Understanding the Development of Joint Attention from a Viewpoint of Cognitive Developmental Robotics

（共同注意発達の理解－認知発達ロボティクスの視点から）

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

Yukie Nagai

January 2004

Thesis Supervisor:	Minoru Asada
Title:	Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University
Thesis Committee:	Minoru Asada, Chair Yoshiaki Shirai Hiroshi Ishiguro Koh Hosoda

Preface

This work has been carried out by Yukie Nagai from 1999 to 2003 under the supervision of Professor Minoru Asada at the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Japan.

Abstract

The objectives of the study described in this dissertation are

- to construct models for a robot to acquire the ability of joint attention through interactions with its environment based on knowledge from cognitive developmental science, and

- to understand the process by which human infants acquire the ability of joint attention through the realization process of the first objective.

Joint attention is a process to look at an object which someone else is looking at. The ability of joint attention is a cornerstone for infants to interact with others and to acquire social communication abilities. A number of studies in cognitive developmental science have explained the developmental phenomena of infants' joint attention. However, their developmental mechanisms have not been revealed yet. On the other hand, cognitive developmental robotics, which is a new interdisciplinary research field between cognitive developmental science and robotics, has a potential to reveal how infants acquire the ability of joint attention by constructing an artificial model for a robot to learn joint attention. A robot that is designed to learn through its experiences is expected to acquire more various and adaptive capabilities than a robot that is fully-programmed by a designer.

This dissertation presents two kinds of constructivist models by which a robot acquires the ability of joint attention through interactions with its environment including a human caregiver from a viewpoint of cognitive developmental robotics. The proposed models are based on knowledge from cognitive developmental science that caregiver's evaluation makes a significant difference in infant's learning. In the case

that an infant learns *with* caregiver's evaluation, it should be discussed how the caregiver evaluates the infant for facilitating his/her learning of joint attention. On the other hand, in the case that an infant learns *without* any external evaluation, it becomes an issue what capabilities the infant should have for acquiring the ability of joint attention. Focusing on these points, the following two approaches are proposed.

1. The first approach is *a developmental learning model* with caregiver's evaluation. This model is based on the knowledge that a caregiver can facilitate an infant learning by adjusting the evaluation criterion according to the performance of the infant. At the same time, it is known that an infant matures his/her internal mechanisms so that it makes his/her own learning easier. These changes are generically called developments. Based on the knowledge, the developmental learning model for joint attention consists of a learning mechanism with caregiver's evaluation and developmental mechanisms of a caregiver and a robot. Experimental results show that the caregiver's development accelerates the learning of joint attention, and the robot's development improves the final task performance of itself.

2. The second approach is *a bootstrap learning model* without any external evaluation. This model is based on the knowledge that an infant inherently has various capabilities, e.g. preferences for salient visual stimuli and contingency learning, and such capabilities enable the infant to acquire new abilities. The scheme of learning that is based on only the learner's innate or pre-acquired capabilities without any external evaluation is called bootstrap learning. The bootstrap learning model for joint attention embeds the mechanisms of visual attention and learning with self-evaluation on visual attention into a robot. Experimental results show that the proposed model enables the robot to acquire the ability of joint attention without any external evaluation and to reproduce a similar staged learning process to the infants'.

It is expected that the proposed models demonstrate the knowledge from cognitive developmental science and give some new suggestions to the understanding of how infants acquire the ability of joint attention.

Acknowledgments

I give an address of my gratitude here to all the people who have encouraged me and made this work possible.

First of all, I would like to appreciate Professor Minoru Asada, who is a principal adviser of this work. He has given me the best research environment where I can work on high level studies and meet many great researchers. The new research field of “cognitive developmental robotics,” which I am interested in and tackling now, was advocated by him and some researchers. His powerful leading and big supports have enabled me to accomplish this work.

Associate professor Koh Hosoda has given me detailed suggestions through this work. His valuable advice and constant encouragement have enabled me to fulfill my Ph.D. study. This work would not be possible without his support. I would like to extend my gratitude to him.

Professor Yoshiaki Shirai and Professor Hiroshi Ishiguro have spared their valuable time to discuss the meanings of the research on cognitive developmental robotics. They made me reacknowledge the difficulties and the significance to study cognitive developmental science from a viewpoint of robotics. Their meaningful suggestions gave me a chance to reaffirm my stance on the study. I am grateful to them.

I also thank all the present and the past members in Emergent Robotics Area at Osaka University. Dr. Yasutake Takahashi has listened to my problems and encouraged me to carry on my study. Dr. Noriaki Mitsunaga has cleared some hardware problems of the experimental setup. Mr. Takashi Minato has given me valuable suggestions since I entered this laboratory. Mr. Masaki Ogino has supported me in diverse ways as a member who joined the laboratory in the same year. Mr. Yuichiro Yoshikawa has spent the time for meaningful discussions with me about the study on cognitive developmental robotics. Mr. Akio Morita has assisted me to conduct some preliminary experiments. Other members who are not mentioned here also have supported me to accomplish my study.

