}$, ${}^{lL_D}\boldsymbol{F}$, ${}^{lL_D}\boldsymbol{F}$, ${}^{lL_D}\boldsymbol{F}$, and ${}^{L_DR_D}\boldsymbol{F}$. As mentioned in the section 2.3.1, sufficient number of matched pairs of the projected points is needed to estimate the fundamental matrices. However, the learner does not know directly the matched point in V_D^D. The basic idea to estimate the fundamental matrices is to utilize its own views to observe its self-body as an alternative to the demonstrator's ones.

Since it is assumed that the learner has the same body structure and the camera parameters as the demonstrator does, and that the initial postures of the both are also the same, the learner's body projected on V_L^L is congruent with the demonstrator's one in V_D^D (see Fig. 2.4). Therefore, instead of using unknown matched points in V_D^D, fundamental matrixes can be estimated by using points which is considered to be matched in V_L^L.

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

$$\begin{aligned} {}^{L_D}\boldsymbol{m}_i &= {}^{L}\boldsymbol{m}_i, \\ {}^{R_D}\boldsymbol{m}_i &= {}^{R}\boldsymbol{m}_i. \end{aligned} \tag{2.19}$$

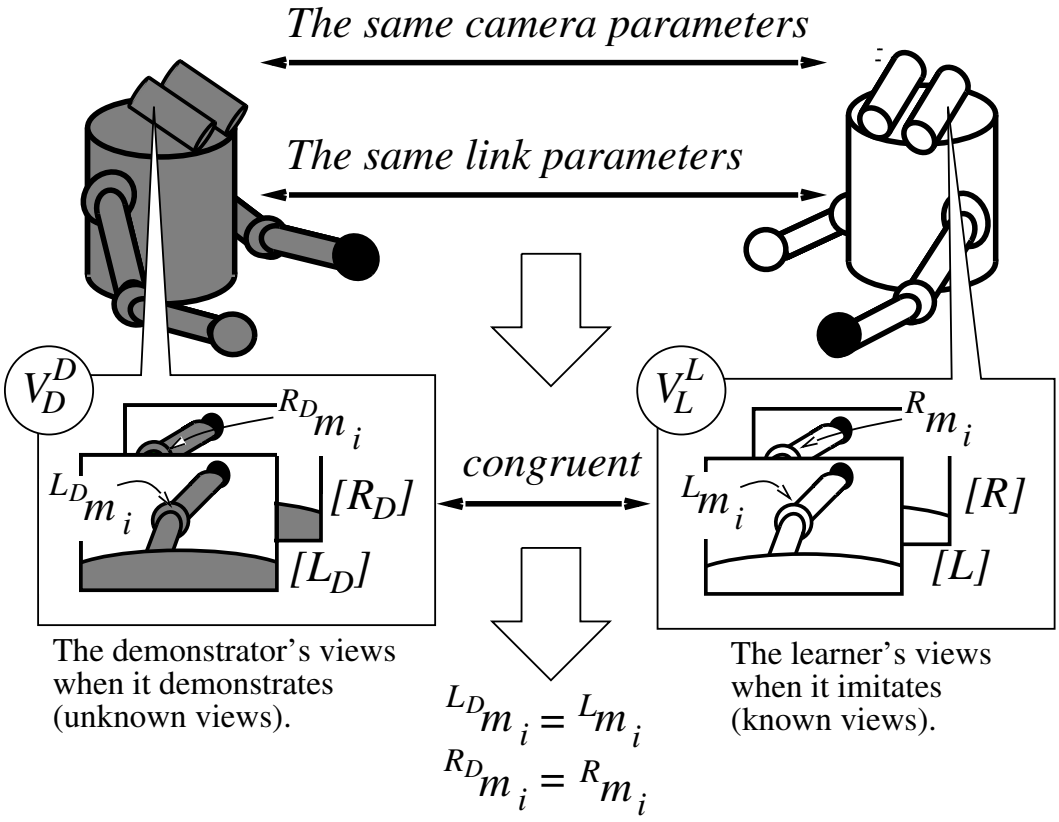


Figure 2.4: The key idea of the method of the parameter estimation. Because it is assumed that the link parameter, camera parameters, and their postures of the learner and the demonstrator are the same, the views observing itself (V_L^L and V_D^D) become congruent with each other. According to the congruence, the image coordinates of matched body parts are the same (i. e. ${}^{L_D}\boldsymbol{m}_i = {}^{L}\boldsymbol{m}_i$, and ${}^{L_D}\boldsymbol{m}_i = {}^{L}\boldsymbol{m}_i$).

It means that the learner can alternate ${}^{L}\boldsymbol{m}_i$ and ${}^{R}\boldsymbol{m}_i$ with ${}^{L_D}\boldsymbol{m}_i$ and ${}^{R_D}\boldsymbol{m}_i$, in order to estimate the fundamental matrices by the method mentioned in the section 2.3.1.

Recovery of the demonstration

Then, the method to recover the trajectory of the demonstration in unknown views is shown. Suppose that the learner observes a trajectory of the demonstrator's end-effector ${}^{l}\boldsymbol{m}_e(t)$ and ${}^{r}\boldsymbol{m}_e(t)$ in V_L^D, and fundamental matrices of five epipolar equations, ${}^{lL_D}\boldsymbol{F}$, ${}^{rL_D}\boldsymbol{F}$, ${}^{lR_D}\boldsymbol{F}$, ${}^{rR_D}\boldsymbol{F}$, and ${}^{L_DR_D}\boldsymbol{F}$ have been already estimated. Utilizing view transformation mechanism mentioned in the previous section, the trajectory of the end-effector ${}^{L_D}\boldsymbol{m}_e(t), {}^{R_D}\boldsymbol{m}_e(t)$ in V_D^D are estimated such as,

$$\begin{bmatrix} {}^{L}\boldsymbol{m}_e \\ {}^{R}\boldsymbol{m}_e \end{bmatrix} = \boldsymbol{A}'({}^{l}\boldsymbol{m}_e(t), {}^{l}\boldsymbol{m}_e(t), {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}, {}^{lR_D}\boldsymbol{F}, {}^{rR_D}\boldsymbol{F}, {}^{L_DR_D}\boldsymbol{F})^{+} \cdot \boldsymbol{d}'({}^{l}\boldsymbol{m}_e(t), {}^{l}\boldsymbol{m}_e(t), {}^{lL_D}\boldsymbol{F}, {}^{rL_D}\boldsymbol{F}, {}^{lR_D}\boldsymbol{F}, {}^{rR_D}\boldsymbol{F}, {}^{L_DR_D}\boldsymbol{F}), \tag{2.20}$$

Recall that the demonstrator has the same body structure and the same stereo cameras, and that V_D^D are the demonstrator's views when it observes itself during demonstrating in the same posture as the learner observes itself. Therefore, if the learner performs the same motion as the demonstration, the trajectory of the learner's end-effector in V_L^L can be congruent with one of the demonstrator's in V_D^D. Thus, the leaner can reproduce the demonstration by following its end-effector to the recovered trajectory ${}^{L_D}\boldsymbol{m}_e(t), {}^{R_D}\boldsymbol{m}_e(t)$ as the desired one in V_L^L. We call this processes, that is recovering the demonstrator's trajectory in the demonstrator's view, as *demonstrator's view recovery.*

2.3.3 Performing the recovered trajectory by adaptive visual servoing

The control system to perform the recovered trajectory in V_L^L is constructed based on adaptive visual servoing (AVS) [31]. In AVS a feature vector can be controlled with online estimation of a Jacobian matrix of time-derivatives of the quantities of image features with respect to joint angle velocities.

Suppose that $\boldsymbol{\theta} \in \Re^m$ denotes the learner's joint angle, and ${}^{LR}\boldsymbol{x}_e = [{}^{L}\boldsymbol{m}_e^T, {}^{R}\boldsymbol{m}_e^T]^T \in \Re^4$ denotes the image feature vector of the learner's end-effector. If the relation between ${}^{LR}\boldsymbol{x}_i$ and $\boldsymbol{\theta}$ is given by ${}^{LR}\boldsymbol{x}_e = {}^{LR}\boldsymbol{x}_e(\boldsymbol{\theta})$, differentiating it, we obtain a velocity relation such as

$${}^{LR}\dot{\boldsymbol{x}}_e = \boldsymbol{J}(\boldsymbol{\theta})\dot{\boldsymbol{\theta}}, \tag{2.21}$$

where $\boldsymbol{J}(\boldsymbol{\theta}) = (\partial {}^{LR}\boldsymbol{x}_i / \partial \boldsymbol{\theta})^T \in \Re^{4\times m}$ is a Jacobian matrix of time-derivatives of the quantities of image features with respect to those of the joint angles. Assuming that movement of the camera-manipulator system is slow enough to consider the Jacobian matrix $\boldsymbol{J}$ to be constant during the sampling time, we obtain

$$ {}^{LR}\boldsymbol{x}_e(k+1) = {}^{LR}\boldsymbol{x}_e(k) + \boldsymbol{J}(k)\boldsymbol{u}(k), \tag{2.22}$$

as a discrete model of the system, where $\boldsymbol{J}(k)$ and $\boldsymbol{u}(k)(= \dot{\boldsymbol{\theta}}\Delta T)$ denote the constant Jacobian matrix and a control input vector in the k-th step during sampling rate ΔT, respectively.

From eq. (2.22), recurrence formula to estimate $\boldsymbol{J}$ can be derived such as,

$$ \hat{\boldsymbol{j}}_i^T(k+1) - \hat{\boldsymbol{j}}_i^T(k) = \frac{\{{}^{LR}\boldsymbol{x}_e(k+1)_i - {}^{LR}\boldsymbol{x}_e(k)_i - \hat{\boldsymbol{j}}(k)_i^T\boldsymbol{u}(k)\}}{\rho + \boldsymbol{u}^T(k)\boldsymbol{P}(k)\boldsymbol{u}(k)}\boldsymbol{P}(k)\boldsymbol{u}(k), \tag{2.23}$$

where $\boldsymbol{j}_i^T$ is the i-th row vector of $\boldsymbol{J}$, ρ is a forgetting factor in the range $0 < \rho \leq 1$, and $\boldsymbol{P}$ is a covariance matrix which is also estimated in recurrence formula such as

$$ \boldsymbol{P}(k) = \frac{1}{\rho}\left[\boldsymbol{P}(k-1) - \frac{\boldsymbol{P}(k-1)\boldsymbol{u}(k-1)\boldsymbol{u}(k-1)^T\boldsymbol{P}(k-1)}{\rho + \boldsymbol{u}(k-1)^T\boldsymbol{P}(k-1)\boldsymbol{u}(k-1)}\right]. \tag{2.24}$$

The recurrence formula to estimate $\boldsymbol{J}$ and $\boldsymbol{P}$ is a solution based on recursive least-mean-square method which minimizes the weighted residual of eq. (2.22).

With online estimation of $\boldsymbol{J}(k)$ in eq. (2.23) and of $\boldsymbol{P}(k)$ in eq. (2.24), the desired trajectory of feature points ${}^{LR}\boldsymbol{x}_{ed}(t)$ is realized by the control law,

$$\begin{aligned} \boldsymbol{u}(k) &= \hat{\boldsymbol{J}}(k)^+\{{}^{LR}\boldsymbol{x}_{ed}(k+1) - {}^{LR}\boldsymbol{x}_{ed}(k)\} \\ &\quad +\{\boldsymbol{I}_m - \hat{\boldsymbol{J}}(k)^+\hat{\boldsymbol{J}}(k)\}\boldsymbol{k}_r \\ &\quad +\boldsymbol{K}\hat{\boldsymbol{J}}(k)^T\{{}^{LR}\boldsymbol{x}_{ed}(k+1) - {}^{LR}\boldsymbol{x}_e(k)\}, \end{aligned} \tag{2.25}$$

where $\hat{\boldsymbol{J}}(k)^+$, $\boldsymbol{I}_m$, $\boldsymbol{k}_r$ and $\boldsymbol{K}$ denote a pseudo-inverse matrix of $\hat{\boldsymbol{J}}(k)$, an $m \times m$ identity matrix, an arbitrary vector, and a positive-definite gain matrix, respectively.

2.3.4 An overview of process in imitation

Finally, an overview of the imitation process based on demonstrator's view recovery is shown in Fig. 2.5. The process consists of

1. estimating the parameters of view transformation from the learner's views V_L^D to the demonstrator's views V_D^D. The estimation is realized by using image points in V_L^L which are supposed to correspond to those in V_D^D,
2. recovering the demonstrator's trajectory of the end-effector in V_D^D by the view transformation in eq. (2.20), and
3. controlling the learner's end-effector to follow the recovered trajectory of the demonstration by adaptive visual servoing system in eq. (2.25).

2.4 Experiment

To consider the validity of the proposed method, real robot experiment was conducted. Two identical manipulators were assumed to be bodies of the learner and the demonstrator, respectively. The learner possess two cameras. (see Fig. 2.6).

An experimental setup

Two identical manipulators (PA10, MHI) were used as the bodies of both the learner and the demonstrator, each of which has 7 DoFs and 2 of them were used in this experiment. The stereo cameras (CCB-EX37, SONY) on the movable camera head located near the learner's body, assuming they were the learner's head and eyes and the manipulator as the learner's arm (see Fig. 2.6 and 2.7(a)).

It obtained two video images through the stereo cameras, and then, they were combined to one (image size: 640[pixel] $\times$ 480[pixel]) by compressing each images into the half along the vertical axis (640[pixel] $\times$ 240[pixel]) in video mixing device [45], and then sent to a tracking module equipped with a high-speed correlation processor utilizing SAD (Sum of Absolute Difference) manufactured by Fujitsu. Before starting an experiment, we specified target images to be tracked by the model. During the experiment, the module tracked the target images, and it fed the image coordinates of the targets to CPU. The CPU calculated a desired joints angle velocity of the manipulator and sent it to the manipulator controller through real-time network (ARCNET, 5.0 Mbps). Using this experimental equipment and writing programs using C language on VxWorks (Wind River), the sampling ratio was 30[Hz]. Note that a different set of the CPU and the manipulator controller was assigned to control the demonstrator robot motion.

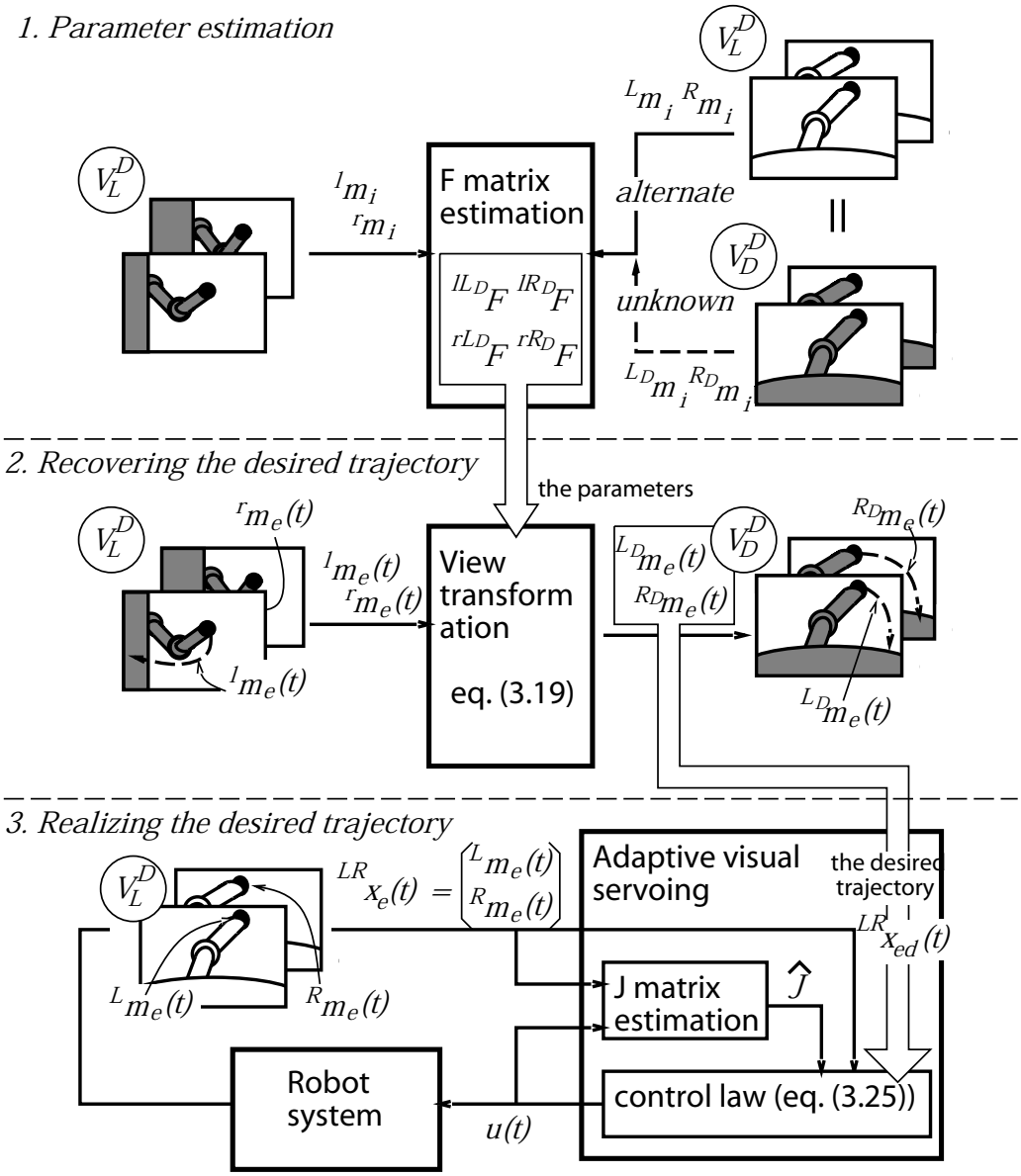


Figure 2.5: An overview of the proposed mechanism of imitation based on demonstrator's view recovery.

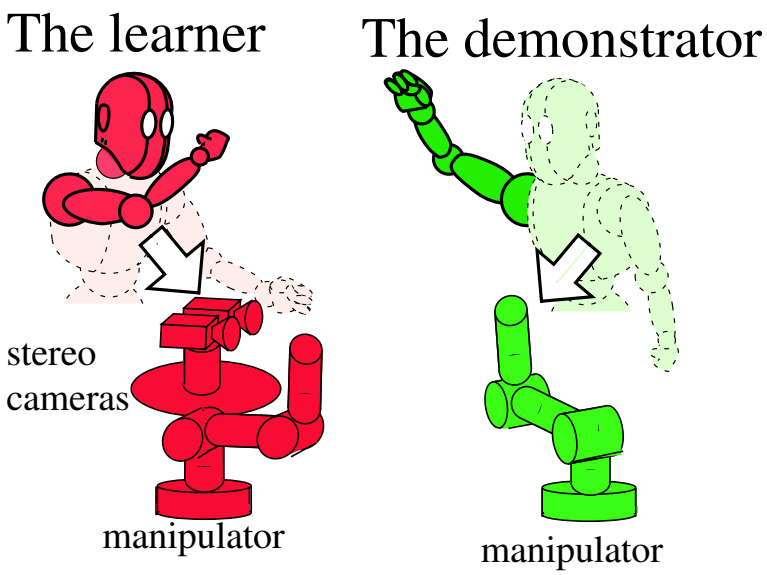


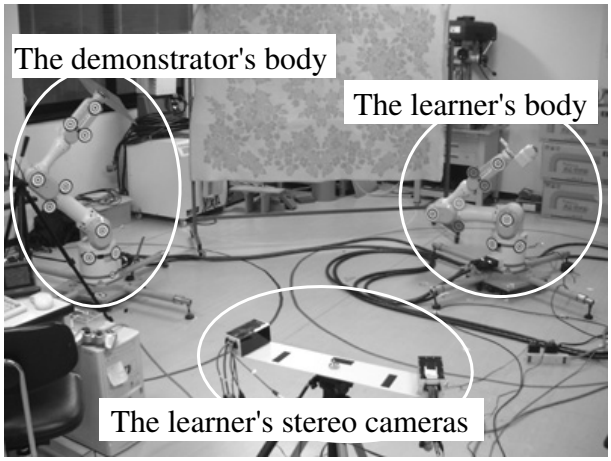
Figure 2.6: Assuming the learner and the demonstrator as two identical manipulators.

An trial of Imitation

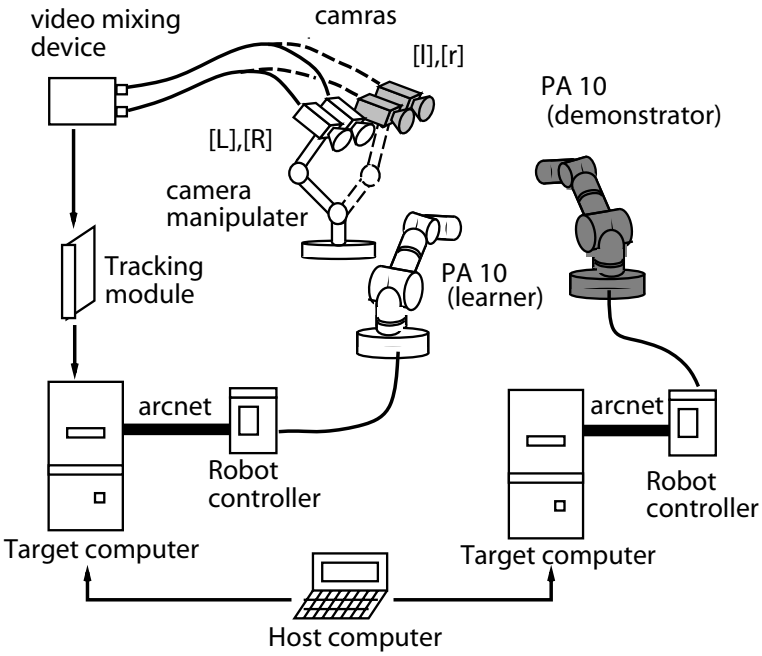
Fig. 2.8 illustrates the process of an trial of imitation. After the parameters of epipolar geometry were estimated, the demonstrator demonstrated the motion that shaped a triangle by its end-effector. The demonstration was performed by sending the series of joint angle velocities to the controller of the demonstrator. At first, the learner detected the end-effector trajectory of the demonstration in its view (V_L^D) by using the tracking module (see Fig. 2.9). Then, it generated the desired trajectory (dots) in its view (V_L^L) by the proposed method of demonstrator's view recovery (see Fig. 2.10), and finally reproduced it by the adaptive visual servoing (see Fig. 2.11).