The last half part of this study was funded by the Advanced and Innovational Research program in Life Sciences from the Ministry of Education, Culture, Sports, Science, and Technology of the Japanese Government.

Professor Ken Tomiyama at Aoyama Gakuin University gave me the first opportunity to tackle the robotics research. He told me the enjoyment of investigating new problems and teaching the acquired knowledge to other people. Thanks to him, I could find a goal in the future that is to become a good teacher as well as a researcher. I am sincerely grateful to him.

I appreciate Ms. Yoko Maekawa for her kind help. She improved my English in this dissertation.

I would like to thank to many friends from university days. They have made a big push for me to fulfill my dream.

Last, but most importantly, I would like to say “どうもありがとう” a lot to my family. My parents, Miyoshi and Toshie, have allowed and supported me to spend a long student life and encouraged me to complete my study. They have warmly watched over self-willed me all the time. My big sister, Sachiko, has given me a good example. Her efforts in studying have directed my efforts. None of this work and my life would be possible without their great love. Thank you so much.

January 7, 2004
Yukie Nagai

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Human infants are born with various capabilities. Innate capabilities of infants might be determined genetically through the evolutionary process of human beings, or might be acquired when they were in their mothers' wombs. In either case, such capabilities enable infants to interact with their surroundings, especially their caregivers, and consequently to acquire more various and advanced cognitive functions through the interactions (see Figure 1.1). The cognitive developmental processes of infants in the first year of their lives are extremely complicated. The researchers in cognitive science, developmental science, psychology, and philosophy have investigated the cognitive developments of infants for a long time. As science and technology progress, the researchers in neuroscience have started to examine the cognitive mechanisms in infants' brain. Recently, many robotics researchers have referred to the cognitive developmental mechanisms of infants for building human-like intelligent robots from engineering standpoints. The issues in which these researchers are interested are organized as follows.

(1) What ability do infants acquire?

(2) When do infants acquire the ability?

(3) Who can acquire the ability and who cannot?

(4) Where in a brain do infants acquire the ability?

(5) How do infants acquire the ability?

Figure 1.1: A human infant and his/her caregiver [Newman and Newman, 2003]. The infant interacts with the caregiver based on his/her innate capabilities and acquires more various and advanced ones through the interactions.

About the first three issues, “what,” “when,” and “who,” the researchers in cognitive science, developmental science, and psychology have made a number of findings through years of studies. The researchers observed infants in both usual living environments and experimental ones, and analyzed the infants’ behaviors when the infants were performing various tasks. Their cross-sectional and longitudinal studies have revealed (1) what abilities infants acquire through their experiences and (2) when infants acquire the abilities. At the same time, the researchers in cognitive science and developmental science examined a variety of infants, who had lived in various surroundings and/or had some disorders. The efforts from various viewpoints enabled the researchers to explain (3) who can acquire the abilities and who cannot, that is, what environments allow infants to acquire the abilities and what disorders prevent that. The findings in these studies have suggested what genetic factors and environmental ones are significant for infants to develop their cognitive functions.

Neuroscientist have newly came up with answers to the fourth question "where" above the findings in cognitive science, developmental science, and psychology. The researchers in neuroscience have measured the brain activities of infants by optical topography. This study has revealed (4) where in a brain infants acquire various capabilities, i.e. cerebral functional localization. It has been indicated that the cerebral functional localization is determined in almost the same manner in all infants even though they grow up in different surroundings. However, it has been also suggested that the area in a brain without any inputs may bear other cognitive functions. In other words, the brain of infants has plasticity. The findings in neuroscience have represented a step on the way to understand the cognitive developments of infants not as the phenomena but as the mechanisms.

About the last issue "how," the researchers in the new research field of developmental cognitive neuroscience [Johnson, 1997] are seeking its answers. Developmental cognitive neuroscience is located at the interface between developmental psychology and cognitive neuroscience. The former investigates the change of cognitive functions during infancy from a biological viewpoint, and the latter examines the construction of an increasingly complex brain. By organically integrating these two fields, the researchers in developmental cognitive neuroscience expect to reveal (5) how infants acquire various abilities as a brain system.

The researchers in cognitive developmental robotics [Asada *et al.*, 2001] are also addressing the issue "how" as well as those in developmental cognitive neuroscience. This research field is a new interdisciplinary one between cognitive developmental science and robotics. The researchers in this field design cognitive developmental models for their robots and environments based on knowledge from cognitive developmental science, and then implement the models into the robots. The robots interact with the environments based on the embedded capabilities, and develop and learn through the interactions like human infants. The validity of the cognitive developmental models is evaluated from both viewpoints of the developmental phenomena and the mechanisms. This approach, called a constructivist approach, is a basic idea of cognitive developmental robotics and enables us to understand the developmental mechanisms of infants through constructing artificial models more clearly.

This study addresses the issue (5) how human infants acquire their cognitive functions through interactions with their environments from a viewpoint of cognitive developmental robotics. It aims at understanding the cognitive developmental mechanisms of infants through constructing artificial models by which a robot develops cognitive functions like infants inspired by knowledge from cognitive science, developmental science, and neuroscience. At the same time, it is expected to realize more intelligent and adaptive robots like human beings than the robots that are fully-programmed by designers.