Fig. 2.10 also shows the true one (boxes) which is measured in advance by sending the same series of those as the demonstrator's one to the controller of the learner. The proposed method of the view transformation works well since the recovered one and the true one are almost the same.

To confirm whether the adaptive visual servoing system to follow the recovered trajectory works well, Fig. 2.12 shows three trajectories of the end-effector in the learner's view (V_L^L), which are the performed one, the true one, and the recovered one, respectively. It is confirmed that the learner performs imitation

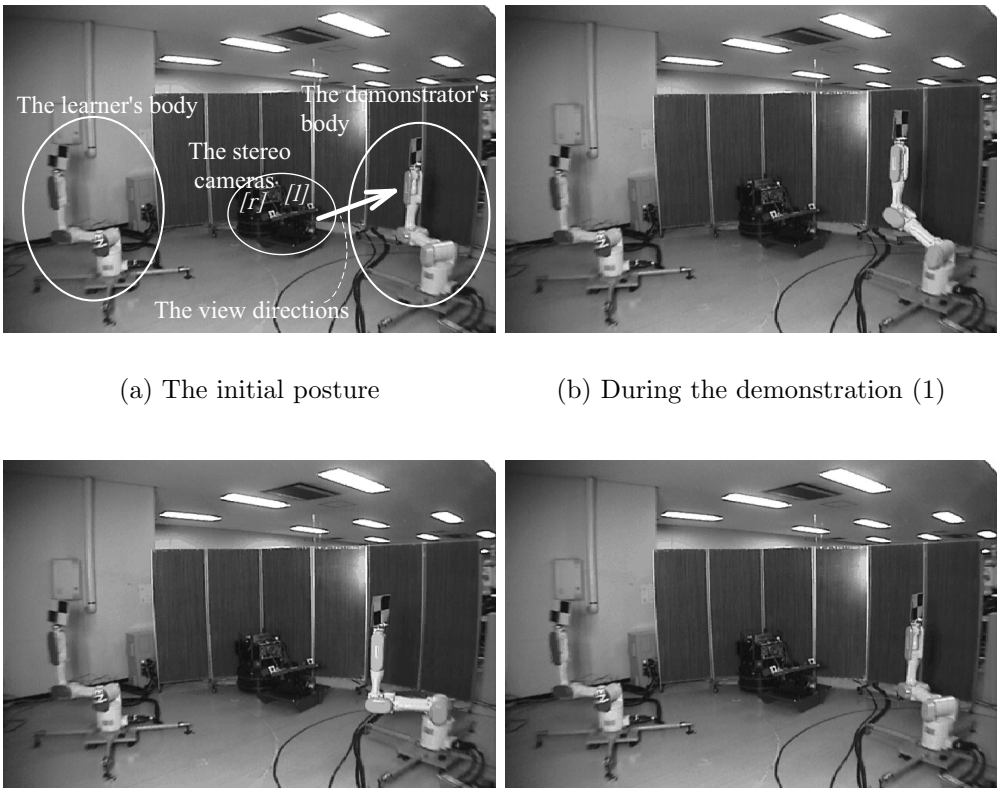


(a) An appearance of the experimental setup



(b) A network of information in experimental system

Figure 2.7: An overview of the experimental setup



(a) The initial posture

(b) During the demonstration (1)

(c) During the demonstration (2)

(d) The final posture

Figure 2.8: Appearances of the demonstrator's motion.

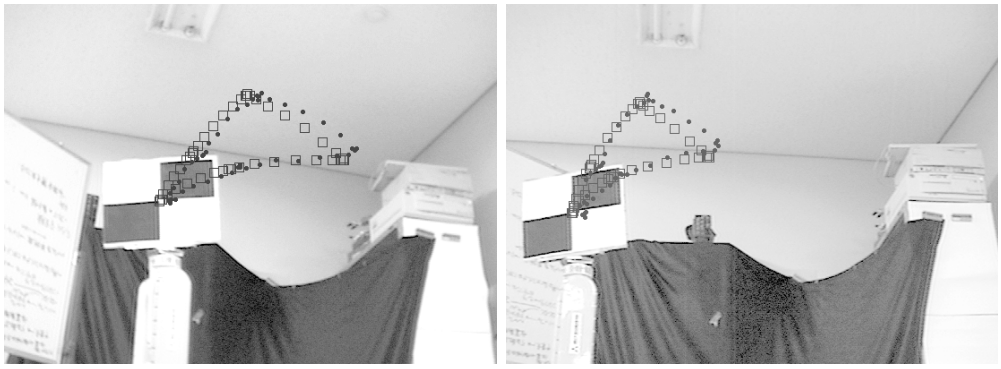
well since they are almost the same.

2.5 Estimation of Fundamental Matrixes by Conflict Resolution with Epipolar Geometry

To recover the demonstrator's view, we need the true fundamental matrices between the views. However, when the learner's posture is different from the demonstrator's one, we cannot estimate them by the proposed method in section 2.3.1. In this section, the method to release the assumption of the same

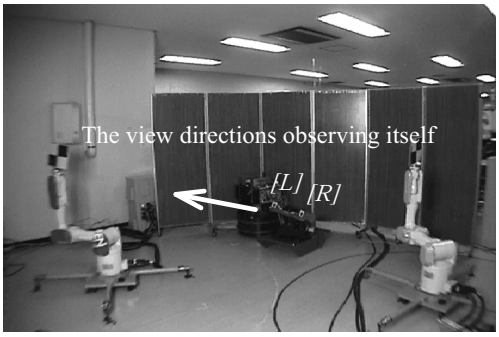
(a) [l] (b) [r]

Figure 2.9: The trajectories of the demonstration in the learner's views observing the demonstrator (V_L^D).



(a) [L] (b) [R]

Figure 2.10: The recovered trajectories (dots) and true trajectories (boxes) in the learner's views observing itself (V_L^L).



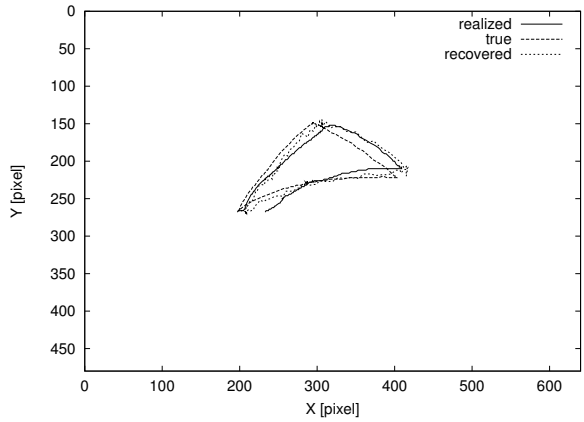
(a) The initial posture

(b) During the imitation (1)

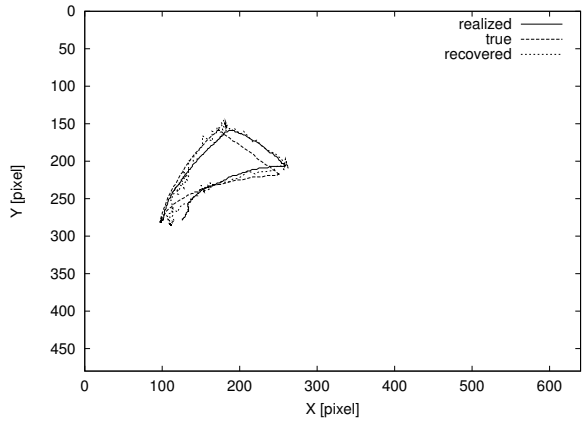
(c) During the imitation (2)

(d) The final posture

Figure 2.11: Appearances of the learner's imitation behavior.



(a) [L]



(b) [R]

Figure 2.12: The trajectories in imitation. The performed one, the true one measured in advance, and the recovered one.

posture is given.

If the fundamental matrices are correctly estimated, the recovered points of the demonstrator's body parts are coincident with the corresponding learner's body parts. Else, the corresponding points are shifted from each other on the image plane since the estimated fundamental matrices conflict with true epipolar geometry. In this section, we derive the evaluation function of this conflict, analyze its behavior based on the affine camera model, and propose the method to resolve the conflict by the control to minimize the evaluation function.

2.5.1 An evaluation function

Suppose that the learner observes the demonstrator's posture and knows N image coordinates of the feature points on the demonstrator's body ${}^{l}\boldsymbol{m}_i$ and ${}^{r}\boldsymbol{m}_i$, $(i = 1, \cdots, N)$ in its stereo views V_L^D. N should be more than 4 which is minimum number to estimate a fundamental matrix of the affine cameras. Then, the learner changes the gaze direction and watches at its own body to know the image coordinate of body part those correspond to ${}^{L}\boldsymbol{m}_i$ and ${}^{R}\boldsymbol{m}_i$.

Confliction of the estimated fundamental matrices with epipolar geometry The fundamental matrices estimated by the method in the previous section are no longer true since both initial postures are different and consequently eq. (2.19) is not satisfied. Therefore, if we use the false fundamental matrices for view transformation, the recovered views are shifted from the true ones of the demonstrator. In other words, image coordinates of the recovered points on the demonstrator's body parts (${}^{L_D}\hat{\boldsymbol{m}}_i$ and ${}^{R_D}\hat{\boldsymbol{m}}_i$) are shifted from the true ones (${}^{L_D}\boldsymbol{m}_i$ and ${}^{R_D}\boldsymbol{m}_i$).

Furthermore, they are also shifted from the projected points of the learner's body parts (${}^{L}\boldsymbol{m}_i$ and ${}^{R}\boldsymbol{m}_i$) which correspond to demonstrator's ones (${}^{L_D}\boldsymbol{m}_i$ and ${}^{R_D}\boldsymbol{m}_i$). It might be felt strange because the estimation minimizes the sum of residuals of epipolar equations between $\{{}^{l}\boldsymbol{m}_i$ or ${}^{r}\boldsymbol{m}_i\}$ and $\{{}^{L}\boldsymbol{m}_i$ or ${}^{R}\boldsymbol{m}_i\}$. However, it is true because the sum of redisuals cannot become zero in this case since we assume that the number of points (N) which is used for estimation is more than the minimum number (4) for it. Therefore, the transformations using the estimated parameters with non-zero residuals are not consistent with data used for estimation (see Fig. 2.13).

The evaluation function of the transforming error caused by the inaccuracy of the estimated fundamental matrices when the learner's posture is different

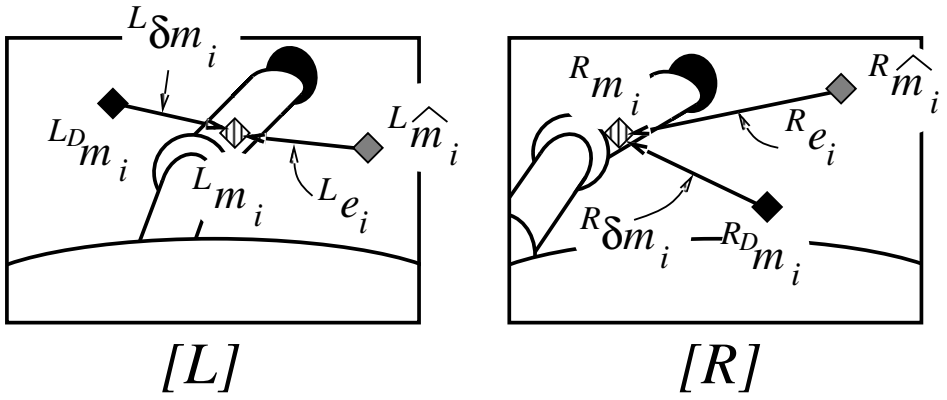


Figure 2.13: An overview of confliction.

from the demonstrator's one is defined as

$$E = \frac{1}{N}\sum_{i=1}^{N} {}^{LR}\boldsymbol{e}_i^T {}^{LR}\boldsymbol{e}_i, \tag{2.26}$$

where

$${}^{LR}\boldsymbol{e}_i = \begin{bmatrix} {}^{L}\boldsymbol{m}_i \\ {}^{R}\boldsymbol{m}_i \end{bmatrix} - \begin{bmatrix} {}^{L_D}\hat{\boldsymbol{m}}_i \\ {}^{R_D}\hat{\boldsymbol{m}}_i \end{bmatrix} \tag{2.27}$$

denotes a vector from a recovered point to the corresponding body part of the learner (see Fig. 2.13). Note that this evaluation function can be calculated from the learner's own viewpoint.

2.5.2 Analyzing the evaluation function

Since the projected points ${}^{LR}\boldsymbol{x}_i$ of the learner's body parts are the functions in terms of the learner's posture ($\boldsymbol{\theta} \in \Re^m$), they can be represented by ${}^{LR}\boldsymbol{x}_i = {}^{LR}\boldsymbol{x}_i(\boldsymbol{\theta})$. Supposing that the learner's posture is given by $\boldsymbol{\theta} = \boldsymbol{\theta}_D + \boldsymbol{\delta\theta}$, where $\boldsymbol{\theta}_D$ is the posture of the demonstrator, the projected point ${}^{LR}\boldsymbol{x}_i$ is given by the following equation including the perturbation ${}^{LR}\boldsymbol{\delta x}_i$,

$${}^{LR}\boldsymbol{x}_i = {}^{LR}\boldsymbol{x}_{Di} + {}^{LR}\boldsymbol{\delta x}_i, \tag{2.28}$$

where ${}^{LR}\boldsymbol{x}_{Di} = {}^{LR}\boldsymbol{x}_i(\boldsymbol{\theta}_D) = [{}^{L}\boldsymbol{m}_{Di}^T, {}^{R}\boldsymbol{m}_{Di}^T]^T$.

As long as ${}^{LR}\boldsymbol{\delta x}_i$ is small, the relationship between $\boldsymbol{\theta}$ and ${}^{LR}\boldsymbol{x}_i$ can be approximated by

$${}^{LR}\boldsymbol{\delta x}_i = \boldsymbol{J}_{\boldsymbol{x}i}\boldsymbol{\delta\theta}, \tag{2.29}$$

where $\boldsymbol{J}_{\boldsymbol{x}i} = (\partial {}^{LR}\boldsymbol{x}_i / \partial \boldsymbol{\theta})^T \in \Re^{4\times m}$ is a Jacobian matrix of time-derivatives of the projected point vector with respect to the joint angle.

Based on the perturbation theory for eigenvalue, the estimated fundamental matrix ${}^{pq}\boldsymbol{f}$ is given by the following equation including a function of perturbation ${}^{pq}\boldsymbol{\delta f}$ in terms of ${}^{LR}\boldsymbol{\delta x}_i (i = 1, \cdots, n)$,

$$ {}^{pq}\boldsymbol{f} = {}^{pq}\boldsymbol{f}_{true} + {}^{pq}\boldsymbol{\delta f}({}^{LR}\boldsymbol{\delta x}_1, \cdots, {}^{LR}\boldsymbol{\delta x}_N), \tag{2.30} $$

where ${}^{pq}\boldsymbol{f}_{true}$ is the true fundamental matrix. From eq. (2.29), the second term of right-hand side in eq. (2.30) is given by a function in terms of $\boldsymbol{\delta\theta}$, such as ${}^{pq}\boldsymbol{g}(\boldsymbol{\delta\theta})$. Therefore ${}^{pq}\boldsymbol{f}$ is approximated by including it,

$$ {}^{pq}\boldsymbol{f} = {}^{pq}\boldsymbol{f}_{true} + {}^{pq}\boldsymbol{g}(\boldsymbol{\delta\theta}) \tag{2.31} $$

Substituting eq. (2.31) to eq. (2.17), and developing the above formulation algebraically in focusing on the dominant term, we obtain

$$ E = \boldsymbol{\delta\theta}^T \boldsymbol{Q} \boldsymbol{\delta\theta}, \tag{2.32} $$

where $\boldsymbol{Q} \in \Re^{m\times m}$ is a positive-semidefinite matrix.

The positive-semidefinite matrix $\boldsymbol{Q}$ can be regarded as a positive-definite matrix because the evaluation function usually becomes zero only when the learner's posture corresponds to the demonstrator's one. Thus, the proposed evaluation function is expected to be a convex function and have a local minimum at which the learner's posture corresponds to the demonstrator's one.

2.5.3 Control to resolve the confliction of the estimation

In this section, a control system to resolve the confliction of estimation by minimizing the evaluation function is proposed. Since it can be regarded as a convex function in terms of joint angles, it is expected to be able to minimize the proposed evaluation function by a gradient method. In order to use the gradient method, we need the gradient vector, which is unknown in our case. Then again, adaptive visual servoing method (AVS) [31] is applied to estimate the gradient vector of the unknown system.

The relation between the learner's joint angle $\boldsymbol{\theta}$ and the evaluation function E is given by $E = E(\boldsymbol{\theta})$ from eq. (2.32). Differentiating it, we obtain a velocity relation,

$$ \dot{E} = \boldsymbol{J}_E \dot{\boldsymbol{\theta}}, \tag{2.33} $$

where $\boldsymbol{J_E}$ is a Jacobian matrix of time-derivatives of the evaluation function with respect to joint angle velocity. Using AVS, the Jacobian matrix $\hat{\boldsymbol{J}}_E$ can be estimated based on the recursive weighted least square method as introduced in section 2.3.3.

In order to minimize E, we can determine the control input vector $\boldsymbol{u}_{input} \in \Re^m$ in the following equation,

$$\boldsymbol{u}_{input} = -\boldsymbol{K}\hat{\boldsymbol{J}}_E^T E \tag{2.34}$$

where $\boldsymbol{K} \in \Re^{m \times m}$ is a positive-definite gain matrix.

2.5.4 Overview of process in imitation

In this section, an overview of the imitation process based on the demonstrator's view recovery in the situation when the learner's posture is different from the demonstrator's one is shown (see Fig. 2.14). Note that only estimating process is different from previous method (see Fig. 2.5). In the process, the learner

1. estimates the parameters of view transformation by iterating the following processes;
   (a) estimating them by regarding the learner's initial posture is the same as the demonstrator's even though they are actually different and by applying the method in the section 2.3.2,
   (b) recovering the demonstrator's body parts in V_D^D by the view transformation in eq. (2.20) with the estimated parameters and calculating the evaluation function in eq. (2.26),
   (c) controlling the learner's joint angles to minimize the evaluation function by the control law in eq. (2.34) with online estimation of Jacobian matrix,
   (d) judging the termination of the control by checking whether the evaluation function is converged to smaller value than a threshold determined in advance, if not, return to (1a),
2. after the evaluation function is converged, recovers the demonstrator's trajectory of the end-effector in V_D^D by the view transformation in eq. (2.20), and then,
3. controls the learner's end-effector to follow the recovered trajectory of the demonstration by adaptive visual servoing system in eq. (2.34).

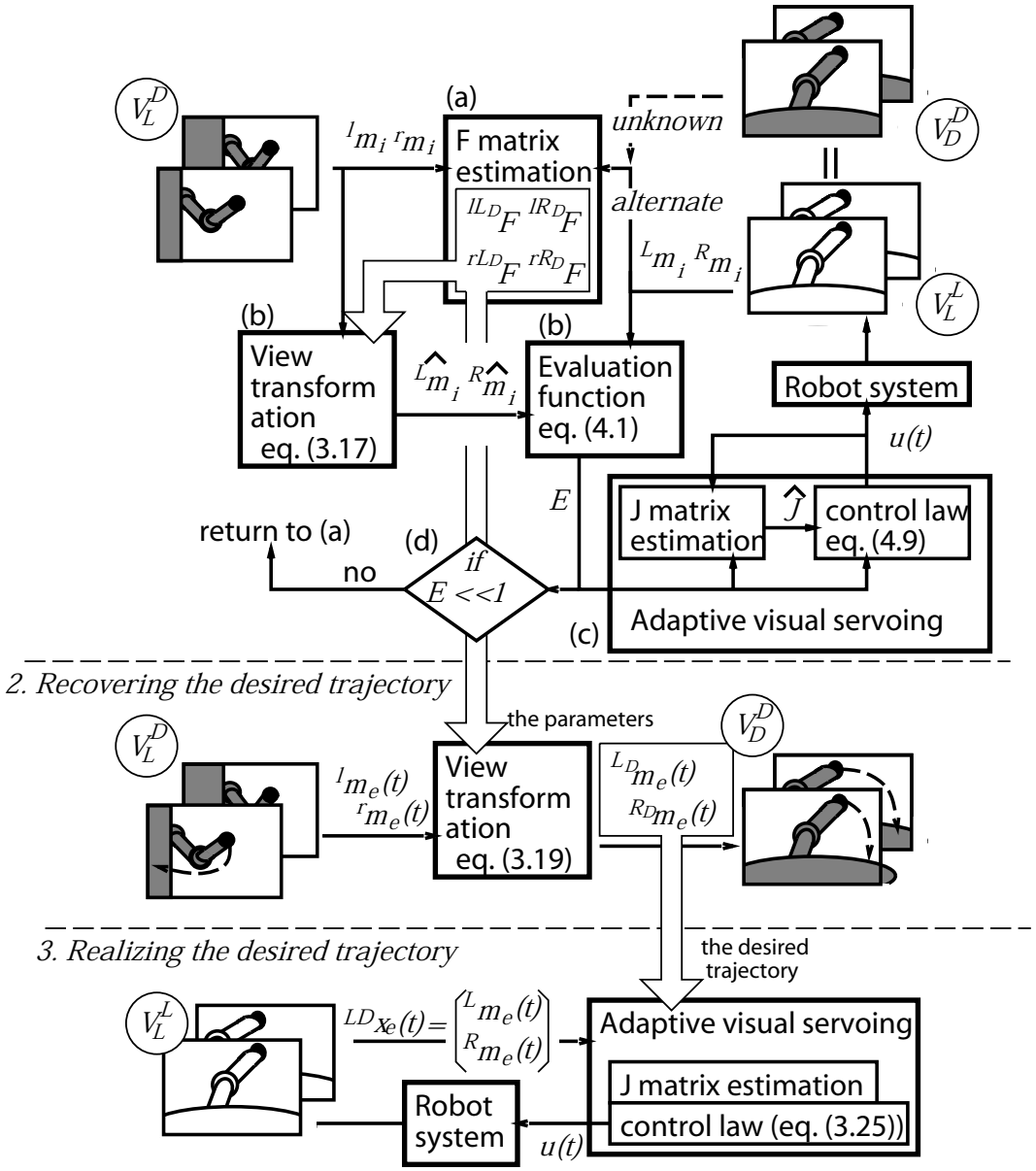


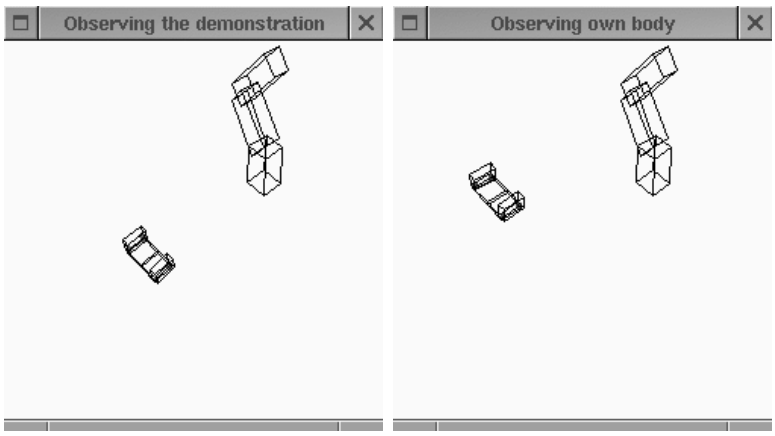
Figure 2.14: An overview of the mechanism of the imitation in the situation where the demonstrator's initial posture is different from the learner's one.

2.6 Experiments

To show the validity of the proposed method, some experimental results are given in this section. As well as in the previous experiments (section 2.4), two identical manipulators are assumed as bodies of the learner and the demonstrator, respectively. A pair of focal points of the stereo cameras are assumed as learner's view points (see Fig. 2.6).

2.6.1 Behaviors of the evaluation function by computer simulation

To examine the behavior of the evaluation function, the computer simulation is performed. A manipulator, which has the same link parameters as the one used in the real robot experiments mentioned later, is simulated as the bodies of the learner and the demonstrator. The camera projection is modeled by a pin hole camera. The pair of stereo cameras locates near the learner's body, assuming it is its head and the manipulator as its arm (see Fig. 2.6 and 2.15). An overview of the simulation is shown (see Fig. 2.15)



(a) Observing the demonstrator

(b) Observing its own body

Figure 2.15: An overview of the computer simulation. The position and orientation of the learner's stereo cameras relative to the demonstrator's body, when it observes the demonstration (a), and when it imitates (b).

When the demonstrator's joint angle are $(\theta_2, \theta_4) = (45°, -90°)$, the learner observes the demonstrator's posture and detects the ten feature points ($N = 10$) on the demonstrator's body in its stereo views (see Fig. 2.15(a)). Then the learner changes the gaze direction and watches its own matched feature points (see Fig. 2.15(b)). Fig. 2.16 shows the calculated evaluation function when moving its joint angles, $\theta_2 = 0° \sim 90°$, and $\theta_4 = 0° \sim -180°$.

The calculated evaluation function seems almost a convex one which has a local minimum at $(\theta_2, \theta_4) = (41°, -90°)$. The local minimum is close to the point which the learner's posture corresponds to demonstrator's one.

2.6.2 Experiments using a real robot

To examine the behavior of the evaluation function and to show a validity of proposed control system, the results in the real experiments are shown in this section. They are examined in the same experimental setup as used in section 2.4 (see Fig. 2.7).

(a) Behaviors of the evaluation function in a real experiment

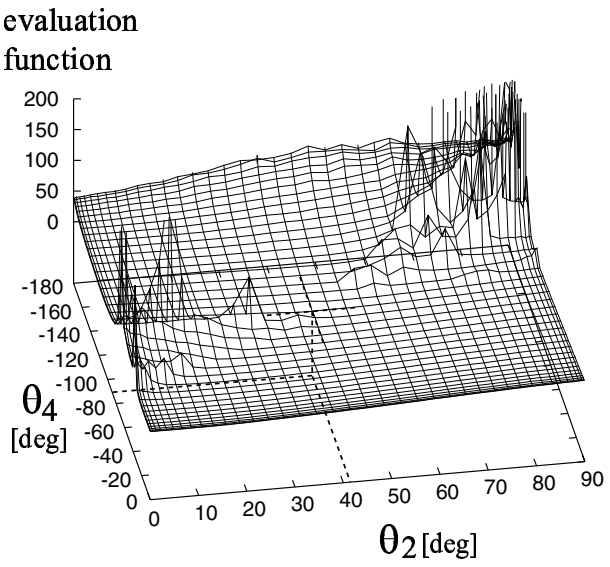
Fig. 2.17 shows the calculated evaluation function in the same manner as in the computer simulation. In the real experiment, the learner moves its joint angles, $\theta_2 = 25°, 35°, 45°, 55°, 65°$ and $75°$, and $\theta_4 = -25° \sim -135°$. Similar to the computer simulation, the calculated evaluation function can be regarded as a convex one which has a local minimum at $(\theta_2, \theta_4) = (45°, -86°)$.

From the results of the computer simulation and the real experiment, we may conclude that the proposed evaluation function is a convex one and has a local minimum at which the learner's posture corresponds to the demonstrator's one.

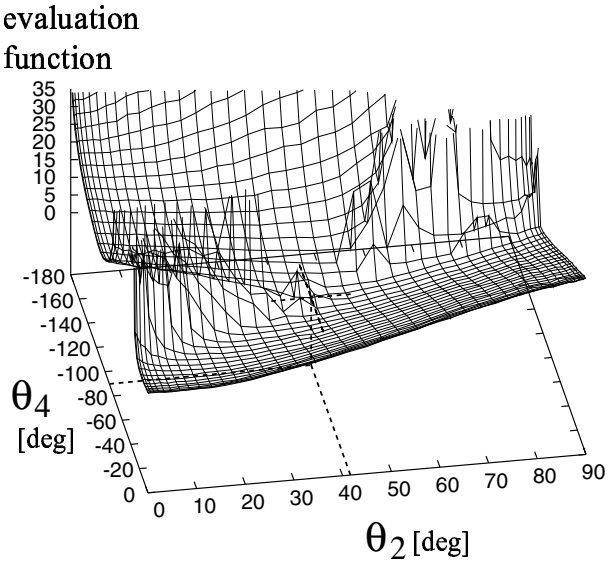
(b) Posture imitation by resolving conflict

To show the validity of the method to resolve the conflict of the estimated affine fundamental matrix with epipolar geometry, the experimental result with real robots is shown in this section. Adaptive visual servoing is applied as one of a gradient method to estimate the true epipolar geometry and then to perform the imitation. The control input is determined by eq. (2.34). The initial values of the Jacobian matrix to be estimated are arbitrarily chosen as

$$\hat{\boldsymbol{J}}_E = \begin{bmatrix} -1.0 & 1.0 \end{bmatrix}, \tag{2.35}$$

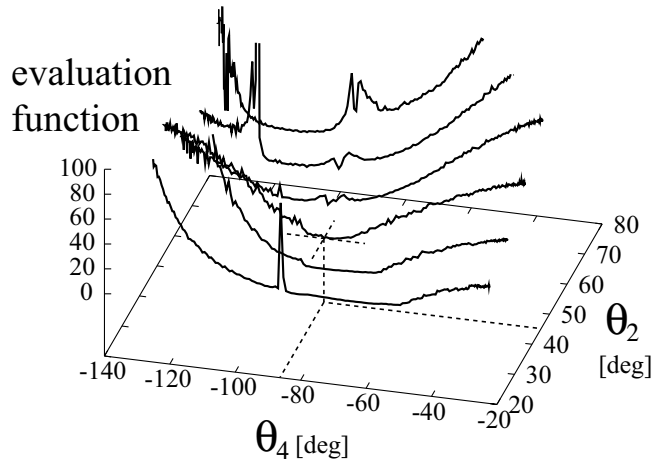


(a) Global view

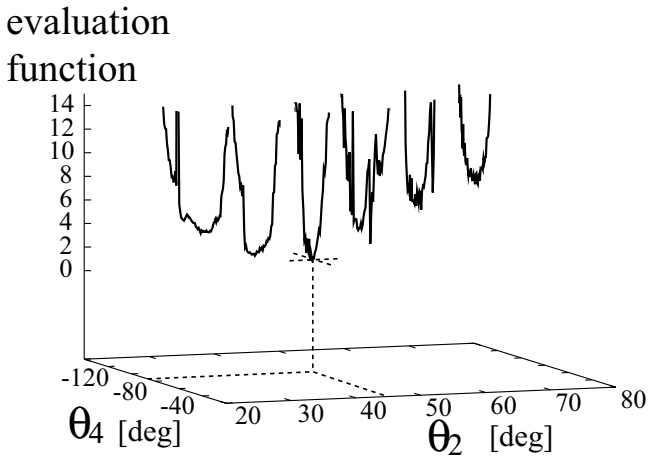


(b) Close up

Figure 2.16: Evaluation function in the computer simulation; global view (left) and close up (right) of the evaluation function. It can be regard that it is almost a convex function which has a local minimum at $(\theta_2, \theta_4) = (41°, -90°)$.



(a) Global view



(b) Close up

Figure 2.17: An evaluation function in the real experiment; a global view (left) and its close up (right) of the evaluation function. It can be regard that it is almost a convex function which has a local minimum at $(\theta_2, \theta_4) = (45°, -86°)$.

and forgetting factor ρ and the positive-definite gain matrix $\boldsymbol{K}$ are 0.95, and $\mathrm{diag}(0.5 \times 10^{-2}, 0.5 \times 10^{-2})$, respectively.

After observing ten feature points on the demonstrator's body, the learner minimizes the evaluation function by using AVS. Fig. 2.18 (a) shows the trajectories of two joint angles θ_2 and θ_4, respectively, during the control. They are evidently converged to the demonstrator's posture (broken lines). Fig. 2.18 (b) shows that the evaluation function also converges to zero by the method.

(c) Trajectory imitation based on estimated epipolar geometry

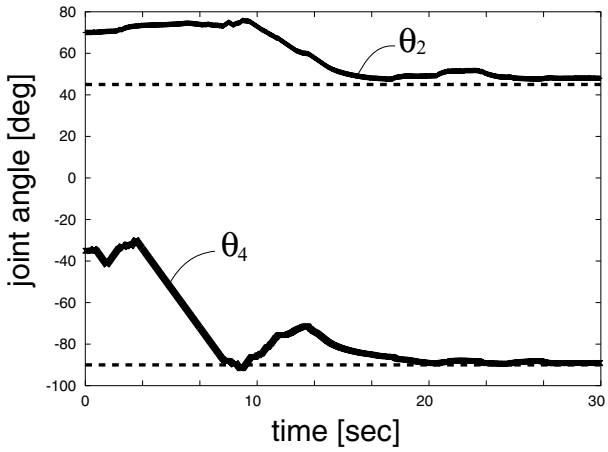
To test the estimated affine fundamental matrix through minimization process in the real experiment, the learner imitates the demonstration by using the estimated parameters. The learner stores the trajectory of the demonstrator's end-effector. Then, it recovers the observed trajectory on its view of monitoring the self motion using the method described in section 2.3, and the learner imitates by reproducing the recovered trajectory based on feedback control using AVS.

Fig. 2.19 shows the trajectories in the learner's view. Each figure contains two trajectories; one is reproduced by the learner's and the other is the true one. Since two trajectories are almost overlapped with each other, we might conclude that the estimation of affine epipolar geometry is sufficient to imitate the demonstrator's motion.

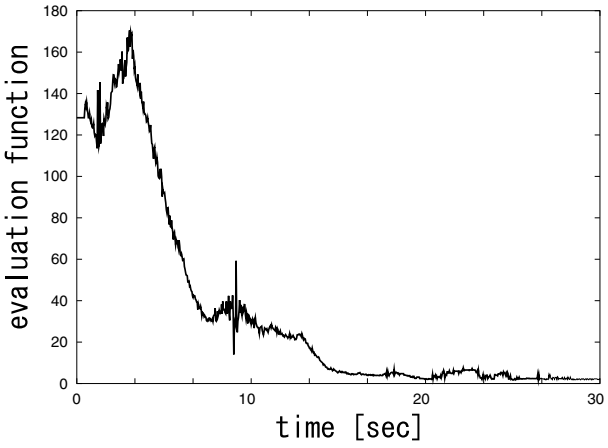
2.7 Summary and discussion

In this chapter, the methods based on the paradigm what we call demonstrator's view recovery were introduced to perfrorm imitation from subjective viewpoint. By the proposed methods, the learner robot could reproduce the demonstration of another robot that has the same body structure as the learner. In this section, we discuss the future works on demonstrator's view reovery.

Abstracting and segmenting the reproduced motion By the proposed method, the learner could just reproduce the demonstration. However, the reproduced motion should be abstracted or segmented into some primitive motions as addressed in some previous work (ex. [8, 12, 34, 35]). Such process might help to optimize the acquired behavior. Furthermore, the acquired representation of motion in this process might be utilized to recognize other demonstration by using it as primitive of recognition.

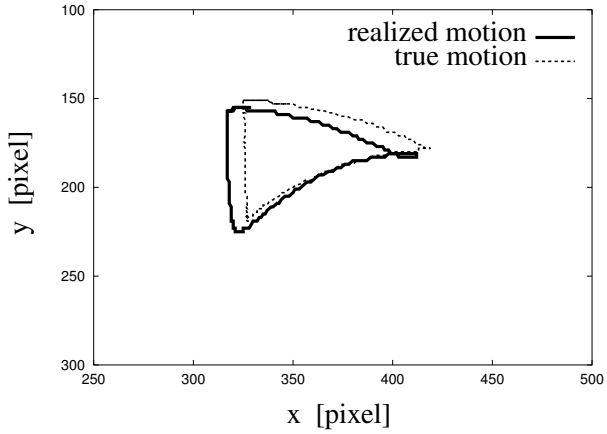


(a) The changes of the learner's joint angles

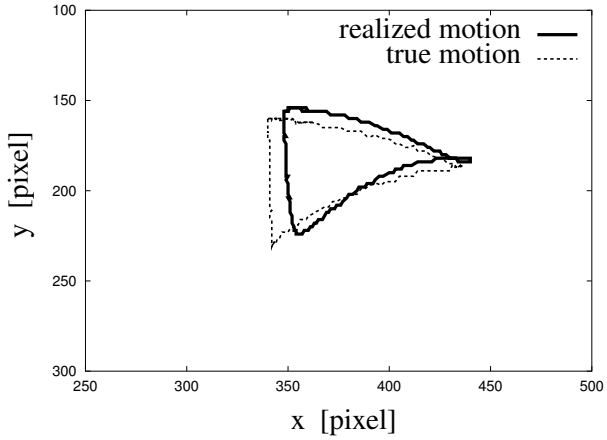


(b) The change of the evaluation function

Figure 2.18: An experimental results: [left] the changes of the learner joint angles (solid line) and the desired one (broken line); [right] the change of the evaluation function.

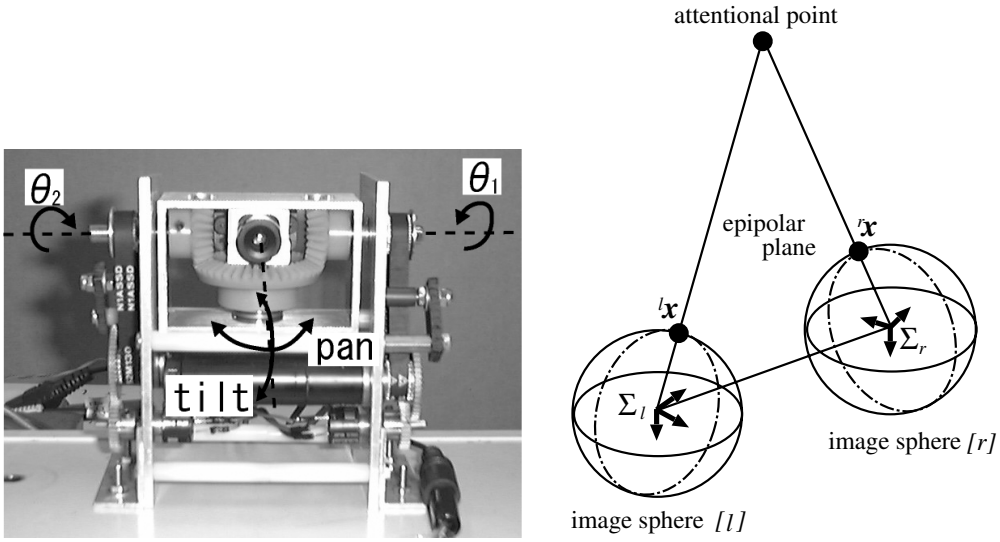


(a) $[L]$



(b) $[R]$

Figure 2.19: The trajectories of imitated motion (solid lines) and those (broken lines) of desired in the learner's view obtained by the same control input as the demonstrator.



(a) Rotation invariant camera

(b) Epipolar geometory between image shperes

Figure 2.20: Epipoloar geometry in the rotation invariant cameras

Extending the visible region We assumed that both bodies of the learner and the demonstrator which are used to estimate the parameters of view transformation and/or to control are visible for the learner in the experiments. However, they could go out of the visible region if the cameras of the learner are close to its body as usual in a humanoid robot. In such a case, it should be able to swing its cameras to keep tracking its end-effector or to find sufficient number of points to estimate the parameter of view transformation. Therefore, the proposed methods should be extended to cope with such swinging. We think that it can be realized by using rotation invariant cameras which can rotate without changing the position of their focal points (see Fig. 2.20(a)). Due to the invariance of the focal points, the rotation of the cameras does not changes epipolar geometry between spherical image coordinate system (see Fig. 2.20(b)). Such coordinate system can be accessed from the robot's own viewpoint by alternately using the joint angle for the camera rotation as the image coordinate.

Development of spatial perception By the proposed method, the learner performed the demonstrator's view recovery based on epipolar geometry. However, it is also a formidable issue to consider how the learner can realize the existence of such geometrical constraint. The demonstrator's view recovery can be regard as the mental processes to imagine that it locomotes to stand at the demonstrator's position and obtains the view from the position. We think that the capability of the demonstrator's view recovery can be acquired through the experiences of self locomotion. This idea might correspond to the idea in a psychological study that experiences of locomotion and visually tracking are important factor for child's development of spatial knowledge [46]. Therefore, how can a robot learn to perfrom such mental locomotion through the experiences of locomotion might be the first step of this issue. Furthermore, it must have some representations about its body to mentaly operate in the virtual locomotion. Concerning the acquisition of the body representation is addressed in the next chapter.

Chapter 3

Body finding based on the invariance in the multiple sensor data

3.1 Introduction

In the previous robotics, since the task of the robot is usually given in the coordinate system that is introduced from the viewpoint of external observer, the representation of the body needed for its task should be defined and calibrated in the same coordinate system (ex.[47, 48]). However, in such approach to define the body in the coordinate system introduced from the designer's viewpoint, the robot's adaptability to the changes in the environment and robot body itself is limited to what extent the designer has supposed in advance. On the other hand, it is suggested that a robot can structure its own sensory input through the interaction with the environment, and thereby induce regularities that simplify learning [3, 4]. According to these backgrounds, I believe that we might be able to build a highly adaptive robot based on the approach where it learns how to perform its task defined in its own sensory space instead of using the coordinate system introduced by the designer. When we follow this approach, how to construct the body representation needed to solve the task in the robot's own sensory space.

On the other hands, the representation of the body in human beings, that is so-called *body scheme* or *body image*, has been focused as a basis of the mechanism of motion and cognition in human beings such as tool-use [49, 50], imitation [51], and sense of self [52]. It seems adaptive as reported that a human can adapt the body representation in his/her brain to the changes in

his/her body, for example amputation [53]. However, the process to acquire the body representation has not been revealed yet. Therefore, building a robot that acquires its body representation from its own viewpoint is an interesting issue from a viewpoint of cognitive developmental robotics [22] that aims at both establishing the design principle of an intelligent robot and understanding intelligence of human beings.

The previous studies on the acquisition of the body representation in robotics can be classified in following three groups according to the assumptions of them.

(1) methods to find body

In the previous work on motor learning [54, 55, 56, 57, 58, 59, 60] or to what extent its body occupies [61, 62], it was usually assumed that the method to extract its body from the sensory data is given. However, it had better autonomously find its body from its own sensory data since it is difficult for the designer to prepare the universal method that can be applied for various kinds of body and environment.

(2) *a priori* knowledge on body structure

It has been proposed that a robot can autonomously find its body based on the correlation between its motion and the changes in sensory data since its motion induces the correlated optical flow with the motor commands [63, 64, 65, 66]. Based on this idea, it can find its body in its view by estimating the correlation coefficient of visual sensory data with its motor commands [63, 64, 65] or by identifying the state transition model [66]. However, in these methods, its visual sensor should be fixed to the environment since the movement of the visual sensor induces optical flow in all over the view. In other words, it was tacitly assumed that it had known which degrees of freedom contributed on the motion of the visual sensor. However, to perform finding body independently of the implementation of the body, it should be able to find the body without such *a priori* knowledge on the body structure.

(3) the constraint of body in sensation

Since the sensory data of a robot reflect the constraints of body, the robot is expected to find its body by utilizing the statistics of the sensory data instead of *a priori* knowledge of the body structure. Based on the statistics of sensory data, there already exist some previous work concerning how to construct the topography of the sensors distributed on

the body [67, 68] or how to find the concept of space where its body exists [69]. Although they are considered to be bases of body representation, it has not been evident to find which part of the sensory data indicates its body.

In this chapter, I address an issue to find which part of the sensory data indicates the robot's body based on the constraints of body in its sensation instead of assuming *a priori* knowledge on body structure. According to a conjecture that the sensations about its self body are independent of environment, it is supposed that the body can be defined by judging whether the sensory data is invariant in terms of the observing posture. The sensory data might perturb for the changes in observing posture and sensory noise both of which depend on the combination of sensory data and the properties of the perceived body surface. Therefore, I introduce a mixture of Gaussian distributions to model such invariance in multiple sensory data. It can discriminate the body from non-body by judging which distribution likely causes the variance of sensory data in the current observing posture.

In the rest of this chapter, first I describe the problem to be tackled and the basic idea to find the body based on the invariance of sensory data. Then, introduce the method to discriminate the robot's body from non-body based on the invariance. To confirm the validity of the proposed method, the experimental results with real robots are shown. Finally, I would like to discuss the limitation of the proposed method and the future work on this topic.

3.2 Body finding based on the invariance

The problem Suppose that the robot has a body structure in which it can observe also its own body by the sensor to observe the external world such as cameras or touch sensors, and that it can distinguish its proprioception to measure its posture from several kinds of sensor data those are obtained through the processes of feature extraction (see Fig. 3.1). Note that the way to extract its body from the sensor data is assumed to be unknown. For example, the designer does not teach what kinds of color or what kinds of texture its body has. Under such assumptions, the task of the robot is to judge whether the sensation in the current posture is caused by observing the self-body or the environment.