This chapter first describes the essence of human intelligence. It is discussed what kind of essence should be considered in investigating human intelligence from a viewpoint of cognitive developmental robotics. Then, the fundamental idea and the approach of cognitive developmental robotics are explained. This study focuses on the problem of the development of joint attention among various cognitive developments of infants. The following section describes the features of joint attention and shows the significance to address the problem of the development of joint attention. Further, the objective and the overview of this dissertation are given.

1.1 Human Intelligence

Human intelligence has various features. The features should be carefully considered in studying the cognitive development of humans from a standpoint of robotics. Cognitive developmental robotics, which is a new methodology to investigate human intelligence from a constructivist approach, places special emphasis not on the intelligence of adults but on that of infants since they develop and learn more than adults. It is discussed what kind of essence should be focused on to investigate human intelligence from a viewpoint of cognitive developmental robotics.

Brooks *et al.* [1998] and Pfeifer and Scheier [1999] have proposed the essence of human intelligence to which engineers should pay their attention when they design human-like intelligent robots inspired by knowledge from cognitive science, developmental science, and neuroscience. They have pointed out the importance that intelligence has a physical body, i.e. embodiment. This section describes embodiment and three more features of human intelligence: development, learning, and social interaction, based on the discussions in [Brooks *et al.*, 1998; Pfeifer and Scheier, 1999].

Embodiment

A human being has a physical body and interacts with environments through the body. The intelligence of humans can be evaluated only when the humans generate their behaviors and interact with environments. It means that human intelligence is defined not only by programs in the brain but also by the complexity of behaviors, which could be changed by environments. Intelligence does not make any sense without a body. Such a relationship between intelligence and a body is called "embodiment" [Brooks, 1991; Pfeifer and Scheier, 1999]. Humans have embodiment and interact with environments in various ways. Embodiment is significant for humans not only to express their intelligence but also to acquire intelligence. Humans develop and learn through interactions with environments. The interactions, which enable the humans to acquire a variety of cognitive functions, are also based on embodiment.

Inversely, humans' bodies also require intelligence. The bodies have appropriate functions for the intelligence. The functions are determined so that they make the best use of the intelligence and enable humans to acquire more advanced intelligence. It should be not too complex or too simple but suitable for intelligence. It means that humans' bodies develop as their intelligence develops. In investigating human intelligence from a viewpoint of cognitive developmental robotics, the researchers are required to utilize human-like robots which have the similar structures in the degree of freedom, sensing capabilities, and actuating ones, and develop these functions.

Development and Learning

Human beings develop and learn throughout their entire lives. Although humans are born with a variety of innate capabilities, the capabilities are not sufficient for the humans to adapt themselves to a human society and live in. Therefore, humans incrementally acquire new capabilities through experiences in environments based on their innate capabilities. For example, humans develop physically and functionally as time goes on. The height and the weight of a body increase, and the accuracy to control own body becomes better. Perceptual functions such as vision and auditory also gradually improve. All of these changes seem to shift toward more differentiated, complicated, and advanced states. Such a change is called "development." Development depends on not only genetic factors but also environmental ones. The degrees of humans' developments are determined based on their experiences in the environments as well as genetic informations. A variety of experiences in the environments are significant for the developments of humans.

Another essence of human intelligence is "learning." Humans acquire various capabilities through learning. The ability to use language is the greatest one that humans acquire through learning. The ability of joint attention, which is discussed in this dissertation, is also acquired through learning based on interactions with environments. These abilities can not be acquired individually. In other words, a higher capability is acquired based on lower capabilities. The ability of joint attention is suggested to lead to the acquisition of the ability of language [Baldwin, 1995; Mundy and Gomes, 1998; Morales *et al.*, 1998; Morales, 2000], theory of mind [Moore and Corkum, 1994; Baron-Cohen, 1995; Charman *et al.*, 2000], the ability of imitation [Kumashiro *et al.*, 2003], and so on. It should be taken into account that learning as well as development play significant roles in human intelligence.

Social Interaction

Human beings "socially interact" with environments, especially, humans interact with other humans. Interactions with others are crucially important for human intelligence. Humans cannot always acquire their intelligence by themselves through interactions with environments. Rather, they acquire their intelligence through the teaching and the evaluation by others. Infants are known to learn a variety of capabilities from their caregivers. The ability of joint attention is one of the capabilities that infants acquire through interactions with their caregivers, and also the ability to use language. Social interactions have a crucial role in the development of human intelligence.

As described in the paragraph of embodiment, human intelligence is evaluated through interactions, that is, the intelligence of humans is evaluated not by themselves but by others. Others evaluate the value of the intelligence of humans based on interactions between them. Such evaluation form has an advantage/disadvantage to estimate the intelligence at high/low levels depending on the intelligence of the evaluators. In other words, the relative strength of the interactions defines the value of intelligence. Social interactions play important roles in the evaluation of intelligence as well as the development.