The invariance of the sensory data in observing self-body The sensory data in observing self-body is considered to be invariant in terms of ob-

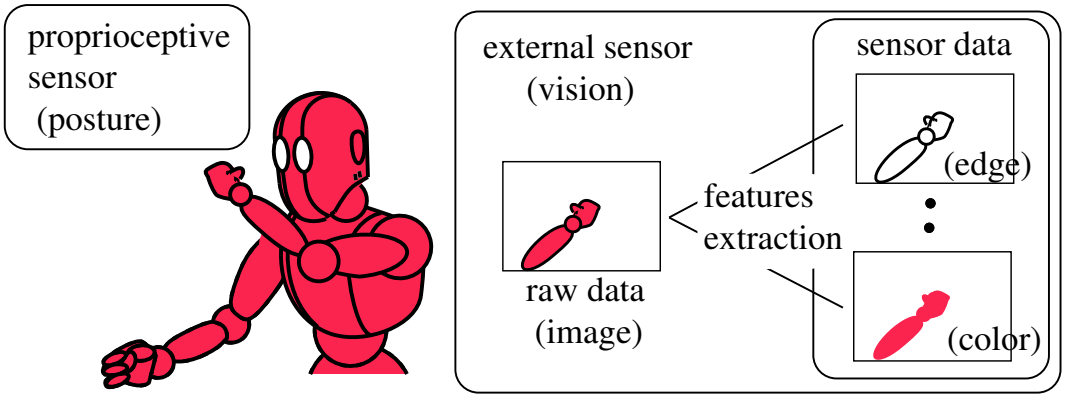
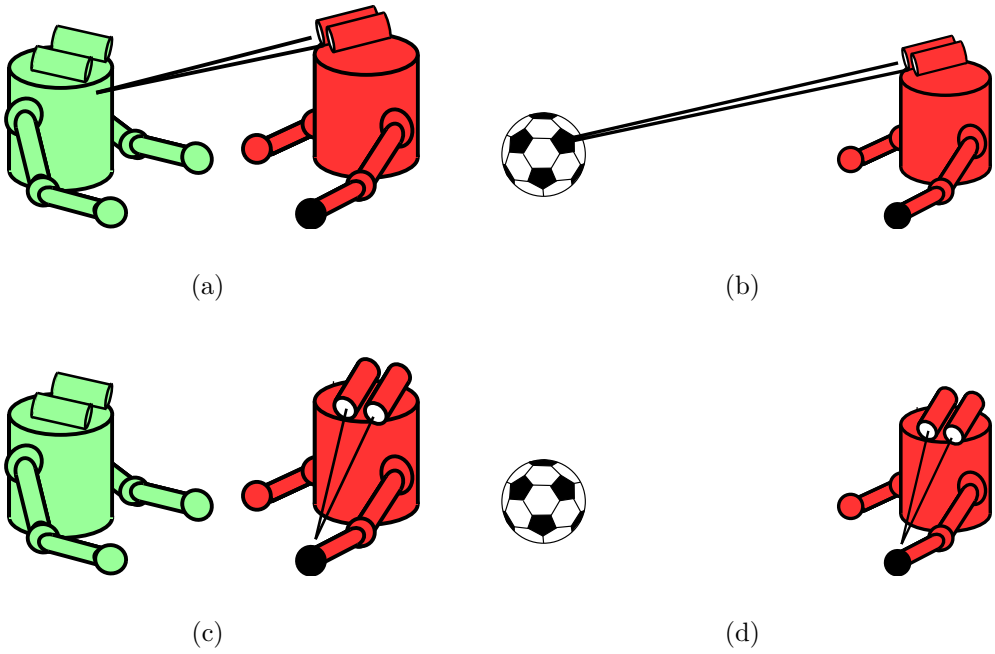


Figure 3.1: Assumptions: The robot can observe both its body and environment, thereby obtain several kinds of sensor data from the external sensor, and know the current posture from the proprioceptive sensor.

serving posture. For example, when a robot observes the environment in a certain posture, its sensation depends on environment even if it takes the same posture (see Fig. 3.2(a) and (b)). On the other hand, when it observes the self-body, it can observe the same part of the body independently of the environment as long as it takes the same posture (see Fig. 3.2(c) and (d)). As well as in the case of using touch sensor, the sensory data in touching its own body is invariant unlike in touching the external objects is not (see Fig. 3.3). Therefore, it is considered to be able to judge whether the sensory data in the current posture causes observing the self-body or the environment by judging whether the sensory data is invariant.

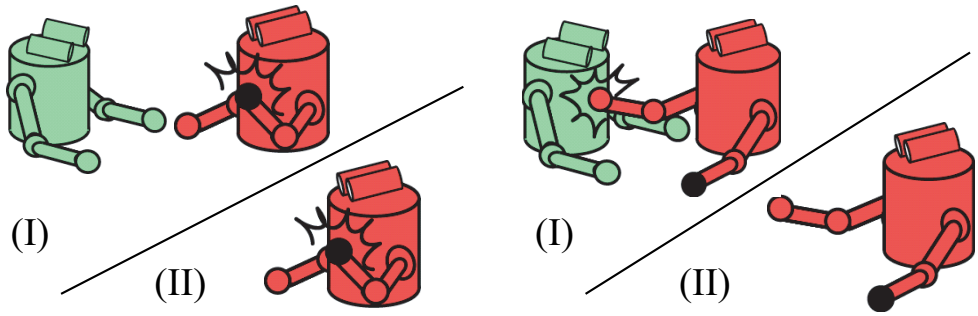
Furthermore, even when it observes the self-body, it cannot necessarily find out the invariance because of the perturbation of the sensory data for the changes in observing posture and sensory noise both of which depend on the combination of sensory data and the properties of the perceived body surface. Therefore, if it relies on sensory data only of one kind of sensory attributes, it might regard only a part of the body as its body or mis-regard a part of the environment as its body. For example, consider about disparity and luminance pattern as sensory data. When it observes the object that continuously occupies a certain region of space, such as the body, disparity is not sensitive for the changes in observing posture while it is usually difficult to measure disparity of coarse texture. On the other hand, although luminance pattern can be measured independently of the texture of the object, it is sensitive for the changes in observing posture when it observes the body part with fine



(a) (b)

(c) (d)

Figure 3.2: The invariance/variance of the view in terms of postures: when it observes the environment ((a) and (b)), what it observes depends on the environment. When it observes its body ((c) and (d)), it can observe the same part of the body independently of the environment.



(a) a posture with the invariant touch (b) a posture with the variant touch

Figure 3.3: The invariance/variance of the tactile sensations in terms of postures: when it touches its own body (a), it can always touch the same parts of the body. When it touches the environment (b), what it touches depend on the environment.

texture. Therefore, to discriminate the body from non-body independently of the texture it is expected to utilize the multiple sensory data, that is applying disparity for the body parts with fine texture and luminance pattern for one with coarse texture. In other words, the multiple sensory data should be complementarily utilized to find the body.

3.3 Body-nonbody discrimination based on a statistical model of invariance

3.3.1 Mixture of Gaussian distribution model

To perform finding body by complimentarily utilizing the multiple sensory data, mixture of Gaussian distribution is introduced to model the invariance in the sensory data in observing self-body. Suppose that the robot can observe the fixated object with D types of the sensory data such as disparity, luminance patterns, color, and so on. Denote the i-th sensory data in observing posture $\theta \in \Re^N$ as $\boldsymbol{x}_i(\boldsymbol{\theta}) \in \Re^{M_i}, (i = 1, \cdots, D)$ where N is the degrees of freedom of the robot's posture and M_i is the dimensionality of the i-th sensory data. The variance of the i-th sensory data $\sigma_i^2(\boldsymbol{\theta})$ is defined as the trace of the covariance

matrix $\boldsymbol{C}(\boldsymbol{\theta}) \in \Re^{M_i}$ of $\boldsymbol{x}_i(\boldsymbol{\theta})$, that is

$$\sigma_i^2(\boldsymbol{\theta}) \triangleq \mathrm{tr}\{\boldsymbol{C}(\boldsymbol{\theta})\}. \tag{3.1}$$

A vector that consists of D number of variance,

$$\boldsymbol{z}(\boldsymbol{\theta}) = [\tilde{\sigma}_1(\boldsymbol{\theta})^2, \cdots, \tilde{\sigma}_D(\boldsymbol{\theta})^2]^T \in \Re^D, \tag{3.2}$$

is called observing variance vector where $\tilde{\sigma}_i^2(\boldsymbol{\theta})$ is the normalized value of $\sigma_i^2(\boldsymbol{\theta})$ in this case the logarithm of $\sigma_i^2(\boldsymbol{\theta})$ is shifted so that

$$0 \leq \tilde{\sigma}_i^2(\boldsymbol{\theta}) < 1 \tag{3.3}$$

is satisfied.

Since sensation can be caused by observing the body or the external world while the sensations of the body involves less variance than one of the external world, it is conjectured that the distribution of observing variance vectors can be regard as a mixture of two Gaussian distributions for observing the body and the nonbody, respectively (see Fig. 3.4). In other words, the distribution of $\boldsymbol{z}$ is given by

$$p(\boldsymbol{z}; \alpha) = w_b \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_b) + w_e \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e) \tag{3.4}$$

where $\mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a normalized Gaussian distribution of $\boldsymbol{z}$ with the average $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ and suffices b and e indicate the body and the environment, respectively. The weights w_b and w_e satisfies

$$\begin{aligned} w_b + w_e &= 1, \\ 0 \leq w_b, w_e &\leq 1. \end{aligned} \tag{3.5}$$

3.3.2 Estimation of the distribution

Since the robot can measure $\boldsymbol{z}(\boldsymbol{\theta})$ but does not know which of two distributions generates the measured $\boldsymbol{z}(\boldsymbol{\theta})$, it must estimate the distribution (eq. 3.4) from the incomplete data, that is $Z = \{\boldsymbol{z}(\boldsymbol{\theta}_1), \cdots, \boldsymbol{z}(\boldsymbol{\theta}_{q_\theta})\}$. Therefore, we apply an EM algorithm [70] to this problem, which is a theoretical paradigm to estimate the maximum likelihood parameters from an incomplete data.

According to the EM algorithm, to obtain the parameters that maximize the logarithmic likelihood function such as

$$\mathcal{L} = \log p(Z|\alpha), \tag{3.6}$$

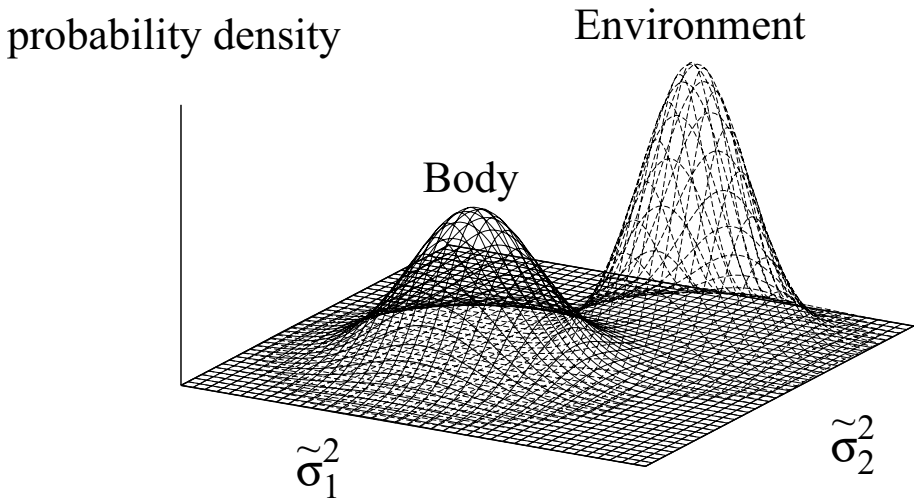


Figure 3.4: Mixture of Gaussian distribution model of observing variace vector

the expectation process and the maximization process are iterately performed until they converge for given initial parameters. In the expectation process, the expectation of the logarithmic likelihood function of the complete data in a given condition of Z and $\alpha^{(t)}$,

$$Q(\alpha|\alpha^{(t)}) = E_Z\{\log p(Z, H|\alpha)|Z, \alpha^{(t)}\} \tag{3.7}$$

where $\alpha = \{w_b, \boldsymbol{\mu}_b, \Sigma_b, w_e, \boldsymbol{\mu}_e, \Sigma_e\}$, is calculated where $\alpha^{(t)}$ is the estimated parameter set until the t-step and H is a set of hidden parameters that identify which of two distributions generates $\boldsymbol{z}(\boldsymbol{\theta})$. In the maximization process, α is updated so that the new α maximize $Q(\alpha|\alpha^{(t)})$. It is guaranteed that each iteration of the expectation and maximization process of the EM algorithm increases the logarithm likelihood function [70].

The task of the robot is judging whether the current sensation in $\boldsymbol{\theta}$ is caused by observing the self body or the environment. It can be performed by judging which estimated distribution likely causes the variance of sensory data in the current observing posture.

3.4 Experiments

In this section, I will introduce experimental results with real robots. To show to what extent the proposed method work, we use two robots which have

different body appearance from each other. One has very robotic body surface (see Fig. 3.5), while the other has infant-like body surface (see Fig. 3.14).

In the first experiments, to test whether the proposed method works independently of the robot's embodiment, we paste different textures, one is fine and the other is coarse, on the different parts of the arm as shown in an egocentric view of the robot (see Fig. 3.6). It consists of two cameras on the camera head which can rotate in the pan and tilt axes, the 4-DoF arms, and the mobile base. It obtains four kinds of visual sensory data at the center region of the left camera, namely, disparity, luminance pattern, chroma, and direction of edges. In the following experiments, however, we only show the case where the arm is fixed in a certain posture as shown in Fig. 3.5 to make result easy to understand.

Through the exploration by randomly changing its posture both of its camera head, it corrects the sensory data to obtain the averages and the variances of them in terms of each posture quantum. During the exploring process, we let the robot move around to make the external world varies. Fig. 3.7 schematically illustrates the exploring process. Note that the correction image in Fig. 3.7 is slightly different from an egocentric view in Fig. 3.6 since the latter is a entire image of the camera while the former is a correction of a part of the image in various postures of the camera head. The task of the robot is extracting the part of the sensory data that is caused by observing its body. Fig. 3.8 is the desired extraction performed by the experimenter. Note that although we, external observer, can easily distinguish the body and the environment by looking at the correction image in Fig. 3.7 and therefore correctly extract the body like in Fig. 3.8, it is formidable for a robot since it is not given any *a priori* knowledge about what its body is.

3.4.1 Body-nonbody discrimination with luminance pattern ($D = 1$)

First we test the proposed method only with luminance pattern of the image elements at the center region (8×8 [pixel]) of the left camera as the sensory data. Fig. 3.9(a) shows the distribution of variances of the luminance pattern in terms of each posture (histogram) and the estimated mixture of Gaussian distributions (solid curve). The distribution at lower variance is considered to correspond to the one in observing self body while the other is considered to correspond to the one in observing the external world. Fig. 3.9(b) illustrates the averaged luminance pattern of the extracted body where the variance is regarded to be caused by the distribution of self-observation.

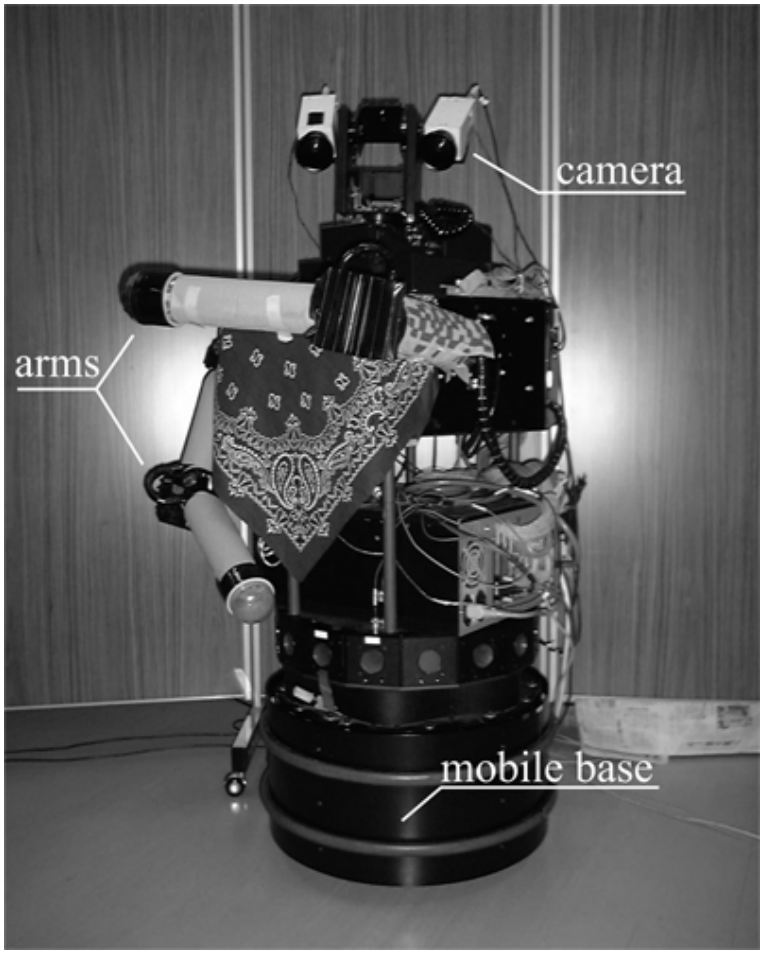


Figure 3.5: Appearance of the robotic test-bed

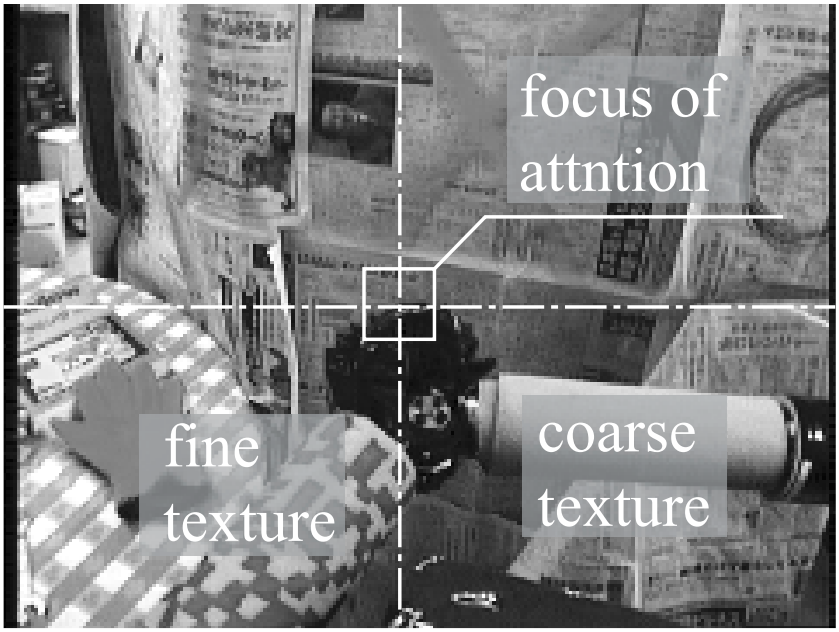


Figure 3.6: An egocentric of the robotic test-bed

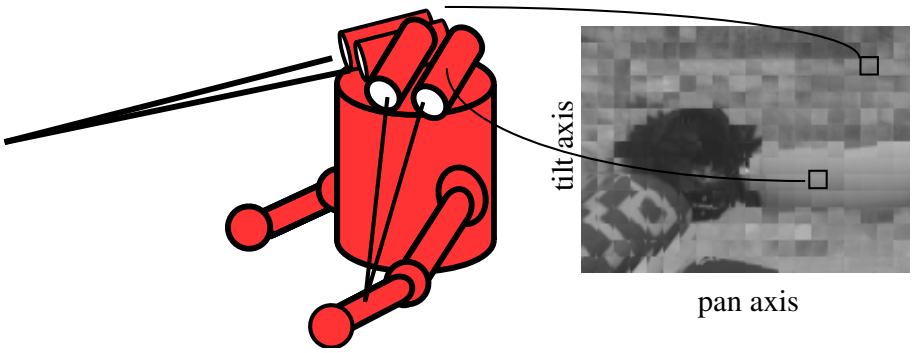


Figure 3.7: An schematic explanation of the learning process: the robot corrects the average and variance of the sensory data for each observing posture.

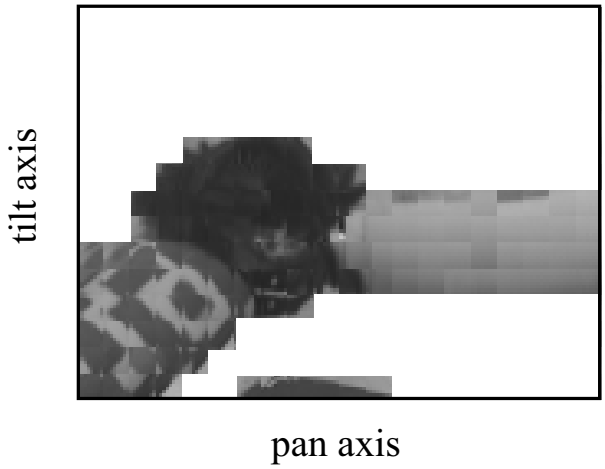


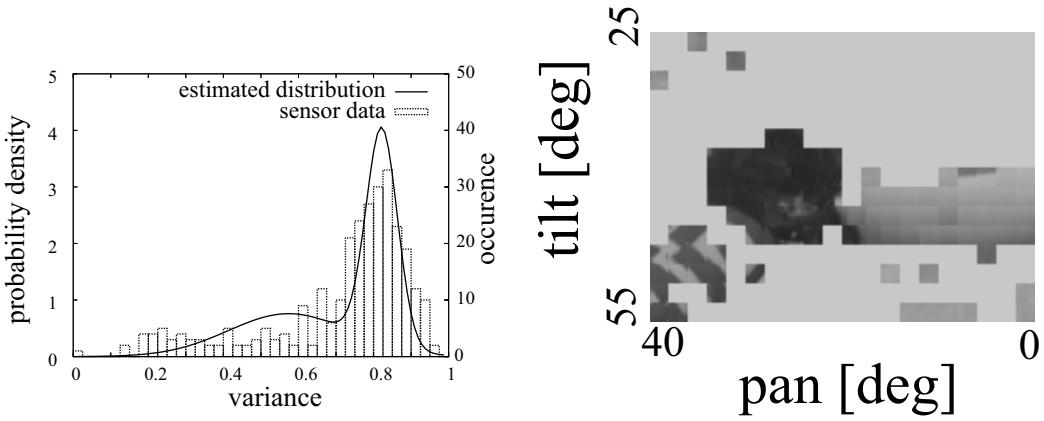
Figure 3.8: The desired extraction of the body

By comparing Fig. 3.8(b) and Fig. 3.9(b), we can see that the body part with coarse texture is correctly extracted while one with fine texture is not. It seems because the observed luminance pattern of the fine texture sensitively varies with the slight changes of observing posture.