This section has described the essence of human intelligence: embodiment, development, learning, and social interaction. Of course, human intelligence has various features beyond them. This study focuses on these four issues since they are considered to be the most significant for the cognitive development of human infants. Next section describes the basic principle of cognitive developmental robotics.

1.2 Cognitive Developmental Robotics

The new research field *cognitive developmental robotics* was advocated by Asada *et al.* [2001]. Cognitive developmental robotics is an interdisciplinary field between cognitive developmental science and robotics, and strives to organically integrate them. The aims of studies in cognitive developmental robotics are organized as follows:

- to reveal the cognitive developmental process of humans, and
- to build robots that develop their cognitive functions like humans.

About the first issue, the cognitive developmental process of humans, especially human infants, have been investigated in the research fields of cognitive science, developmental science, psychology, neuroscience, and so on. Studies in these areas have revealed a number of findings about the cognitive development of infants from observational approaches and/or analytical ones. The former makes inferences about the development of cognitive functions of infants based on enormous quantity of data of their behavioral experiments. On the other hand, the latter seeks the brain mapping of their cognitive functions by measuring their brain activities when they are performing some tasks. These approaches, however, have some problems in understanding the cognitive developments of infants as their mechanisms. To observe the behaviors of infants is not enough to picture their internal mechanisms. The brain activities of infants could be interpreted in several ways; furthermore, the brains are not sufficient to define their intelligence. In contrast with observational approaches and analytical ones, cognitive developmental robotics takes constructivist approaches to reveal the cognitive developments of infants. Through the approaches, the researchers construct artificial models for the development of cognitive functions based on knowledge from cognitive developmental science, then implement them into robots, and finally understand the cognitive development of infants by observing and analyzing the process of interactions between the robots and environments from both phenomenological and functional viewpoints. To construct artificial models means that designers have the thorough understanding of the mechanisms. Furthermore, to implement the models into robots enables the designers to understand the robots' intelligence not only as programs but also as their behaviors. Cognitive developmental robotics aims at understanding the developmental mechanisms of humans' cognitive functions through the constructivist approaches.

About the second issue, a number of studies to build robots or systems with human-like intelligence have been conducted in the research fields of robotics and artificial intelligence. These studies have realized robots or systems which show high performance in particular tasks. However, the most studies have aimed at building adult-like robots that already acquired intelligence, not infant-like robots that develop their intelligence. As mentioned in the previous section, development is one of striking features of human intelligence. Intelligence without development is not complete. Besides, to construct artificial intelligence just like adults' seems more difficult than to construct intelligence like infants'. It is considered that the cognitive functions of infants are relatively simple and easy to be modeled in robots. In addition, it is extremely interesting to investigate how robots which have only primary cognitive functions like infants acquire advanced ones like adults through interactions with environments. The acquired cognitive functions are expected to be more intelligent and adaptive than those fully-programmed by designers. For these reasons, to build robots that develop their cognitive functions is challenging problem in cognitive science, developmental science, neuroscience, and robotics. The principle to design such robots is shown in Figure 1.2. In studies of cognitive developmental robotics, it should be discussed (a) how to design the mechanisms embedded in a robot so that it can develop and learn and (b) how to build an environment that supports the

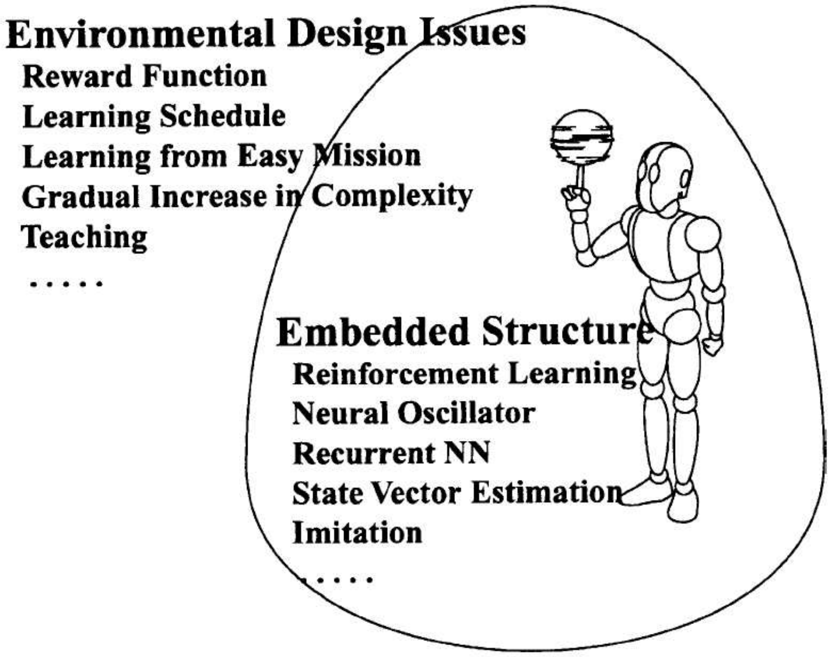


Figure 1.2: The design principle of cognitive developmental robotics [Asada *et al.*, 2001]. From the standpoint of engineering, it should be discussed (a) how to design the mechanisms embedded in a robot so that it can develop and learn and (b) how to build an environment that supports the robot.

robot. Only when both points are designed appropriately, the robot becomes possible to develop and learn its cognitive functions.