3.4.2 Body-nonbody discrimination with disparity ($D = 1$)

Fig. 3.10(a) shows the distribution of observing variances of the disparity (histogram) and the estimated mixture of Gaussian distributions (solid curve). As well as in the previous case, the distribution at lower variance is considered to correspond to the one in observing self body while the other is considered to correspond to the one in observing the external world. Fig. 3.10(b) illustrates the average luminance of the extracted body in the posture where the observing variance is regarded to be caused by the distribution of self-observation.

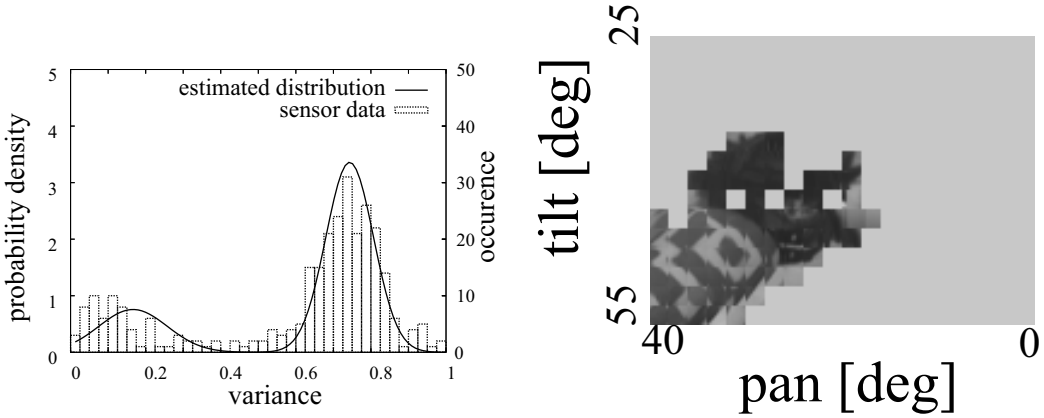
By comparing Fig. 3.8(b) and Fig. 3.10(b), we can see that the body part with fine texture is correctly extracted while one with coarse texture is not. It seems because the robot tends to fail in stereo matching needed to detect the disparity at the coarse texture.



(a) distribution of observing variance and its estimation

(b) the found body

Figure 3.9: Body-nonbody discrimination with luminance pattern



(a) distribution of observing variance and its estimation

(b) the found body

Figure 3.10: Body-nonbody discrimination with disparity

3.4.3 Body-nonbody discrimination with luminance pattern and disparity ($D = 2$)

In the former experiments, we saw that the parts of the body which is correctly extracted by the proposed method depended on what type of sensory data are used. Then, we test whether the performance of body-nonbody discrimination is improved by complementarily utilizing both of these two kinds of sensory data.

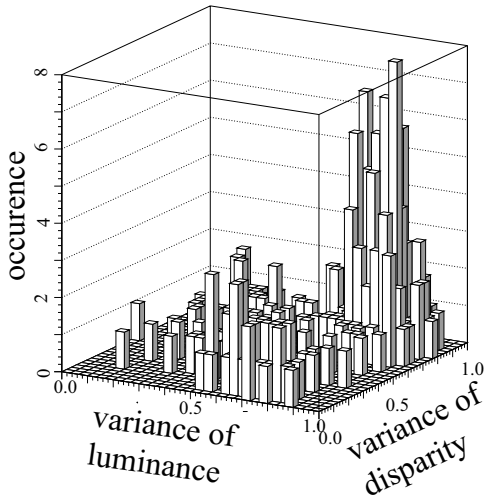
Fig. 3.11 illustrates the experimental result with luminance pattern and the disparity at the center region of the left camera. Fig. 3.11(a) shows the distribution of observing variance vectors each of which consists of the variance of disparity and one of luminance pattern for a certain posture. Fig. 3.11(b) is the estimated mixture of Gaussian distributions. As well as in the case of using single sensory data, we can see that the distribution at lower observing variance corresponds to the one for self-observation while the other corresponds to the one for observing the external world. Fig. 3.11(c) illustrates the average luminance pattern of the extracted body where observing variance is regarded to be caused by the distribution of self-observation.

From Fig. 3.11(c), we can see that the body parts both with coarse and fine textures are almost extracted. The extracted body is closer to the desired one (Fig. 3.8) compared to the results with single sensory data (Fig. 3.9(b) and 3.10 (b)). Therefore, it is considered that two kinds of sensory data complementarily contribute to discriminate body from non-body.

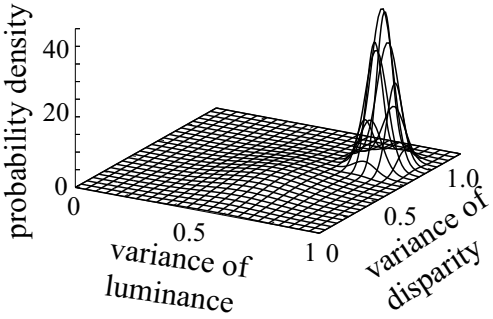
3.4.4 Body-nonbody discrimination with luminance, disparity, color and edge direction ($D = 4$)

Furthermore, we conduct another experiment by adding two kinds of sensory data, namely color and edge direction. Fig. 3.12(a) shows the result of body-nonbody discrimination by using color and Fig. 3.12(b) shows one by using edge direction. We can see that it mis-regard the environment as its body in more observing postures than one in the previous experiment with single sensory data (Fig. 3.9(b) and Fig. 3.10(b)). By the body-nonbody discrimination with four kinds of sensory data including these sensory data, that is luminance, disparity, color, edge direction, it succeeded in discriminating body from non-body at least in the same extent (Fig. 3.13) as, or slightly wider region than, in the previous result with two kinds of sensory data, that is luminance and disparity (Fig. 3.11(b)).

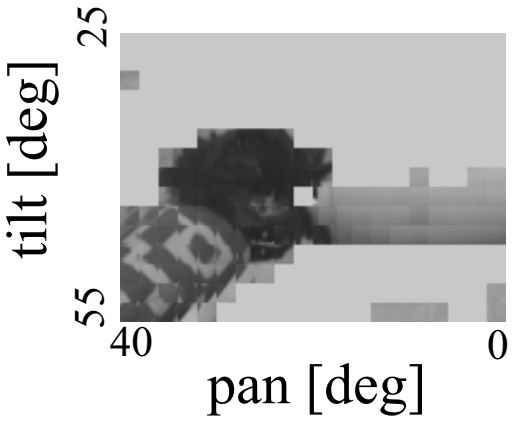
Therefore, we may conclude that the proposed method works even if the



(a) distribution of observing variance

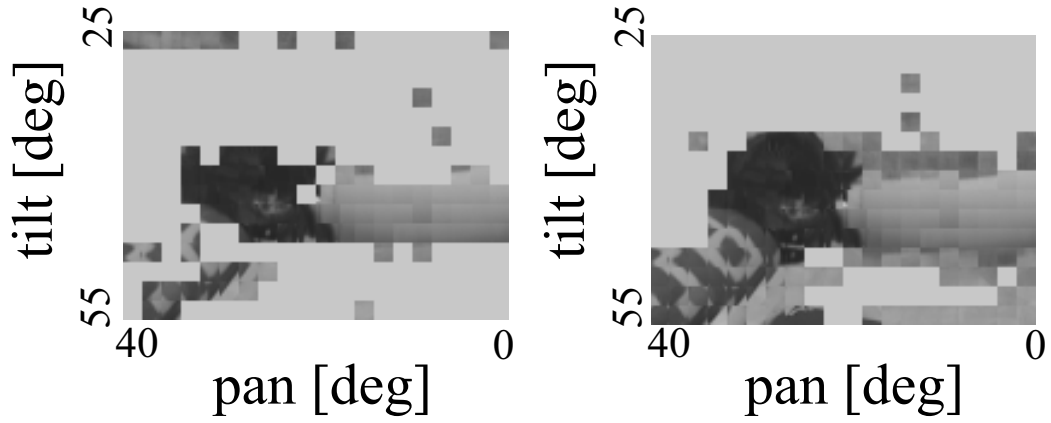


(b) the estimated distribution



(c) the extracted body

Figure 3.11: Body-nonbody discrimination with luminance and disparity



(a) found body with color (b) found body with direction of edge

Figure 3.12: Body-nonbody discrimination with additional sensory data

robot use some inappropriate sensory data with which incorrect extraction of the body is caused.

3.4.5 Discriminating body from non-body by a robot with infant-like body appearance ($D = 2$)

To confirm that the proposed method can apply to other robot with different body surface, we conducted another experiment by using a robot with completely different appearance from the previous one (Fig. 3.5). Fig. 3.14 shows an overview of the test-bed robot that has infant-like body appearance while only the head is replaced with a camera. It obtains two kinds of sensory data ($D = 2$), namely luminance pattern and color of the center of the view (Fig. 3.15). The camera can rotate in both pan and tilt axes which corresponds to its DoFs. Therefore $N = 2$ in this experiment. It corrects the sensory data by randomly changing the posture of the camera and calculate their average and variance. The environment varies during the learning process since the experimenter randomly move the robot.

Fig. 3.16(a) is the result of extracting the body by the proposed method with both luminance and disparity. We can see that it almost succeed in extracting the body although there exist regions where it mis-regard the environment as its body. Fig. 3.16(b) and (c) shows the distribution of the variance of luminance and color, and the estimated mixture of Gaussian distribution,

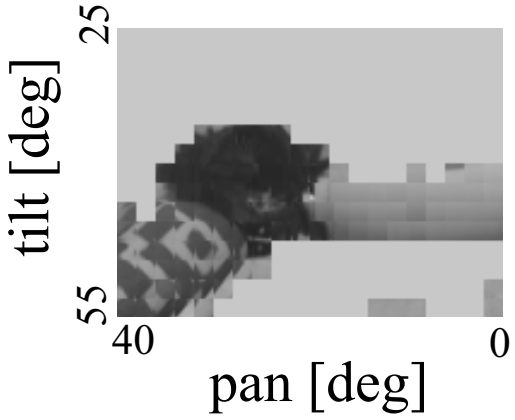


Figure 3.13: Body-nonbody discrimination with four kinds of sensory data, luminance, disparity, color, and direction of edge

respectively. From the landscape in each axis, it can be seen that there are some overlaps between two distributions that seem to cause the failure when it tries to discriminate its body from non-body with each of them. However, from the landscape in the combined axes, they look being located separately. Therefore, we may conclude that the proposed method work well by complementarily utilizing the multiple sensory data even in the robot with infant-like body surface.

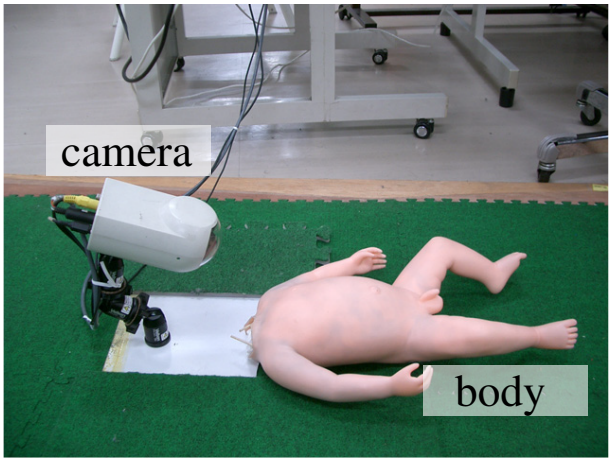
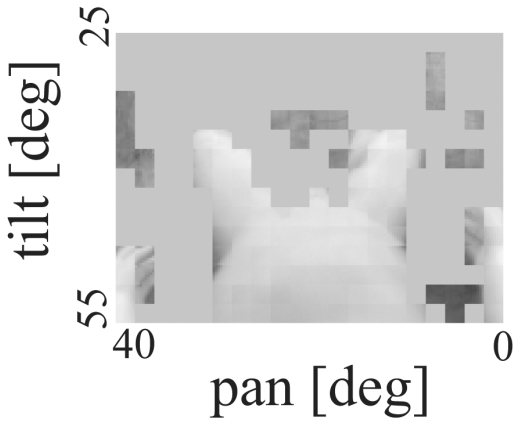
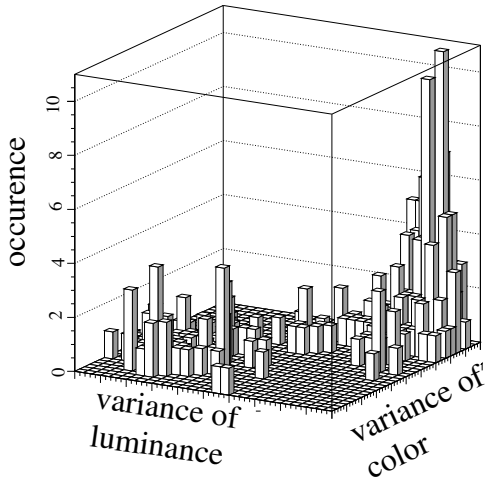


Figure 3.14: An appearance of the body of a test-bed robot with an infant-like body surface

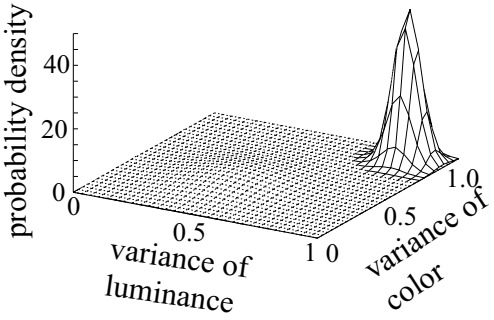
Figure 3.15: An example of egocentric view of the test-bed robot



(a) found body



(b) the variance



(c) the estimated distribution

Figure 3.16: Body-nonbody discrimination by an infant-like robot with luminance and color

3.5 Summary and discussion

In this chapter, I introduced the method to discriminate the robot's body based on an idea that the body can be defined by the invariance of the sensory data. In the proposed method, the distribution of the variance of sensory data in terms of observing posture were approximated by the mixture of two Gaussian distributions each of which corresponds to the observation of the body and the environment, respectively. It can discriminate its body from non-body by judging which distribution likely causes the variance of the sensory data in the current observing posture.

By the experiments with real robots, it was considered that the proposed method could work well independently of the properties of body surface by complementarily utilizing the multiple sensory data. Although some combination of sensory data causes the overlap of two distributions, it is expected to be able to make them separate by adding different type of sensory data. However, note that increase of the dimension of sensor data causes the increase of the search space to estimate the mixture of Gaussian distributions.

Multimodal representation of body Since what the robot become to perform by the proposed method is just judging whether it observes its body or environment in the current observing posture, and therefore, is not sufficient for the body representation to attain the its task. To approach it, the issue on how to acquire the body representation in the multi modalities is one of our future work. When a robot tries to integrate the sensation in multi modalities, it faces with the problem where different modalities often receive different objects since the receptive fields of them are limited. we think that the the invariance in self-perception can be also utilize for this problem and has already propose a method to match the sensations in touching and watching [61, 71].

Biological plausibility It is reported that a human neonate can distinguish double-touching from being touched by the other in the study on rooting reflex [72]. It might be explained by the proposed method as double-touching is invariant in terms of touching posture unlike being touched. Infants seem to learn the invariance through touching their mouth with their hand in the womb as reported in a clinical study [73].

As mentioned in the introduction, the method to find the body based on the sensory correlation with its motion [63, 64, 65, 66] needs *a priori* knowledge on body structure. Furthermore, it cannot be applied to find the link in-between the visual sensor and the environment such as the trunk or the

legs since the visual sensor should be fixed. However, from the viewpoint of the modeling the cognitive developmental process of human being, this approach might be more plausible. For example, it is reported that infants can discriminate self-produced motion based on visual-proprioceptive contingency [52, 74, 75] or that even a neonate seems interested in the relationship between its hand motion and changes in its view [76]. Although the basic idea of the proposed method in this chapter is different from one in the correlation based approach, I think they do not conflict each other but can complementarily work. Therefore, I would explore the possible extension to involve the information of motion in the proposed method. Furthermore, verifying the proposed model from the viewpoint of a constructivist approach to model the human cognitive developmental process should be also considered.

Chapter 4

Vowel acquisition through interaction with human caregiver

4.1 Introduction

Vocalization, making sound through modulating a source sound, is one of the most promising means of human-robot communication because humans rely on speech in their daily life. To communicate through vocalizations, an agent should share with its interlocutor common phonemes, which are elemental units of vocal communication. Human infants acquire phonemes of their mother tongue and finally their mother tongue itself through interactions with their caregivers without having the capability to articulate, nor having *a priori* knowledge about the relationship between the sensorimotor system and phonemes. In this study, we aim to build a robot that learns to vocalize with a human caregiver in the conditions similar to those of human infants. As Asada et al. have suggested in their discussion of a constructivist approach to cognitive developmental robotics [22], building this kind of a robot may help us to model the human developmental process of phoneme acquisition.

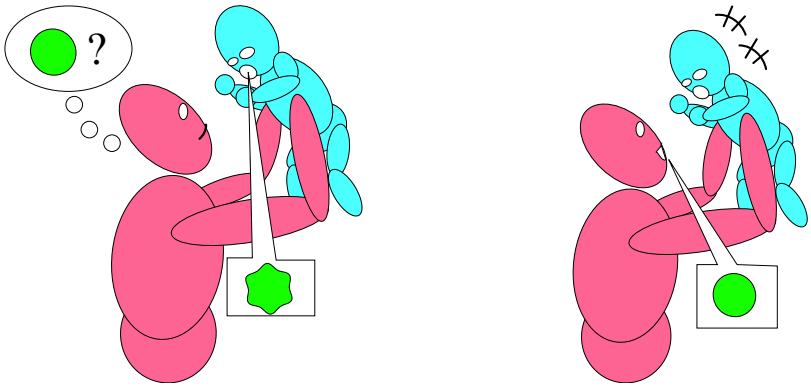
We assume that a robot can acquire phonemes without any knowledge about the relations between phonemes and its sensorimotor system. Thus, it must obtain information for learning them through interactions with its environment, namely its caregiver. Previous studies showed that a population of computer simulated agents with a vocal tract and cochlea can acquire shared vowels by self-organization through interactions with each other [77, 78]. Although they did not assume *a priori* knowledge about vowels, there was an

assumption that the agents can reproduce sounds similar to those of other agents so that "imitation game [77]" or "magnet effect [78]" leads to share vowels in population. However, we should take infant immaturity into account for modeling the vowel acquisition process since infants cannot reproduce the caregiver's utterances as they are.

To build a robot that has different capability of vocalizing from its caregiver and acquires phonemes through interactions with the caregiver, we have to cope with two main design issues: what are the interactive mechanisms involved and what should be the behavior of the caregiver/teacher? We observe that maternal imitation effectively reinforces infant vocalization [24, 25] and that its speech-like cooing tends to lead utterances of its mother [79]. Therefore, we hypothesize that imitation by the caregiver, which is repetition of infant's vocalization with adult phonemes, plays an important role in phoneme acquisition through interactions. The purpose of this study is to build a robot that acquires phonemes through random vocal articulations and interactions with a caregiver who repeats the robot's vocalizations.

In this chapter, we address the issue to build a robot that acquires vowels through interactions with its caregiver. The robot can vocalize by a vocal apparatus similar to the one used in Higashimoto and Sawada [80]. We assume that the robot has the capability of extracting formants which are well-known sound features to distinguish vowels [81]. Detecting formants seems a basic element of perception since many species show an ability to use formant as a perceptual cue [82]. The learning mechanism consists of interconnected auditory and articulation layers which clusters the input, formants and its own articulation parameters, respectively, by self-organization [83]. The connection weights between them are updated by means of simple Hebbian learning during interactions with the caregiver. The caregiver's repetitive utterances enables the robot to acquire vowels by matching its articulations with the caregiver's vowels. To resolve the arbitrariness in selecting the proper articulations, we introduce "subjective criteria" into the learning rule that considers the toil involved in the articulation, in this case the torque to deform the tract and its resultant deformation, based on the specific parameters to the individual.

The rest of this chapter is organized as follows: First, we explain how to design interactions to enable the learning of vowels. Then, we describe how learning works. After showing the configuration of the experimental robot and reporting a preliminary experiment to verify its acoustic properties, we discuss how the proposed method works.



(a) (b)

Figure 4.1: Conceptual figures of mother-infant interaction: (a) the infant/robot coos randomly, (b) if the caregiver is able to understand the sounds as being vowels then the caregiver repeats them with his or her own vowel categories.

4.2 The environmental design for interaction

For a robot to learn to vocalize vowels without *a priori* knowledge about the relationship between the vowels and its sensorimotor system, it should interact with a caregiver. Our approach is to set up a situation in which the robot and the caregiver interact in a way that promotes the robot's learning.

Based on a study of mother-infant interaction [24, 25, 79], we conjecture that maternal imitation of an infant's vocalizations plays a key role in the vowel acquisition of an infant. To build a robot that reproduces the observed interaction in the developmental studies, we embed in the robot a mechanism for producing random cooing. At the same time, the caregiver utters the matched vowel if the robot's vowel can be regarded as a human vowel, but does not otherwise (see Fig. 4.1). Note that the caregiver utters his or her own vowels.

By designing the experimental situation in this way, the robot obtains the invariant pairs of its articulation and the corresponding vowel so that vowel acquisition can succeed based on a simple learning rule despite the difference in the articulation parameters.

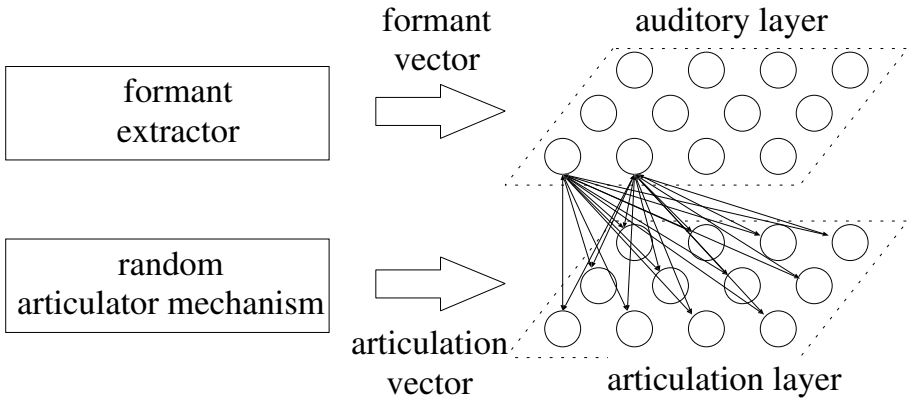


Figure 4.2: The learning mechanism that consists of interconnected auditory and articulation layers: the auditory layer receives formant vector of the caregiver's utterances while the articulation layer receives the articulation vector to deform its vocal tract.

4.3 Learning mechanism

The robot's learning mechanism consists of two layers and connection between them (see Fig. 4.2). After describing the processing in these two layers, we give two learning rules to connect between them.

4.3.1 Auditory layer

Frequency peaks in the soundwaves are effective to distinguish vowels; the peaks are called formants [81]. The auditory layer receives *formant vectors* from the formant extractor. Each vector consists of the frequencies corresponding to the lowest four peaks of the caregivers' utterances; the auditory layer clusters them in a self-organizing manner, using a Kohonen map [83].

The auditory layer consists of N_f units. The i-th unit has a codebook vector $\boldsymbol{f}_i \in \Re^4$ and a position vector $\boldsymbol{r}_i \in \Re^2$ that indicates the position in the layer. When the auditory layer receives a formant vector $\boldsymbol{f} \in \Re^4$, units with closer codebook vectors activate more, and the most active unit suppresses the other units. Finally, an activation a_i^f of the i-th unit is calculated by

$$a_i^f = \begin{cases} g(\boldsymbol{f}_i^T \boldsymbol{f} - h) & \text{if} \quad i = \arg_j \max \boldsymbol{f}_j^T \boldsymbol{f}, \\ 0 & \text{otherwise,} \end{cases} \tag{4.1}$$

where $g(x)$ and h are a step function of scalar x and a scalar threshold, re-

spectively. The most active unit is called the winner and labeled win^f.

In the Kohonen map algorithm, codebook vectors of some units near the winner are modified to be closer to the input vector. The updating rule is defined as,

$$\begin{aligned} \boldsymbol{f}_i(t) &= \boldsymbol{f}_i(t-1) + \alpha(t) \cdot \\ & \quad \Phi(\boldsymbol{r}_i^f, \boldsymbol{r}_{win^f}^f)(\boldsymbol{f}(t) - \boldsymbol{f}_i(t-1)), \end{aligned} \tag{4.2}$$

where $\alpha(t)$ is the time dependent scalar learning rate and the neighborhood function $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is a monotonically decreasing function with respect to the distance between vector $\boldsymbol{x}$ and $\boldsymbol{y}$ that is calculated by

$$\Phi(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{|\boldsymbol{x} - \boldsymbol{y}|}{2\sigma^2(t)}\right), \tag{4.3}$$

where $\sigma(t)$ is a time dependent scalar that determines how much nearby units learn according to a distance metric. At the start of learning, $\sigma(t)$ is set to such a high value that $\Phi(\boldsymbol{x}, \boldsymbol{y})$ is high in a wide region of the auditory layer and gradually decreases so that $\Phi(\boldsymbol{x}, \boldsymbol{y})$ of nearby units only have high values at the end of learning. This learning rule clusters codebook vectors close to frequently observed input vectors in the auditory layer — in this case, frequently heard vowels.

4.3.2 Articulation layer

The articulation layer receives an *articulation vector* $\boldsymbol{m} \in \Re^5$ from the random articulation mechanism. The vector consists of five motor commands to deform the vocal tract. The articulation layer clusters the vectors by the same method used in the auditory layer.

The articulation layer consists of N_f units. The i-th unit has a codebook vector $\boldsymbol{m}_i \in \Re^5$ and a position vector $\boldsymbol{r}_i \in \Re^2$. For each i-th unit, the calculation of the activation a_i^m and update of the codebook vector are performed in the same manner as the auditory layer.

4.3.3 How to learn weights connecting two layers

The connecting weights between the auditory and the articulation layers are updated by the Hebbian learning rule. Therefore, connections between simultaneously active neurons in the auditory and articulation layers are strengthened while others are weakened. Let w_{ij} be a connecting weight between the

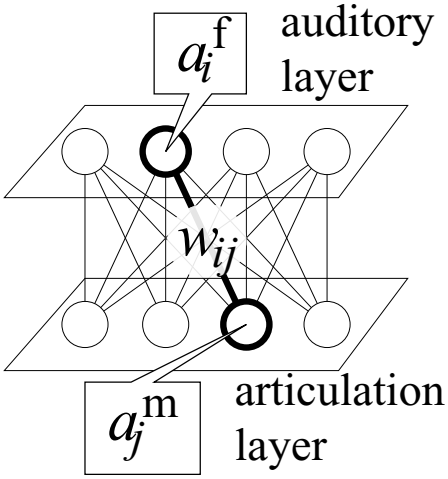


Figure 4.3: Descriptions of the variables: the activation of the i-th unit in the auditory layer a_i^f, that of the j-th units in the articulation layer a_j^m, and a connection weight between them w_{ij}

i-th unit in the auditory layer and j-th unit in the articulation one. The learning rule is defined as

$$\tau \dot{w}_{ij} = -w_{ij} + \alpha a_i^f a_j^m, \tag{4.4}$$

where τ is a time constant of learning and α is the learning rate. Based on eq. (4.4), w_{ij} will converge to

$$w_{ij} = \alpha E\{a_i^f a_j^m\}, \tag{4.5}$$

where $E\{a_i^f a_j^m\}$ is the average of $a_i^f a_j^m$ [84]. We use the discrete version of the updating rule (eq. (4.4)) such that,

$$w_{ij}(t+1) = w_{ij}(t) + \frac{1}{\tau}(\alpha a_i^f(t) a_j^m(t) - w_{ij}(t)), \tag{4.6}$$

where t denotes the time step.

Employing the initially random articulation mechanism causes invariant pairs of units to activate in both layers simultaneously since the caregiver is engaged in repetitive utterances. Therefore, through the learning process, clusters of articulation vectors are matched with corresponding vowels as connections between both layers.

However, interactions may connect multiple articulation units with a corresponding vowel since the caregiver may interpret some vocalizations caused by different articulations as the same vowel. To match a heard vowel with a unique articulation in order to vocalize it, we introduce subjective criteria into the learning rule so that the articulation involving less toil is selected — that is, the articulation vectors involving less toil obtains stronger connection from the auditory layer and vice versa. Therefore, the learning rule for the connections is slightly modified:

$$w_{ij}(t+1) = w_{ij}(t) + \frac{1}{\tau}(c\eta(\boldsymbol{m})a_i^f(t)a_j^m(t) - w_{ij}(t)), \tag{4.7}$$

where $\eta(\boldsymbol{m})$ is a function that evaluates the toil involved in the articulation calculated by

$$\eta(\boldsymbol{m}) = \exp\Big(-\frac{C_{trq}(\boldsymbol{m})}{\sigma_{trq}^2}\Big) \cdot \exp\Big(-\frac{C_{dfm}(\boldsymbol{m})}{\sigma_{dfm}^2}\Big), \tag{4.8}$$

where σ_{trq} and σ_{dfm} are scalar constants that are the specific parameters to the individual, and C_{trq} and C_{dfm} are the cost functions of the torque to deform the tract and its resultant deformation, respectively. σ_{trq} and σ_{dfm} are chosen by trial and error. The cost functions are defined as

$$\begin{aligned} C_{trq}(\boldsymbol{x}) &= \boldsymbol{x}^T\boldsymbol{x}, \text{and} \\ C_{dfm}(\boldsymbol{x}) &= \sum_{k=1}^{4}(x_k - x_{k+1})^2, \end{aligned} \tag{4.9}$$

where x_k is the k-th element of the vector $\boldsymbol{x}$.

4.4 A robotic test bed

Vocalization is generally well-known as an outcome from a modulation of a source of sound energy by a filter function determined by the shape of the vocal tract; this is often referred to as the "source-filter theory of speech production" [85]. We implement the source-filter theory by using a vibrator as a sound source and silicon rubber tube as a vocal tract whose shape is deformed by five electric motors. This implementation is similar to that of Higashimoto and Sawada [80] except that they use an artificial vocal cord while we use a membrane that a vibrator oscillates at a fundamental frequency.

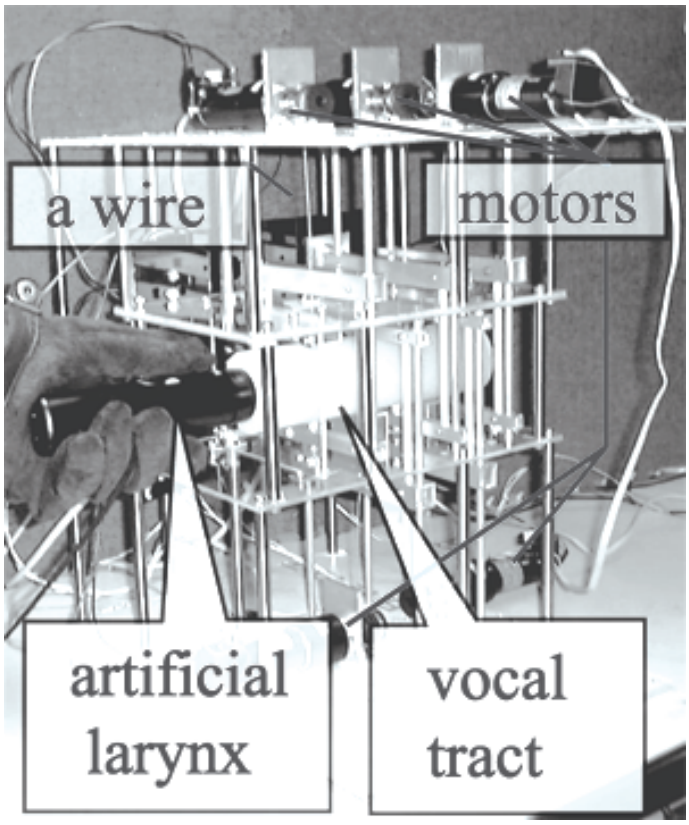
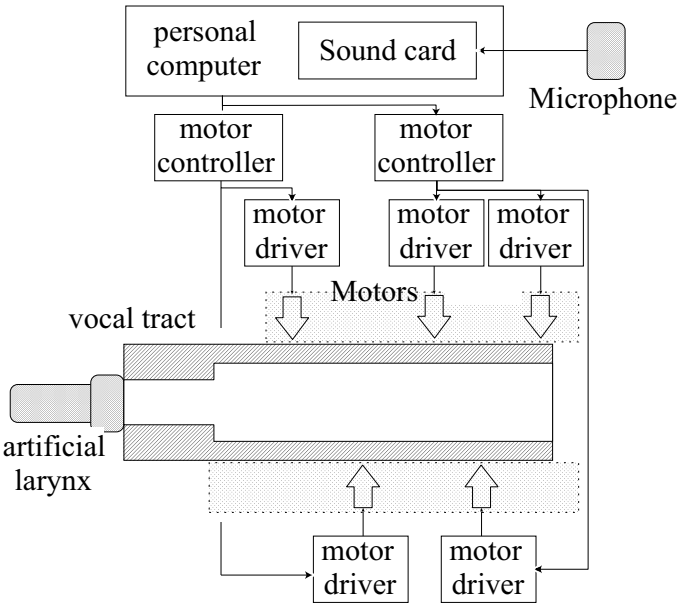


Figure 4.4: An appearance of the robotic test bed

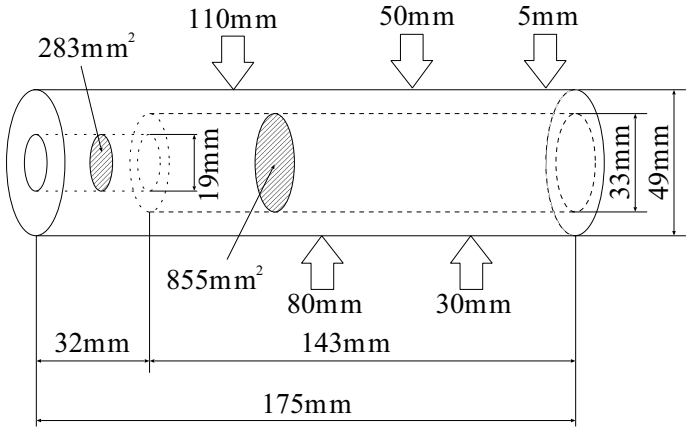
Figures 4.4 and 4.5 depict the robot hardware. Five electric motors are bound to their respective attachments by wires. The motors pull at the attachments to deform the tract. They are controlled by motor controllers (usbMC01, iXs Research Corp.) according to control commands from the host computer. An artificial larynx (Myvoice, Secom Medical System Co. Ltd.) is used as a sound source that generates a soundwave with the fixed property of frequency (see Fig. 4.6). The host computer receives signals through a microphone and calculates their formants.

4.4.1 Preliminary experiment

We conducted a preliminary experiment to confirm the acoustic property of the silicon vocal tract in the robot. We let the robot vocalize by sending various articulation vectors and measured formants of the vocalized sound if



(a) The connections of the elements



(b) The size of the vocal tract and the positions to be pulled by the motors

Figure 4.5: An overview of the system: empty arrows indicate the positions to be pulled and the numbers over them are the distances from the right-side end of the tract.

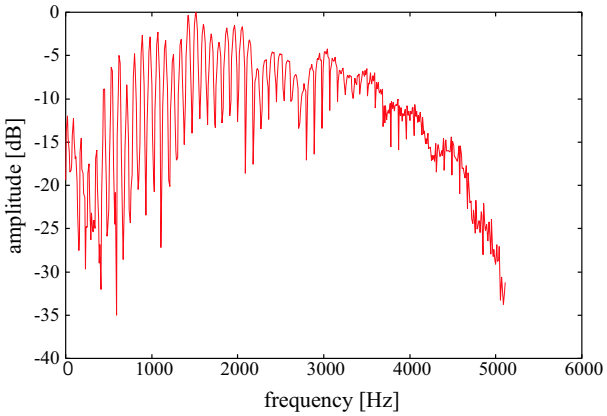


Figure 4.6: The soundwave frequency of the sound source

a Japanese experimenter (hereafter, caregiver) could interpret as vowels. For comparison, formants of the caregiver's vowels were also measured. Figure 4.7 shows two distributions of the calculated formants of the robot's vocalization and the caregiver's. The lowest three formants were sufficient to describe the distributions. Figure 4.8 shows averages for a few hundred samples of (a) the robot's and (b) the caregiver's vowels.

Formant distribution of the robot tends to be higher than that of the caregiver. We can see that the robot cannot reproduce human formants since there is no overlap between both distributions. They are clustered in the formants space. It means that formants are available for recognizing the vowels of the robot in addition to those of human beings. We confirmed that the robot can vocalize four Japanese vowels but not /o/. The robot cannot vocalize /o/ because its vocal tract does not have sufficient degrees of freedom as we determined through some preliminary tests. Therefore, the vowels in the following experiments exclude /o/.

4.5 Experiment

We conducted two learning experiment with and without the toil criterion to test whether the proposed method works. The auditory layer consists of 15×15 units while the articulation layer consists of 10×10 ones. Each element of an

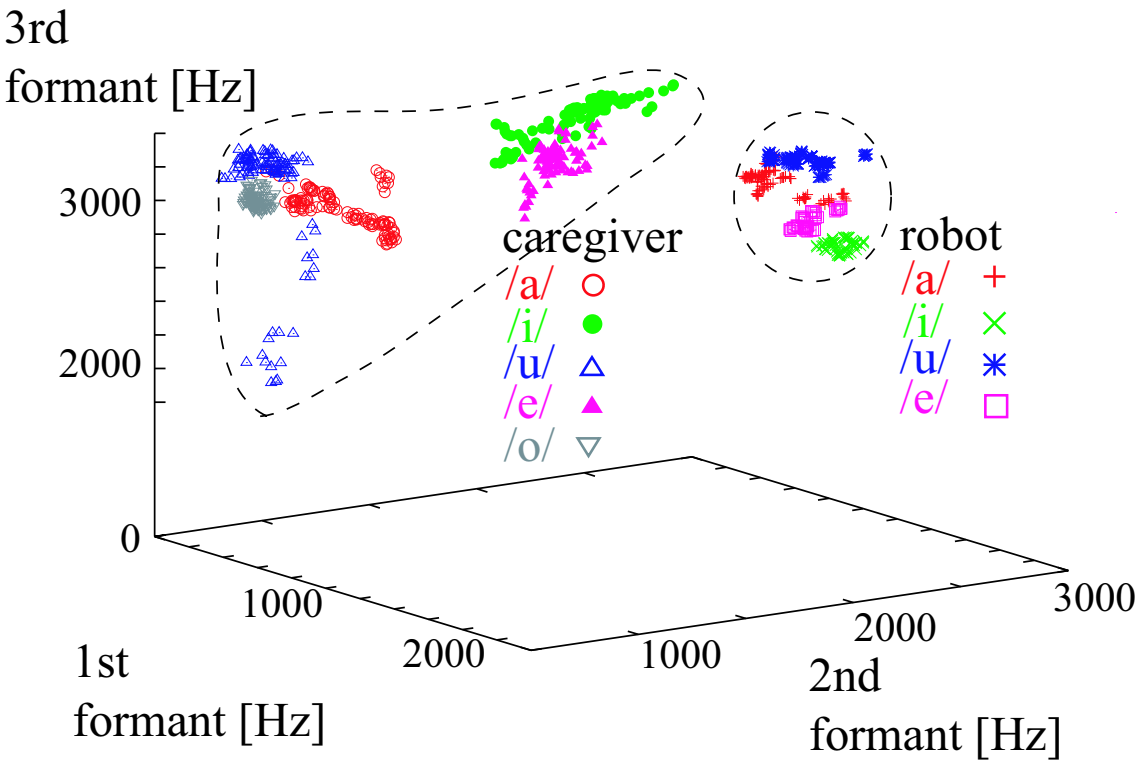
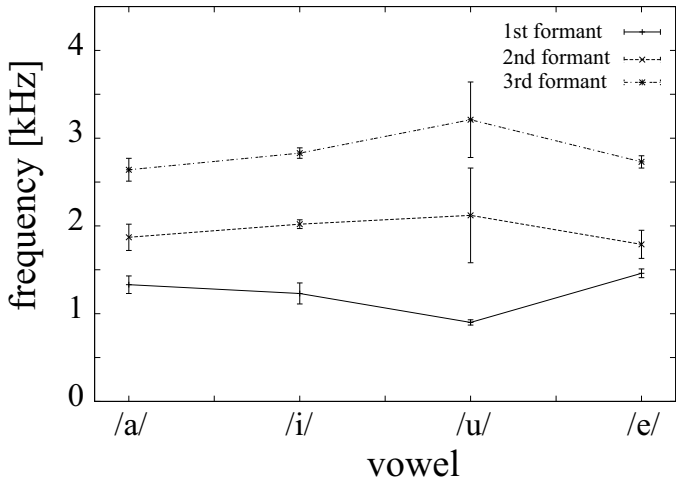
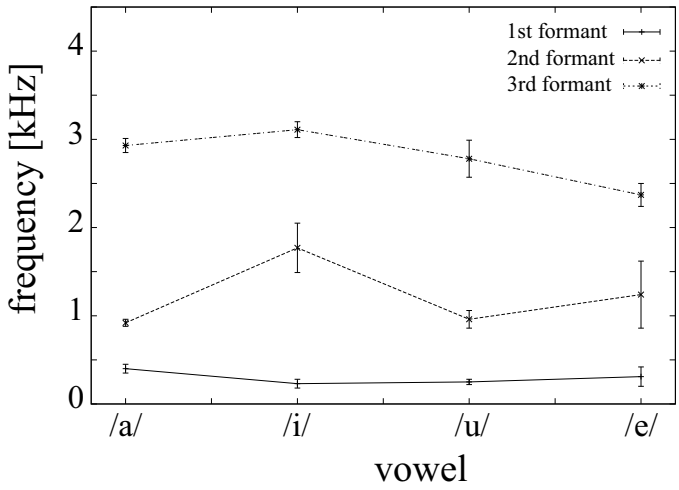


Figure 4.7: Formant distributions of the robot and the caregiver



(a) The robot



(b) The caregiver

Figure 4.8: Averages and standard deviations of formants of vocalized sound by (a) the robot and (b) the caregiver

Figure 4.9: Experimental setup: the robot interacts with a human caregiver.

articulation vector is quantized into five levels; these elements are the motor commands of the random articulation mechanism. If the robot's vocalization sounds like a vowel, the caregiver utters the corresponding vowel (see Fig. 4.9), then the robot calculates formants of the caregiver's utterance and updates the codebook vectors and connections. In the following experiments, the caregiver repeated 39 vocalizations. The 39 training examples were used for learning, iterately 50 times for each.

4.5.1 Learning without the toil criterion

To examine how the robot acquired vowels by the learning rule without the toil criterion, that is based on eq. (4.6), we observe which units in the articulation layer are activated by the propagation from the auditory layer after learning. At first, the caregiver utters one of the four vowels. Then, activations occur in the auditory layer and are propagated to the articulation layer through the connections. We observed which units in the articulation layer are activated by the propagation because the activated units can be regarded as vowels matching those of the caregiver. Fig. 4.10(a) shows distribution of the articulation vectors of the most strongly activated units by 30 input, which are the caregiver's utterances, for each vowel. They are compressed onto a two di-

mensional plane by principal component analysis (PCA). For comparison, the distribution of all articulation vectors that cause vowel-like sounds is shown in Fig. 4.10(b).

We can see that the distribution of the activated articulation vectors are parts of the region in which articulation vectors causes vowel-like sounds. Furthermore, the propagation for the input vowels in the same category activates the articulation vectors in a cluster that causes corresponding sounds. It means that the robot succeed in learning to match its articulations with the caregiver's vowels. However, we can see that there is arbitrariness in selecting the matched articulation with the caregiver's utterances since a unit in the auditory layer has multiple connections.