Analogous principles to cognitive developmental robotics had been advocated before. Turing [1950], who invented a turing test to evaluate artificial intelligence of computers, described in his paper as follows:

> Instead of trying to produce a program to simulate the adult mind, why not rather try to produce one which simulates the child's?
>
> [Turing, 1950]

He divided the problem on building artificial intelligence like human adults into two: child program and education process, and emphasized that they should strongly relate each other. Brooks *et al.* [1998] presented a novel methodology for constructing human-like artificially intelligent systems. They suggested alternative four essence of intelligence: development, social interaction, embodiment, and integration, based on evidence from cognitive science and neuroscience. Kozima and Zlatev [2000] and Metta *et al.* [2000] also proposed similar methodologies to build infant-like robots. All of these approaches have the same principle as that of cognitive developmental robotics. It means that a number of researchers expect to understand human intelligence and to build human-like intelligent systems by addressing the problems in the cognitive development of humans from constructivist approaches.

The following section explains joint attention, which this study focuses on as one of the problems of cognitive developments, and discusses the meanings to address the problem.

1.3 First Step in Social Intelligence: Joint Attention

This study focuses on joint attention, which is one of the primary cognitive functions of human infants, and discusses how infants acquire the ability of joint attention through interactions with their surroundings from a viewpoint of cognitive developmental robotics. Joint attention is defined as a process to look at an object that someone else is looking at [Butterworth, 1991]. Refer to Section 2.1.1 for more detail descriptions. The ability of joint attention is a crucial first step of nonverbal communications and social intelligence, and enables infants to encounter other people, especially to form a triadic interaction with their caregivers. The triadic interactions allow infants to learn various knowledge from their caregivers and to acquire the verbal ability, the mind-reading ability, and so on. This section describes the significance to investigate the development of joint attention in infants from a viewpoint of cognitive developmental robotics.

A Cornerstone for the Development of Social Intelligence

As mentioned above, the ability of joint attention is crucial for infants to develop their cognitive functions from that time. Infants who do not have the ability of joint attention can realize only *dyadic* interaction, not *triadic* interaction. It means that infants who lack the ability of joint attention are not able to learn knowledge from their caregivers. This lack might cause them damage in the acquisition of language [Baldwin, 1995; Mundy and Gomes, 1998; Morales *et al.*, 1998; Morales, 2000].

Humans estimate others' intention, desire, knowledge, and belief by reading the gaze of others. To estimate the internal representations of others facilitate humans communicating with others. Moreover, before the ability, human infants are suggested to find out about others and to acquire self-other consciousness [Reddy, 2003] by detecting the gaze and the approach of others directing toward the infants or objects. It means that the lack of the ability of joint attention poses a significant problem to infants in recognizing others, who have different "minds" from the infants. The ability to attribute others' behaviors to their minds is called *theory of mind*. The development of theory of mind is considered to have a strong relationship with the ability of joint attention [Moore and Corkum, 1994; Baron-Cohen, 1995; Charman *et al.*, 2000] as well as the language development. In fact, infants with autism are conjectured not to have theory of mind because of the lack of the ability of joint attention [Baron-Cohen, 1995]. Infants with Williams syndrome are also known not to have the ability of joint attention. The evidence described here indicates that the ability of joint attention is significant for the development of infants' cognitive functions.

Including Key Issues of the Development of Social Intelligence

It is considered that the developmental process of joint attention includes key issues of the development of other cognitive functions. First, the ability of joint attention is conjectured to be closely related to innate capabilities of infants since the development of joint attention or gaze following, which is a precursor of joint attention, is found in infants shortly after birth. Innate capabilities allow infants to interact with environments and to acquire various cognitive functions, such as joint attention, through the interactions. If infants congenitally lack some innate capabilities, there are negative effects on the development of cognitive functions. Besides, it is a challenging problem not only in cognitive science and developmental science but also in medical science to reveal the relationship between inherent capabilities and posteriori ones. Also, between two different posteriori capabilities, infants have interactive linkages as well as anteroposterior ones. It is indicated that the ability of joint attention develops from birth to 18 months old [Butterworth and Jarrett, 1991] (more detail description is found in Section 2.1.2). At the same time, infants develop their body, sensing and actuating mechanisms, and cognitive functions. Therefore, it is conjectured that these capabilities interact in diverse ways, and it should be discussed how the developments of these abilities affect each other.

Another significant issue of the cognitive developments is interactions between infants and caregivers. When infants learn their cognitive functions, caregivers play important roles in their learning. Caregivers facilitate the infants' learning by teaching and evaluating them. The development of joint attention is considered to be one of the first interactions in which infants learn based on caregivers' evaluation. It should be also discussed how infants acquire their cognitive functions through interactions with their caregivers.