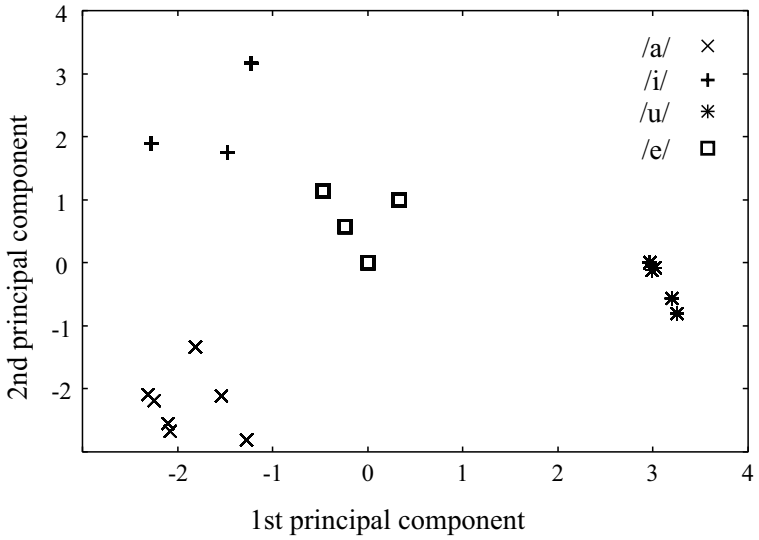
4.5.2 Learning with the toil criterion

We next examine how the robot acquires vowels by the learning rule with the toil criterion that is based on eq.(4.7). As in the previous experiment, we observe which units in the articulation layer are activated by the auditory layer through the learned connections. Fig. 4.11 shows the articulation vectors of the most strongly activated units by 30 inputs for each vowel. They are compressed onto a two dimensional plane by PCA.

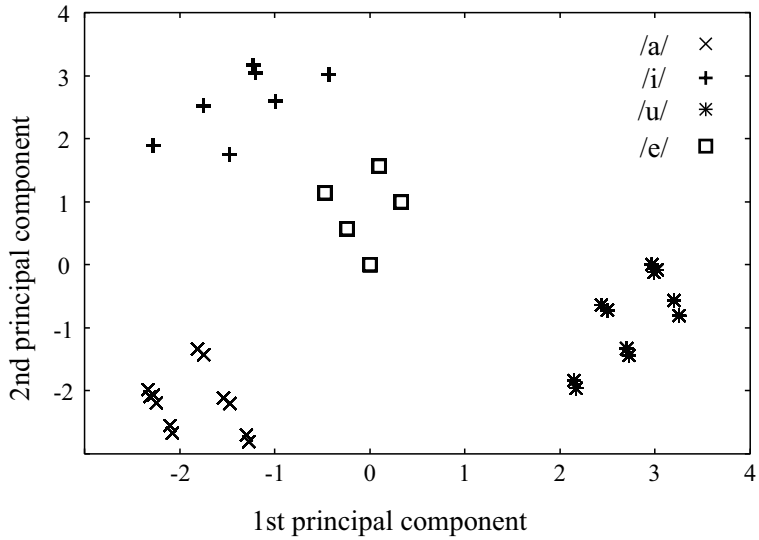
We can see that fewer articulation vectors are selected than in Fig. 4.10(a) while the articulation vectors that causes corresponding vowel-like sounds are selected. Therefore, we confirmed that the robot can match its articulations with the caregiver's vowels by the learning rule with the toil criterion and, furthermore, that the toil criteria decrease the arbitrariness. The selected articulation vectors involving less toil. The acquired articulations are shown in Fig. 4.12.

4.6 Summary and discussion

In this chapter, we have proposed a learning model of vowel acquisition implemented by a robot with different articulation parameters but without *a priori* knowledge about the relationship between the sensorimotor system and phonemes. There are some related studies concerning vocalization in robotics. Nishikawa et al. built a series of anthropomorphic robots which can produce Japanese phonemes including consonant sounds [86, 87]. However, the articulations were tuned manually by the human designer since they have not addressed the issue of acquisition. Higashimoto and Sawada built a system which learns to vocalize human vowels [80]. It consists of a vibrator that



(a) Activated articulation vectors by listening to the caregiver's utterances.



(b) All articulation vectors that sound like vowel

Figure 4.10: Distribution of the articulation vectors in the two major principal component space: (a) that of the activated (selected) units by the propagation of heard caregiver's utterances through the connection learned without the toil criterion, and (b) that of all articulation vectors that causes vowel-like sounds

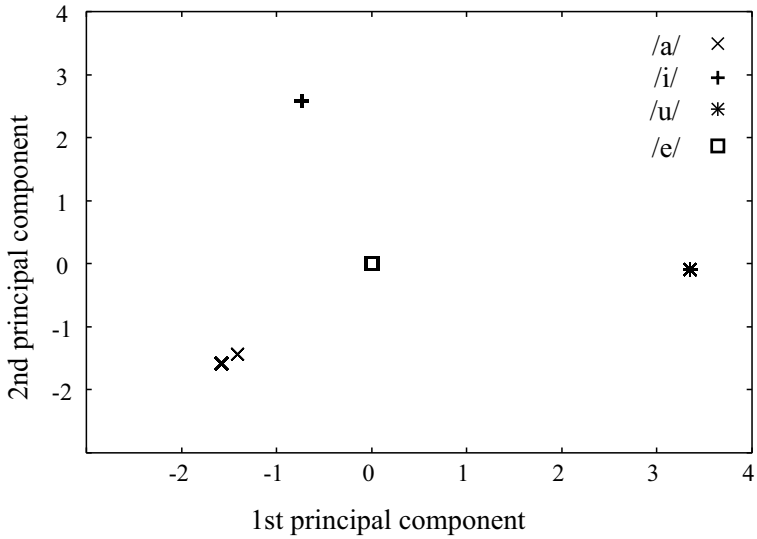
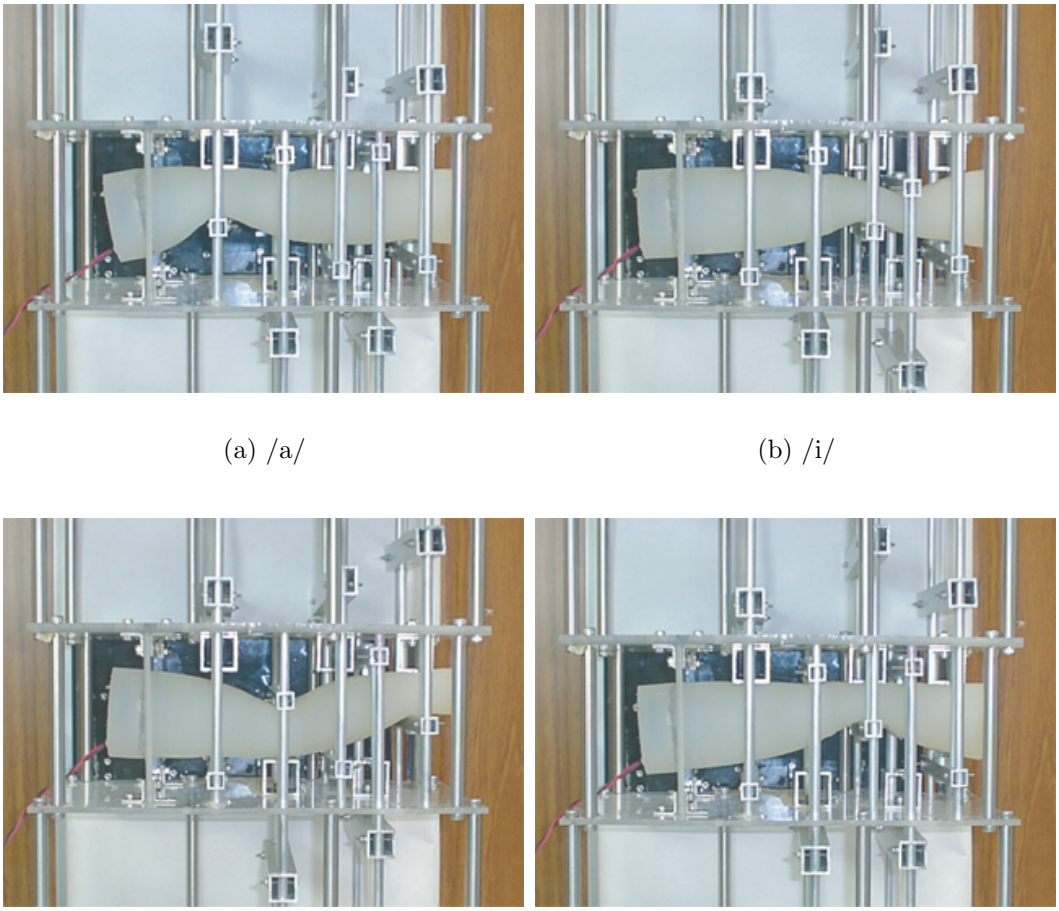


Figure 4.11: Distribution of the activated (selected) articulation vectors in the two major principal component space by the propagation of the heard caregiver's utterances through the connection learned with the toil criterion

generates a source sound and a deformable silicone rubber tube to modulate the source sound. It can vocalize vowels similar to a human's by learning the inverse model of articulation parameters with respect to a spectrum envelop of human utterances. In other words, it performs a imitation based on the similarity of raw soundwave since they assume that the robot can reproduce the sound with the same spectrum envelop as the human utterance. Nishikawa's group has also done the optimization of the parameters of articulation to mimic based on similarity of raw soundwave [88]. Vocalization based on recording/playback systems, for example [89], could be also regarded as imitation in the same vein. However, imitation based on the similarity of raw soundwave cannot be equated with the vowel acquisition processes of infants since they cannot reproduce the caregiver's physical soundwave form as they are because of their immaturity. Unlike the previous studies, we have proposed a more cognitively plausible alternative implemented by a robot that acquires vowels without the capability of reproducing the human physical soundwave forms.

When an agent tries to imitate the behavior of other agent with a different body structure, it needs to abstract observed behavior to some extent since it cannot duplicate it as it is. However, abstraction brings arbitrariness into



(a) /a/ (b) /i/

(c) /u/ (d) /e/

Figure 4.12: Appearances of the acquired articulations

the imitation process — even if the agent acquires pairs describing its own behavior and that of the caregiver. We proposed a method to cope with this arbitrariness by introducing subjective criteria, that is how much toil does the articulation involve. As we showed in the second experiment, the toil criteria reduced the arbitrariness of the matched articulations. This kind of subjective criteria could play an important role in imitation, understanding the behavior of others, and communicative processes between agents that have different bodies since there are arbitrariness to be resolved in these situations.

In the following, we first discuss why we use a real robot approach instead of a synthesizer. Then, to show the directions of our future work, we discuss the validity of the proposed model concerning the internal mechanism of the robot and the behavior of the caregiver.

A real robot approach In this study, we implemented a mechanical system as the mechanism to vocalize vowels because it cannot reproduce human physical soundwave form. Furthermore, such a robotic approach may draw attention to the fact that the robot and the caregiver have different bodies. Even if infants can reproduce the utterances of the caregiver as they are, infants perceive their own reproduced soundwaves differently from the caregiver's original one. Because, these two soundwaves travel different pathways: the caregiver's soundwave travels only through the air while the infant's travels through both the air and his or her bodies (See figure 4.13). As a result, soundwaves at the infant's auditory sensors are different from each other. This fact makes imitation based on the similarity of raw soundwaves difficult. However, the current robot is not designed to address this issue. Making a new robot with a microphone inside its body is a topic for our future work.

An Internal mechanism In the proposed model, we adopt a mechanism for producing random articulations. However, we should consider the motivation of the robot since infants seem to develop owing to their desire to communication. In other words, we cannot ignore the aspect of cooing as a communication media. What kind of motivational force can be applied to the robot that acquires vowels towards communication? Masataka [90] reported that, after vocalizing spontaneously, three- to four-month-old infants tend to pause as if they anticipate the response from their caregivers and vocalize repeatedly in the absence of a response. He also speculated that an infant can adjust the frequency of its sound in reaction to its mother's responses. Therefore, anticipation seems to be an important element of motivational force to explore. Addressing how to implement the anticipation mechanism that drives

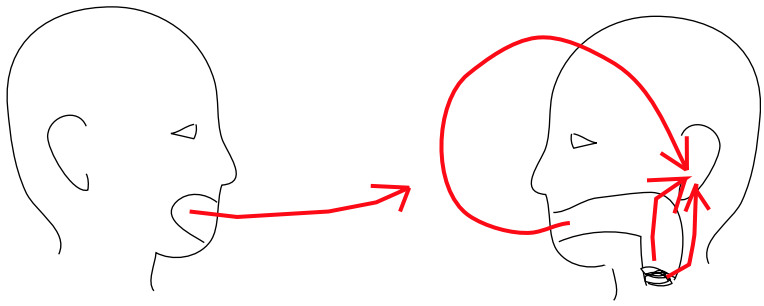


Figure 4.13: Conceptual figure indicating the difference between the pathway of the caregiver's soundwave to the infant's ear and that of the infant's soundwave to his or her own ear. The arrows indicate the soundwaves.

exploration is a topic for our future work. As Kaplan and Oudeyer [91] argued the importance of a non-task-specific value system to last developing, addressing how to modulate a motivational force may be also important. Furthermore, from the computational viewpoint, the strategy of exploration by the random mechanism is not suitable if the system has many degrees of freedom. Issues on reducing the computational cost for exploration are related to issues of motivational force since a specific motivational force can be regarded as a sort of bias for exploration.

The caregiver's behavior In the proposed model, the robot-caregiver interaction is simplified: the caregiver always utters the vowel that matches the cooing of the robot if the cooing can be regarded as a vowel. However, the proposed learning method does not always require the caregiver's repetition since it extracts clusters utilizing the statistical consistency in the data. That is, the method works only if the caregiver tends to be engaged in the repetitive utterances. Furthermore, this simplification is unrealistic because the human caregivers usually talk to infants with adult language, that is words or sentences.

Coping with words is important also from the viewpoint of lip-reading. Since mother-infant interaction occurs face-to-face, the visual information of the caregiver's lip can be matched with vowels based on the proposed learning architecture that learns the invariant abstraction. However, since the vowel sounds are determined by the partially invisible shape of the tract, there may be no visual features that are invariant with specific vowels. In order to resolve the ambiguity of the one-shot vision, it seems a promising way to learn together

the sequence of vowels. Therefore, extending the proposed architecture so that it can cope with sequential vowels is one area for our future work.

Chapter 5

Conclusion and future work

In the previous chapters, we have addressed following three issues concerning a robot that can imitate the demonstration from a subjective viewpoint:

- The question in chapter 2 is how a robot can map observation of the demonstration to its corresponding motion only through mappings between its sensorimotor space. By virtue of the assumption of the similarity of the both bodies and the opt-geometric constraint between views, the demonstrator's view is recovered from the learner's view to observe the demonstration onto the learner's view to observe its self-body. To estimate the parameters to recover, the learner's self-view is utilized as an alternative to the demonstrator's one. Then, the recovered demonstration is reproduced by estimating the relationship between motor command and optical flow in the view.

- The question in chapter 3 is how a robot can find its body from its sensory data without any *a priori* knowledge about its body. The invariance of sensation in self body observation is modeled by using a mixture of two Gaussian distributions, and then, is utilized to judge whether it observes its own body in the current posture. Multiple sensory data can be complementarily utilized in the proposed method.

- The question in chapter 4 is how a robot can acquire behaviors, that is vocalization in this case, common to the partner of interaction without any *a priori* knowledge on the relationship between the behavior and its sensorimotor system or the capability to duplicate the partner's behavior. The robot acquired the mapping of the acoustic feature of interlocutor's vowel onto its articulation parameters by utilizing the in-

terlocutor's imitative response and its subjective criteria of the toil to articulate.

Although the proposed mechanisms cannot be equated with the universal one to imitate a variety of behaviors from a subjective viewpoint, we suppose that these studies are early steps toward it. Therefore, we discuss possible future directions on this topic.

5.1 Body mapping between different bodies

Building a robot that can imitate the demonstration through the observation with visual sensors seems an important issue. It is partially because such kind of capability enables users to provide the robot with behaviors by just showing how to do as usual in teaching other persons, or partially because of synthetic understanding of the brain mechanism of view-based imitation in humans. The method proposed in chapter 2 is limited to the case when the robot has the same body structure as the demonstrator. On the other hand, although it concerns imitation between dissimilar bodies, one proposed in chapter 4 is limited to the case of vowel. How can we extend them to perform view-based imitation between dissimilar bodies? In other words, what kinds of constraints can be utilized to obtain the references to learn mapping between dissimilar bodies in visual space?

Utilizing the imitative response of human beings Deducing from following evidences in the studies of psychology, we speculate that the interaction with a caregiver who imitatively responds to the robot's behavior could provide it with the references to learn mapping. Additional to the evidence of imitative response introduced in chapter 4 (e.g., maternal imitation of cooing [24, 25]), there are other evidences that support the existence of imitative response in the interaction. In psychology, it is well known that there is contagious facial expression [26, 27] or yawing [92] in interaction of humans. Furthermore, even chimpanzees also exhibit contagious yawing [28]. Apart from the facial movement, human beings [27] and even monkeys [29] are said to exhibit contagious postural movement. These evidences might indicate that when an infant interacts with an adult caregiver, the infant receives the matched behavior more frequently than mis-matched ones, and thereby, can utilize the matched experience to learn the mapping between bodies.

Furthermore, this process might be reciprocal, and thereby become more frequent as an infant develops to imitate the caregiver's behavior. Such recipro-

cal enhancement of imitative response, if true, is considered to bootstrap from the well-known neonatal imitation [40, 93]. To utilize the imitative responses, a robot should be able to induce human-directed responses from human beings. For that purpose, it is supposed that a robot should have sufficient close appearance [94] and expression [95, 96] to humans. Furthermore, the behavior and the appearance [94] might be necessary to be balanced [94] or matched with the robot's task [97].

Utilizing object affordance It is observed that infant show the capability to imitate the demonstration directed to object in early stage of his/her development [98, 99]. Object affordance might cause such experiences of sharing the behavior with other person since the constraint of the object help to reduce the candidate behaviors to be performed [100, 101]. If two agents possess similar capability to contact with the object, the possibility of sharing behavior increases compared to the case when they behave without contact. To utilized object affordance to cause the shared behavior between the robot and humans, it had better have human-like body structure that causes equivalent constraint for objects.

Utilizing joint attention *Joint attention* that can be defined as looking where someone else is looking [102], is usually supposed to be play an important role in communication. However, it also seems to play an important role in constructing the mapping between bodies since it causes the experience of not only sharing attention but also sharing the behavior of attention. Interestingly, a recent study on primate reported that Japanese monkeys which usually do not perform imitation start to imitate after training of joint attention [103]. To utilize joint attention as a constraint to learn the mapping of bodies, a robot should be able to perform joint attention with humans. It has been already proposed that it can acquire joint attention from subjective viewpoint by finding the contingency between the caregiver's face image and the position where the salient object exists instead of relying on explicit supervision by the designer [104, 105, 106].

Learning to map bodies from the constrains Even though an infant can experience to share the behavior with the caregiver through the reciprocal imitative response, object affordance, or joint attention, this does not guarantee the success in learning of the mapping between bodies. To perform learning, an infant must realize that he/she engages in sharing behavior with the caregiver since the experience of sharing does not always occur. It is said that older

than 14-month-old infant is sensitive and prefer to be imitated his/her action to objects [107, 108]. Furthermore, it has been reported that even six-week-old infant has acquired the sensitivity to be imitated his/her facial movement [107]. However, in the proposed method in chapter 4 or in the previous work that utilize the postural, imitative responses to learn the mapping [109], it was assumed that the human experimenter always imitatively responded to the robot's behavior or posture during the learning process. We should consider how a robot can pick up only the experiences of matched behaviors and ignore those of mis-matched behaviors in its learning process.

5.2 Multimodal representation of body parts

Some of the shared behaviors argued in the previous section concern the body parts such as a face or a head. This fact makes mapping difficult since these parts of the learner cannot be seen. Therefore, to consider how to construct the mapping involving these parts, the representation of the body should be composed of multi modalities. The multimodal representation of the body is also important from the viewpoint of constructing the representation of body parts, in other words how to group the sensation that originates from receiving the same part of the body. Although sensory features of close region are similar to each other [67], clustering based only on the visual similarity is considered to be problematic because an appearance of a body part would drastically change depending on its posture.

Binding problem in constructing body representation Therefore, constructing the body representation is deeply related to one of the most fundamental cognitive functions called *binding*, that is to find the correspondence of sensations between different modalities such as vision and touch. Although binding problem has been addressed in the field of brain science (see a survey [110]) and constructivist studies (ex. [111, 112]), they have usually focus how to integrate different attributes in the same modality such as color and shape in vision. However, when it tries to integrate the sensation in different modalities, it faces with the following problem: 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. For example, 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. [62, 60]).

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 be a key information to solve the binding problem as utilized in the chapter 3. In other words, 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. We have already proposed a method to match the foci of attention in different modalities, touch and vision, based on the fact that *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 [61, 71]. Although how to extrapolate the acquired mapping to unseen region should be considered, infants might be engaged in developing their body representation from early stage in their lives since neonate [113] or even fetus [73] have already exhibited double-touching.

Subjective criteria to make mapping one-to-one The method called cross-anchoring learning proposed in our previous work [71] is an extended type of Hebbian learning. The learning process is directed to be one-to-one mapping to reduce the many-to-many correspondence involved in the relationship between the occurrences of self-occlusion and of double-touching, and therefore, can be converged to exclusively connect possible pairs of self-occlusion and double-touching. The pressure to be one-to-one mapping is considered as one of the subjective criteria that concerns the simplicity of the system in the informational point of view and might also be utilized for learning the mapping between different bodies. As argued in the previous section, even if the robot can share behaviors with humans, the experience of sharing would not occur in every interaction since the humans would behave as they like. Therefore, it must pick up only the sharing experiences to use them to construct the mapping. Cross-anchoring learning is expected to find the correct one-to-one correspondence of behaviors involved in many-to-many concurrence of behaviors.

Bibliography

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

[2] C.G. Atkeson and S. Schaal. Learning tasks from a single demonstration. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1706–1712, 1997.

[3] S. Nolfi. Adaptation as a more poweful tool than decomposition and integration. In T. Fogarty and G. Venturini, editors, *Proceedings of the Workshop on Evolutionary Computing and Machine Learning, 13th International Conference on Machine Learning*, 1996.

[4] R. Pfeifer and C. Scheier. *Understanding Intelligence*, chapter 12 The Principle of Sensory-Motor Coordination. The MIT Press, 1999.

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

[6] K. Ikeuchi and T. Suehiro. Toward an assembly from observation. *IEEE Transaction on Robotics & Automation*, Vol. 10, pp. 368–385, 1994.

[7] S. Nakaoka, A. Nakazawa, K. Yokoi, H. Hirukawa, and K. Ikeuchi. Generating whole body motions for a biped humanoid robot from captured human dances. In *Proceedings of 2003 IEEE International Conference on Robotics and Automation*, pp. 3905–3910, 2004.

[8] D. C. Bentivegna, C. G. Atkeson, and G. Cheng. Learning tasks from observation and practice. *Robotics and Autonomous Systems*, Vol. 47, pp. 163–169, 2004.