Many Findings about the Development of Joint Attention

Joint attention has attracted interests of the researchers in cognitive science, developmental science, and psychology and has been investigated for a long time [Moore and Dunham, 1995]. The researchers have made a number of findings about joint attention. The findings help us to generate hypotheses of the development of joint attention and to design artificial models for robots. In cognitive developmental robotics, constructing artificial models is a main issue; therefore, all elements in the hypotheses must be based on the findings about infants' cognitive developments. Knowledge about not only joint attention but also its related cognitive functions are necessary to construct artificial models. In this point, the studies on joint attention have been conducted from a variety of viewpoints; thus, it is considered that they have found enough data to construct the developmental models for robots.

After constructing artificial developmental models for joint attention and implementing them into robots, the validity of the models should be evaluated through experiments. It must be examined whether the hypotheses for constructing the models are appropriate, and whether infants could have such models or not. It is useful for the evaluation of the models to compare the developmental process of robots' joint attention with that of infants. The robots which have the developmental models are expected to generate similar developmental process of joint attention as infants. In the evaluation, knowledge acquired in cognitive science and developmental science is helpful. Because of a number of findings about joint attention, the studies on the development of joint attention from a viewpoint of cognitive developmental robotics become meaningful ones.

Various Applications

Robots that have the ability of joint attention are able to be applied in many ways. The ability of joint attention enables robots to communicate with humans without using any language. Humans would attribute *minds* to robots if the robots follow the directions of humans' gaze and look at the same object that the humans are looking at. Of course, the robots do not have any higher cognitive functions such as minds; however, humans' minds allow them to feel the robots' minds. Such advantage of the ability of joint attention and the embodiment of the robots should be utilized in many applications and should be discussed in more detail.

The realization of the robots with the ability of joint attention has a significant role in investigating the developments of more advanced cognitive functions in infants from a viewpoint of cognitive developmental robotics. Evidence from cognitive science and developmental science have suggested that the ability of joint attention allows infants to acquire theory of mind [Moore and Corkum, 1994; Baron-Cohen, 1995; Charman *et al.*, 2000], the language ability [Baldwin, 1995; Mundy and Gomes, 1998; Morales *et al.*, 1998; Morales, 2000], and the ability of imitation [Kumashiro *et al.*, 2003]. These abilities are essential for infants to participate in human society as social agents and to acquire more higher cognitive functions. It is also interesting to investigate how infants acquire these abilities based on that of joint attention.

As described here, studies on the development of infants' joint attention from a viewpoint of cognitive developmental robotics have great meanings for cognitive science, developmental science, and robotics. Our study investigates the development of joint attention from a viewpoint of cognitive developmental robotics in consideration of the significance of joint attention described above.

1.4 Overview

The aims of the study described in this dissertation are the followings:

- to construct the models by which a robot acquires the ability of joint attention based on knowledge from cognitive developmental science, and
- to understand the developmental mechanisms of infants' joint attention through the realization process of the first objective.

The studies in cognitive developmental science have made a number of findings about infant's development and learning. It is known that one of the significant factors in the process of infant's learning is caregiver's evaluation. The caregiver's evaluation makes a great difference in the process of the infant's learning. If an infant learns *with* caregiver's evaluation, the infant will be facilitated the learning owing to the evaluation. In this case, it should be discussed how the caregiver evaluates the infant for more facilitating his/her learning. On the other hand, if an infant learns *without* any external evaluation, the infant might have a difficult time learning compared to the former case with evaluation. However, it is considered that the infant has some potentials to acquire new abilities by himself/herself. In this case, it becomes an issue what capabilities the infant should have for acquiring new abilities.

Focusing on these points, this dissertation presents two kinds of constructivist models by which a robot acquires the ability of joint attention through interactions with a human caregiver from a viewpoint of cognitive developmental robotics. The proposed models are

1. *a developmental learning model with caregiver's evaluation*, and
2. *a bootstrap learning model based on robot's embedded mechanisms.*

The former model is based on the knowledge that a caregiver can facilitate an infant learning by adjusting the evaluation criterion according to the performance of the infant. At the same time, it is known that an infant matures his/her internal mechanisms so that it makes his/her own learning easier. These changes that become more advanced or matured state are called developments. By applying the knowledge, the developmental learning model for joint attention consists of a learning mechanism with caregiver's evaluation and developmental mechanisms of the caregiver and a robot. The experiments of this model evaluate how the caregiver's development and the robot's development facilitate the learning of the robot's joint attention. On the other hand, the latter model is based on the knowledge that an infant inherently has various capabilities, e.g. preferences for salient visual stimuli and contingency learning, and such capabilities enable the infant to acquire new abilities. The scheme of learning which is based on only innate or pre-acquired abilities without any external evaluation is called bootstrap learning. The bootstrap learning model for joint attention embeds the mechanisms of visual attention and learning with self-evaluation on visual attention into a robot. The experiments of this model examine whether the robot can acquire the ability of joint attention based on the proposed model without any external evaluation. Through the realization process of the two models above, it is expected to find some new suggestions for the understanding of the developmental mechanisms of infants' joint attention.

This dissertation consists of six chapters including this one. The outlines of the chapters are the followings:

Chapter 1. Introduction

The grand challenge of this study is to understand human intelligence as well as to realize human-like intelligent robots from a viewpoint of cognitive developmental robotics. In this chapter, first, the essence of human intelligence was discussed. Then, the basic idea of cognitive developmental robotics was explained. As the first step to understand human intelligence, this study investigates the development of joint attention. This chapter described the significance to investigate joint attention and the aim of this study.

Chapter 2. Related Work

The findings about the development of joint attention from cognitive developmental science are described. Human infants are well known to have several innate capabilities and to develop their various cognitive functions based on the capabilities. Such capabilities related to joint attention are explained. Then, this chapter reviews previous work that has investigated social and/or developmental robots. Some of these robots have the embedded ability of joint attention to communicate with humans, and the others have developmental mechanisms like human infants. These studies are compared with our study from a viewpoint of cognitive developmental robotics.

Chapter 3. Joint Attention between a Robot and a Human Caregiver

The task definition of joint attention between a robot and a human caregiver is given. The robot has some functions to obtain several sensor inputs and to output its motor command. The learning objective of the robot is to acquire the sensorimotor coordination to achieve joint attention. Then, this chapter explains the concepts of the proposed two constructivist models by which a robot acquire the ability of joint attention. It is described what kind of knowledge from cognitive developmental science is utilized to construct the proposed models and how the models are constructed for verifying the knowledge.

Chapter 4. Developmental Learning with Caregiver's Evaluation

The first constructivist model, which is called a developmental learning model, is presented. It is known in cognitive developmental science that the development of a caregiver's criterion for task evaluation and that of an infant's internal functions facilitate the infant's learning. The proposed model evaluates how a caregiver's development and a robot's development facilitate the robot learning joint attention. Experimental results show the effectiveness of the proposed model.

Chapter 5. Bootstrap Learning based on Robot's Embedded Mechanisms

The second constructivist model, which is called a bootstrap learning model, is presented. It is suggested that an infant has potentials to develop his/her cognitive functions based on his/her innate capabilities without caregiver's help. The proposed model examines whether a robot can acquire the ability of joint attention based on its embedded mechanisms of visual attention and learning with self-evaluation on visual attention. Experimental results show the validity of the proposed model.

Chapter 6. Conclusions

> Finally, conclusions of this study and future work are given. The constructivist models proposed in this dissertation enable a robot to acquire the ability of joint attention and allow us to evaluate the validity of the knowledge from cognitive developmental science. However, they have several problems to be solved, e.g. online learning, two-way joint attention, and so on. These problems are discussed as near future work.

Joint attention is a small first step to understand human intelligence. However, most cognitive functions of infants are conjectured to be acquired through interactions with their caregivers in the same manner as joint attention. Infants learn much knowledge from their caregivers. This study is expected to be an important first step to understand human intelligence as well as to build human-like intelligent robots.

Chapter 2

Related Work

Development and learning abilities of human infants have interested the researchers in cognitive science, developmental science, psychology, neuroscience, and so on. The researchers in these fields have investigated infants from cross-sectional and longitudinal perspectives, and discovered many findings about infants. In particular, joint attention has been frequently addressed as the first developmental step of infants' social cognitive functions Moreover, the findings in these studies have motivated robotics researchers to build robots that develop and learn like infants. Robots that imitate infants are expected to acquire more advanced intelligence than robots that are fully-programmed by designers.

This chapter presents the knowledge about the developments of human infants, which has been indicated in cognitive developmental science, and reviews engineering approaches to build robots that develop like infants or communicate with humans based on the embedded abilities of joint attention. First, the definition of joint attention and the developmental process of infants' joint attention are explained. Observational findings that seem to relate to the development of joint attention are described. These findings and knowledge serve as foundations for the constructivist models proposed in Chapters 4 and 5. Then, this chapter reviews some robotics approaches to build social and/or developmental robots based on the knowledges from cognitive developmental science.

2.1 Findings about Human Infants from Cognitive Developmental Science

Joint attention and other various capabilities of human infants have been investigated in the research field of cognitive developmental science. This section describes observational findings of these studies. The findings give the bases for the proposed constructivist models by which a robot acquires the ability of joint attention like human infants.

2.1.1 Joint Attention

Joint attention is defined as a process to look at an object that someone else is looking at [Butterworth, 1991]. Strictly, joint attention in this definition means joint *visual* attention. To joint own attention with someone's attention can be achieved not only on a visual sense but also on other modalities, e.g. auditory and tactile senses. However, most publications discuss joint attention on the modality of a visual sense and use the term of "joint attention" instead of the term of "joint visual attention." We also define joint attention as that on the modality of a visual sense.