[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] 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.

[11] T. Inamura, Y. Nakamura, and M. Shimosaki. Integration of behavior recognition and generation processes based on associative memory. In *Proceedings of the 19th Annual Conference of the Robotics Society of Japan (in Japanese)*, pp. 1237–1238, 2001.

[12] A. J. Ijspeert, 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.

[13] A. Billard and M. Mataric. Learning human arm movements by imitation: Evaluation of a biologically inspired connectionist architecture. *Robotics and Autonomous Systems*, Vol. 37, pp. 145–160, 2001.

[14] G. Di Pellegrino, L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Understanding motor events: a neurophysiological study. *Expermental Brain Research*, Vol. 91, pp. 176–180, 1992.

[15] L. Fadiga et al. Motor facilitation during action observation: a magnetic stimulation study. *Journal of Neurophysiology*, Vol. 73, pp. 2608–2611, 1995.

[16] M. Iacobini, R. P. Woods, M. Brass, H. Bekkering, J. C. Mazziotta, and G. Rizzolatti. Cortical mechanizm of human imitation. *Science*, Vol. 286, No. 24, pp. 2526–2528, 1999.

[17] N. Nishitani and R. Hari. Temporal dynamics of cortical representation for action. In *Proceedings of the National Academy of Sciences, USA*, Vol. 97, pp. 913–918, 2000.

[18] G. Buccino, F. Binkofski, G.R. Fink, L. Fadiga, L. Fogassi, V. Gallese, R.J. Seitz, K. Zilles, G. Rizzolatti, and H.J. Freund. Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience*, Vol. 13, pp. 400–404, 2001.

[19] G. Rizzolatti and M. A. Arbib. Language within our grasp. *Trends in Neuroscience*, Vol. 21, p. 188, 1998.

[20] L. Fadiga, L. Fogassi, V. Gallese, and G. Rizzolatti. Visuomotor neurons: ambiguity of the discharge or 'motor' perception? *International Journal of Psychophysiology*, Vol. 35, pp. 165–177, 2000.

[21] V. Gallese and A. Goldman. Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Science*, Vol. 2, No. 12, pp. 493–501, 1998.

[22] 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.

[23] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Develpmental robotics: a survey. *Connection Science*, Vol. 15, No. 4, pp. 151–190, 2003.

[24] M. Pélaez-Nogueras, J. L. Gewirtz, and Michael. M. Markham. Infant vocalizations are conditioned both by maternal imitation and motherese speech. *Infant behavior and development*, Vol. 19, p. 670, 1996.

[25] N. Masataka. *The Onset of Language*, chapter The development of vocal imitation. Cambridge University Press, 2003.

[26] U. Dimberg. Facial reactions to facial expressions. *Psychophysiology*, Vol. 19, No. 6, pp. 643–647, 1982.

[27] U. Hess, P. Philippot, and S. Blairy. Mimicry - facts and fiction. In Philippot et al., editor, *The social context of Nonverbal Behavior*, pp. 213–241. Cambridge University Press, 1999.

[28] J. R. Anderson, M. Myowa-Yamakoshi, and T. Matsuzawa. Contagious yawning in chimpanzees. *Biology letters*, 2004.

[29] K. Nakayama. Observing conspecifics scratching induces a contagion of scratching in japanese monkeys (macaca fuscata). *Journal of Comparative Psychology*, Vol. 118, pp. 20–24, 2004.

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

[31] 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.

[32] Y. Kuniyoshi. The science of imitation -towards physically and socially grounded intelligence-. In *Proceedings in Real World Computing (RWC) Joint Symposium*, 1994.

[33] P. Bakker and Y. Kuniyoshi. Robot see, robot do : An overview of robot imitation. In *Proceedings of Artificial Intelligence and the Simulation of Behaviour (AISB) Workshop on Learning in Robots and Animals*, 1996.

[34] J. Tani. Learning to generate articulated behavior through the bottom-up and the top-down interaction processes. *Neural Networks*, Vol. 16, No. 1, pp. 11–23, 2003.

[35] K. Samejima, K. Doya, and M. Kawato. MOSAIC reinforcement learning architecture: Symbolization by predictability and mimic learning by symbol. *Journal of the Robotics Society of Japan (in Japanese)*, 2001.

[36] G. Hayes and J. Demiris. A robot controller using learning by imitation. In *Proceedings of the 2nd International Symposium on Intelligent Robotic Systems*, pp. 198–204, 1994.

[37] K. Dautenhahn. Getting to know each other - artificial social intelligence for autonomous robots. *Robotics and Autonomous Systems*, Vol. 16, pp. 333–356, 1995.

[38] 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.

[39] Y. Kuniyoshi, Y. Yorozu, M. Inaba, and H. Inoue. From visuo-motor self-learning to early imitation – a neural architecture for humanoid learning. In *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 3132–3139, 2003.

[40] A.N. Meltzoff and M.K. Moore. Imitation of facial and manual gestures by human neonates. *Science*, Vol. 198, pp. 75–78, 1977.

[41] P. Andry, P. Gaussier, and J. Nadel. From visuo-motor development to low-level imitation. In *Proceedings of the second International Workshop on Epigenetic Robotics*, pp. 7–15, 2002.

[42] M. Ogino, S. Matsuyama, J. Ooga, and M. Asada. Motion recognition and generation for humanoid based on visual-somatic field mapping. In *Proceedings of the Third International Conference on Development and Learning*, 2004.

[43] C. Nehaniv and K. Dautenhahn. Mapping between dissimilar bodies: Affordances and the algebraic foundations of imitation. In *Demiris/Birk, Proceedings of the European Workshop on Learning Robots*, pp. 64–72, 1998.

[44] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion and Object Recognition: A Unified Approach.* Kluwer Academic Publisher, 1996.

[45] Y. Matsutmoto, T. Shibata, K. Sakai, M. Inaba, and H. Inoue. Real-time color stereo vision system for a mobile robot based on field multiplexing. In *Proceedings of IEEE International Conference on Robotics and Automation*, pp. 1934–1939, 1997.

[46] K. Hiraki, A. Sashima, and S. Phillips. From egocentric to allocentric spatial behavior : A computational model of spatial development. *Adaptive Behavior*, Vol. 6, No. (3/4), pp. 371–391, 1998.

[47] J.J. Kuffner, S. Kagami, K. Nishiwaki, M. Inaba, and H. Inoue. Dynamically-stable motion planning for humanoid robots. *Autonomous Robots*, Vol. 12, No. 1, pp. 105–118, 2002.

[48] F. Kanehiro, H. Hirukawa, K. Kaneko, S. Kajita, K. Fujiwara, K. Harada, and K. Yokoi. Locomotion planning of humanoid robots to pass through narrow spaces. In *Proceeding of the 2004 IEEE International Conference on Robotics*, pp. 604–609, 2004.

[49] 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.

[50] A. Maravita and A. Iriki. Tools for the body (schema). *Trends in Cognitive Sciences*, Vol. 8, No. 2, pp. 79–86, 2004.

[51] L. Berthouze and S. Itakura. Possibility of self-recognizing robots: From the perspective of research on nonhuman primates. *Japanese Journal of Cognitive Studies: Consciousness: Toward a Cognitive Science of Brain and Mind (Special Issue)*, Vol. 4, No. 3, pp. 120–127, 1997.

[52] P. Rochat. *Blackewell Handbook of infant development*, chapter Origins of Self-concept. Blackwell Publishing Ltd, 2004.

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

[54] J.S. Albus. A new approach to manipulator control: The cerebellar model articulation controller (CMAC). *Transactions of the ASME. Journal of Dynamic System, Measurement, and Control*, Vol. 97, pp. 220–227, 1975.

[55] M. Kawato, K. Furukawa, and R. Suzuki. A hierachical neural-network model for control and learning of voluntary movement. *Biological Cybernetics*, Vol. 57, pp. 169–185, 1987.

[56] M. Kuperstein. Neural model of adaptive hand-eye coordination for single posture. *Science*, Vol. 239, pp. 1308–1311, 1988.

[57] M. Jordan and D.E. Rumelhart. Forward models: Supervised learning with a distal teacher. *Cognitive Science*, Vol. 16, pp. 307–354, 1992.

[58] P. Morasso and V. Sanguineti. Self-organizing body schema for motor planning. *Journal of Motor Behavior*, Vol. 27, No. 1, pp. 52–66, 1995.

[59] D. M. Wolpert and M. Kawato. Multiple paired forward and inverse models for motor control. *Neural Networks*, Vol. 11, pp. 1317–1329, 1998.

[60] A. Sotytchev. Development and extension of the robot body schema. In *Proc. of the third Intl. Workshop on Epigenetic Robotics*, pp. 179–180, 2003.

[61] Y. Yoshikawa, H. Kawanishi, M. Asada, and K. Hosoda. Body scheme acquisition by cross modal map learning among tactile, image, and proprioceptive spaces. In *Proceedings of the Second International Workshop on Epigenetic Robotics*, pp. 181–184, 2002.

[62] K. F. MacDorman, K. Tatani, Y. Miyazaki, and M. Koeda. Protosysmbol emergence. In *Proceedings of the International Conference on Intelligent Robot and Systems*, pp. 1619–1625, 2000.

[63] M. Marjanovic, B. Scassellati, and M. Williamson. Self-taught visually-guided pointing for a humanoid robot. In *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, 1996.

[64] G. Metta and P. Fitzpatrick. Early integration of vision and manipulation. *Adaptive Behavior*, Vol. 11, No. 2, pp. 109–128, 2003.

[65] P. Michel, K. Gold, and B. Scassellati. Motion-based robotic self-recognition. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2763–2768, 2004.

[66] 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.

[67] D. Pierce and B. Kuipers. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence*, Vol. 92, pp. 169–229, 1997.

[68] Y. Kuniyoshi, Y. Yorozu, Y. Ohmura, K. Terada, T. Otani, A. Nagakubo, and T. Yamamoto. From humanoid embodiment to theory of mind. In F. Iida et al., editor, *Embodied Artificial Intelligence*, Vol. 3139 of *LNCS/AI series*, pp. 202–218. Springer, 2004.

[69] D. Philipona, J.K. O'Regan, and J.-P. Nadal. Is there something out there? inferring space from sensorimotor dependencies. *Neural Computation*, Vol. 15, No. 9, pp. 2029–2049, 2003.

[70] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, Vol. B-39, pp. 1–38, 1977.

[71] Y. Yoshikawa, Koh Hosoda, and Minoru Asada. Cross-anchoring for binding tactile and visual sensations via unique association through self-perception. In *Proceedings of the third International Conference on Development and Learning*, 2004.

[72] P. Rochat and S.J. Hespos. Differential rooting response by neonates: Evidence for an early sense of self. *Early Development and Pareting*, Vol. 6, pp. 105–112, 1997.

[73] J. W. Sparling, J. V. Tol, and N. C. Chescheir. Fetal and neonatal hand movement. *Physical Therapy*, Vol. 79, No. 1, pp. 24–39, 1999.

[74] H. Papousek and M. Papousek. Mirror-image and self recognition in young infants: a new method of experimental analysis. *Developmental Psychobiology*, Vol. 7, pp. 149–157, 1974.

[75] L.E. Bahrick and J.S. Watson. Detection of intermodal proprioceptive-visual contingency as a potential basis of self-perception in infancy. *Developmental Psychology*, Vol. 21, pp. 963–973, 1985.

[76] A.L. van der Meer, F.R. van der Weel, and D.N. Lee. The functional significance of arm movements in neonates. *Science*, Vol. 267, No. 3, pp. 693–695, 1995.

[77] B. de Boer. Self-organization in vowel systems. *Journal of Phonetics*, Vol. 28, pp. 441–465, 2000.

[78] P.-Y. Oudeyer. Phonemic coding might result from sensory-motor coupling dynamics. In *Proceedings of the 7th international conference on simulation of adaptive behavior (SAB02)*, pp. 406–416, 2002.

[79] N. Masataka and K. Bloom. Acoustic properties that determine adult's preference for 3-month-old infant vocalization. *Infant Behavior and Development*, Vol. 17, pp. 461–464, 1994.

[80] T. Higashimoto and H. Sawada. Speech production by a mechanical model construction of a vocal tract and its control by neural network. In *Proceedings of the 2002 IEEE International Conference on Robotics & Automation*, pp. 3858–3863, 2002.

[81] R. K. Potter and J. C. Steinberg. Toward the specification of speech. *Journal of the Acoustical Society of America*, Vol. 22, pp. 807–820, 1950.

[82] W. Tecumseh Fitch. The evolution of speech: a comparative review. *Trends in Cognitive Science*, Vol. 4, No. 7, pp. 258–267, 2000.

[83] T. Kohonen. *Self-Organization and Assosiative Memory.* Springer-Verlag, New York, 1984.

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

[85] P. Rubin and E. Vatikiotis-Bateson. *Animal Acoustic Communication*, chapter Measuring and modeling speech production. Springer-Verlag, New York, 1998.

[86] K. Nishikawa, K. Asama, K. Hayashi, and A. Takanishi H. Takanobu. Development of a talking robot. In *Proccedings of the IEEE International Conference on Intelligent Robot and Systems*, pp. 1760–1765, 2000.

[87] K. Nishikawa, H. Takanobu, T. Mochida, M. Honda, and A. Takanishi. Development of a new human-like talking robot having advanced vocal tract mechanisms. In *Proceedings of the 2003 IEEE International Conference on Robotics & Automation*, pp. 1907–1913, 2003.

[88] K. Nishikawa, T. Kuwae, H. Takanobu, T. Mochida, M. Honda, and A. Takanishi. Mimicry of human speech sounds using an anthropomorphic talking robot by auditory feedback. In *Proceedings of the 2003 IEEE International Conference on Robotics & Automation*, pp. 272–278, 2004.

[89] T. Kanda, H. Ishiguro, M. Imai, T. Ono, and K. Mase. A constrcutive approach for developing interactive humanoid robots. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1265–1270, 2002.

[90] N. Masataka. Effects of contingent and noncontingent maternal stimulation of vocal behavior of 3 to 4-month-old japanese infants. *Journal of Child Language*, Vol. 20, pp. 303–312, 1993.

[91] F. Kaplan and P.-Y. Oudeyer. Motivational principle for visual know-how development. In *Proceedings of the Third International Workshop on Epigenetic Robotics*, pp. 73–80, 2003.

[92] S. M. Platek, S. R. Critton, T. E. Myers, and G. G. Gallup Jr. Contagious yawning: the role of self-awareness and mental state attribution. *Cognitive Brain Research*, Vol. 17, pp. 223–227, 2003.

[93] M. Myowa-Yamakoshi, M. Tomonaga, M. Tanaka, and T. Matsuzawa. Imitation in neonatal chimpanzees (pan troglodytes). *Developmental Science*, Vol. 7, No. 4, pp. 437–442, 2004.

[94] T. Minato, M. Shimada, H. Ishiguro, and S. Itakura. Development of an android robot for studying human-robot interaction. In *Innovations in Applied Artificial Intelligence; Proc. of Seventeenth International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 424–434, 2004.

[95] C. Breazeal. Toward sociable robots. *Robotics and Autonomous Systems*, Vol. 42, pp. 167–175, 2003.

[96] H. Kozima, C. Nakagawa, Y. Yasuda, and D. Kosugi. A toy-like robot in the playroom for children with developmental disorder. In *Proceedings of the Third International Conference on Development and Learning*, 2004.

[97] J. Goetz, S. Kiesler, and A. Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *Proceedings of the 12th IEEE Workshop on Robot and Human Interactive Communication. RO-MAN 2003*, 2004.

[98] A. N. Meltzoff. Infant imitation after a 1-week-delay: Long term memory for novel acts and multiple stimuli. *Developmental Psychology*, Vol. 24, pp. 470–476, 1988.

[99] G. Gergely, H. Bekkering, and I. Király. Rational imitation in prevebal infants. *Nature*, Vol. 415, p. 755, 2002.

[100] E. Oztop and M. Arbib. Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics*, Vol. 87, pp. 116–140, 2002.

[101] H. Kozima, C. Nakagawa, and H. Yano. Emergence of imitation mediated by objects. In *Proceedings of the Second International Workshop on Epigenetic Robotics*, 2002.

[102] G. Butterworth. Joint visual attention in infancy. In Gavin Bremner and Alan Fogel, editors, *Blackwell Handbook of Infant Development*, chapter 8, pp. 213–240. Blackwell publishing, 2001.

[103] M. Kumashiro, H. Ishibashi, Y. Uchiyama, S. Itakura, A. Murata, and A. Iriki. Natural imitation induced by joint attention in japanese monkeys. *International Journal of Psychophysiology*, Vol. 50, pp. 81–99, 2003.

[104] Y. Nagai, K. Hosoda, A. Morita, and M. Asada. A constructive model for the development of joint attention. *Connection Science*, Vol. 15, No. 4, pp. 211–229, 2003.

[105] A. Morita, Y. Yoshikawa, K. Hosoda, and M. Asada. Joint attention with strangers based on generalization through joint attention with caregivers. In *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 3744–3749, 2004.

[106] C. Teuscher and J. Triesch. To care or not to care: Analyzing the caregiver in a computational gaze following framework. In *Proceedings of the Third International Conference on Development and Learning*, 2004.

[107] A. N. Meltzoff and M. K. Moore. A theory of the role of imitation in the emergence of self. In P. Rochat, editor, *The Self in Infancy: Theory and Reseach*, chapter 5, pp. 73–93. Elsevier Science, 1995.

[108] B. Agnetta and P. Rochat. Imitative games by 9-, 14-, and 18-month-old infants. *Infancy*, Vol. 6, No. 1, pp. 1–36, 2004.

[109] A. Stoica. Robot fostering techniques for sensory-motor development of humanoid robots. *Robotics and Autonomous Systems*, Vol. 37, pp. 127–143, 2001.

[110] A. Treisman. Solutions to the binding problem: Progress through controversy and convergence. *Neuron*, Vol. 24, pp. 105–110, 1999.

[111] G. Tononi, O. Sporns, and G.M. Edelman. Reentry and the problem of integrating multiple cortical areas: Simulation of dynamic integration in the visual system. *Cerebral Cortex*, Vol. 2, pp. 310–335, 1992.

[112] A.K. Seth, J.L. McKinstry, G.M. Edelman, and J.L. Krichmar. Visual binding, reentry, and neuronal synchrony in a physically situated brain-based device. In *Proceedings of the 3rd International Workshop on Epigenetic Robotics*, pp. 177–178, 2003.

[113] A.F. Korner and H.C. Kraemer. Individual differences in spontaneous oral behavior in neonates. In J.F. Bosma, editor, *Third symposium on oral sensation an perception*, pp. 335–346. US Department of Health Education, and Welfare Publication, 1972.

Publications by the Author

Journal papers

- Y. Yoshikawa, M. Asada, K. Hosoda, and J. Koga, "A Constructive Approach to Infant's Vowel Acquisition through Mother-Infant Interaction", Connection Science, Vol.15, No.4, pp.245-258, 2003.

- Y. Yoshikawa, M. Asada, and K. Hosoda, Imitation based on Demonstrator's View Recovery Utilizing Epipolar Geometry, Journal of the Robotics Society of Japan (in Japanese), Vol.22, No.1, pp.68-74, 2004.

- Y. Yoshikawa, K. Hosoda, M. Asada, and Y. Tsuji, Body finding based on the invariance in multiple sensory data, Journal of the Robotics Society of Japan (in Japanese), (submitted).

Conference papers with reviews

- M. Asada, Y. Yoshikawa, and K. Hosoda. Learning by Observation without Three-Dimensional Reconstruction. In *Proc. of the 6th International Conference on Intelligent Autonomous Systems* , pp. 555-560, 2000.

- Y. Yoshikawa, M. Asada, and K. Hosoda. View-Based Imitation Learning by Conflict Resolution with Epipolar Geometry. In *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 1416-1427, 2001.

- Y. Yoshikawa, M. Asada, and K. Hosoda. Developmental approach to Spatial Perception for Imitation Learning: Incremental Demonstrator's View Recovery by Modular Neural Network. In *Proc. of the 2nd IEEE-RAS International Conference on Humanoid Robots* , pp. 107-114, 2001.

- Y. Yoshikawa, Hiroyoshi Kawanishi, M. Asada, and K. Hosoda, Body Scheme Acquisition by Cross Modal Map Learning among Tactile, Image, and Proprioceptive Spaces, In Proc. of the 2nd International Workshop on Epigenetic Robotics, Scotland, August, pp.181-184, 2002.

- Y. Yoshikawa, Yoshiki Tsuji, M. Asada, and K. Hosoda, View-based Imitation with Rotation Invariant Pan-Tilt Stereo Cameras, In Proc. of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.159-164, 2002.

- Y. Yoshikawa, K. Hosoda, and M. Asada, Does the invariance in multi-modalities represent the body scheme? - a case study with vision and proprioception -, In Proc. of the 2nd Intl. Symposium on Adaptive Motion of Animals and Machines, CD-ROM, Sa-P-II-1, 2003.

- Y. Yoshikawa and J. Koga and M. Asada, and K. Hosoda, A Constructive Model of Mother-Infant Interaction towards Infant's Vowel Articulation, In Proc. of the 3rd International Workshop on Epigenetic Robotics, pp.139-146, 2003.

- Y. Yoshikawa and J. Koga and M. Asada and K. Hosoda, Primary Vowel Imitation between Agents with Different Articulation Parameters by Parrot-like Teaching, In Proc. of the 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.149-154, 2003.

- Y. Yoshikawa, M. Asada and K. Hosoda, Towards Imitation Learning from a Viewpoint of an Internal Observer, F. Iida et al.(Ed.), LNCS/AI series, vol.3139, Embodied Artificial Intelligence, Springer, pp.278-283, 2004.

- Y. Yoshikawa, K. Hosoda, and M. Asada, Binding tactile and visual sensations via unique association by cross-anchoring between double-touching and self-occlusion, In Proc. of the 4th International Workshop on Epigenetic Robotics, pp.135-138, 2004.

- Y. Yoshikawa, Y. Tsuji, K. Hosoda, and M. Asada, Is it my body? - body extraction from uninterpreted sensory data based on the invariance of multiple sensory attributes -, In Proc. of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.2325-2330, 2004.

- A. Morita, Y. Yoshikawa, K. Hosoda, and M. Asada, Joint attention with strangers based on generalization through joint attention with caregivers, In Proc. of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004.
- Y. Yoshikawa, K. Hosoda, and M. Asada, Cross-anchoring for binding tactile and visual sensations via unique association through self-perception, In Proc. of the 3rd International Conference on Development and Learning, 2004.