Joint attention between a human infant and his/her caregiver was first documented by Scaife and Bruner [1975]. They found out that an infant has a tendency to follow his/her caregiver's gaze. Gaze following between an infant and a caregiver is positioned as a preliminary step toward the development of the infant's joint attention. Butterworth and Jarrett [Butterworth and Jarrett, 1991; Butterworth, 1991; 1995] have studied infants' joint attention from a developmental perspective. They suggested that infants develop their ability of joint attention from 6 to 18 months old and show three developmental stages. The detailed description is in the next section. Butterworth [2000] told that joint attention of infants is not based on a theory of *mind* but based on a theory of *body*. It means that infants can realize joint attention without understanding others' minds. Baron-Cohen *et al.* [Baron-Cohen, 1995; Charman *et al.*, 2000] have investigated infants with/without autism and emphasized the importance of the developmental relationship between the ability of joint attention and a theory of mind. Theory of mind is a capability to attribute others' behaviors to their minds. In other words, this capability enables infants to understand others' purpose, knowledge, belief, thought, preference, and so on, and to socially interact with others. Infants with autism are known not to be able to understand others' minds because they do not have a theory of mind. Baron-Cohen [1995] has proposed a mindreading system, which includes the mechanisms of shared attention and a theory of mind, and suggested that the ability of shared attention becomes a preliminary step toward a theory of mind. Moore *et al.* [Moore and Corkum, 1994; Moore and Dunham, 1995] have also indicated that the ability of joint attention becomes a precursor of a theory of mind.

Many researchers have discussed infants' social attention including joint attention from various perspectives. Emery [2000] has classified such social attention as shown in Figure 2.1. The classification includes five kinds of social attention: mutual versus averted gaze, gaze following, joint attention, shared attention, and "theory of mind." This provides clearer definition of joint attention.

A. *Mutual gaze* is that the attention of individuals X and Y is directed to one another. *Averted gaze* is that individual X is looking at Y while the focus of the attention of Y is elsewhere.

B. *Gaze following* is that individual X detects that Y's gaze is not directed to X and follows the line of the gaze direction of Y.

C. *Joint attention* is that individual X follows the gaze direction of Y and looks at the same object that Y is looking at. This is the same process as gaze following except that there is a focus of attention, i.e. the object.

D. *Shared attention* is that individuals X and Y are looking at the same object and are aware of it each other, that is, individual X knows Y is looking at the object, and Y knows X is looking at the object. This is a combination of mutual gaze and joint attention.

E. *Theory of mind* is an ability based on the social attention and enables individual X to understand what Y is thinking about the object that they are looking at. This ability requires a combination of the previous A-D processes.

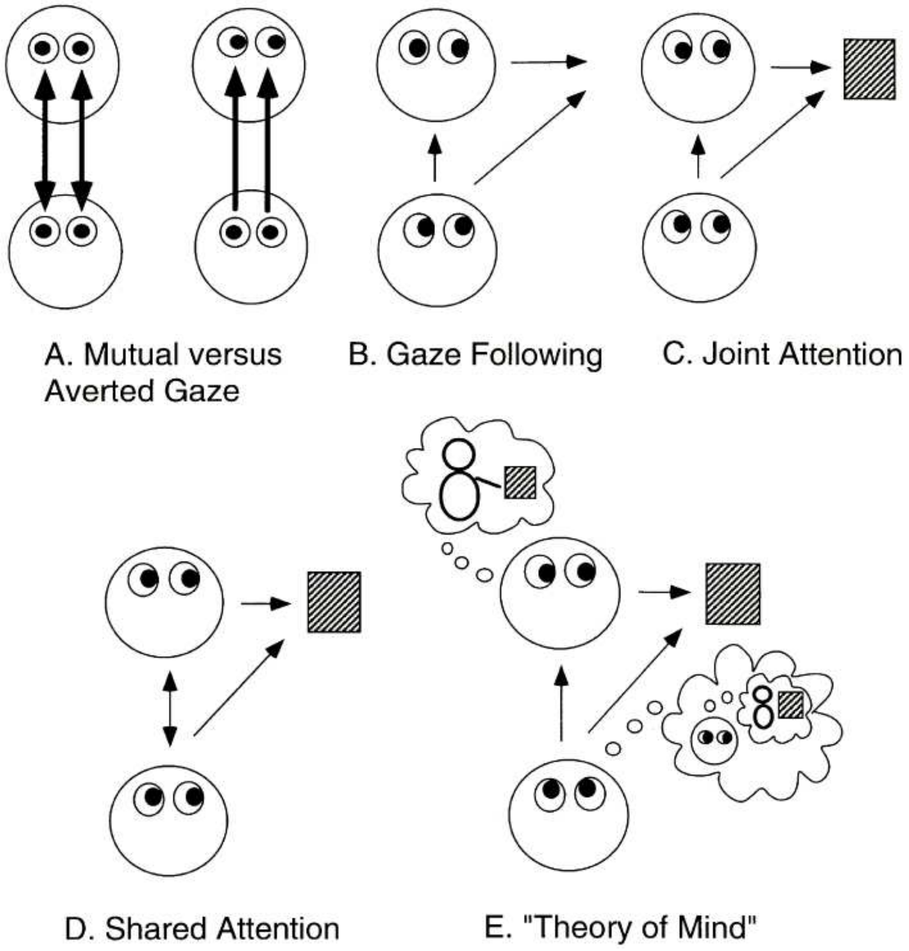


Figure 2.1: The classification of social attention [Emery, 2000]. Joint attention (C) is defined as a process to look at the same object that someone else is looking at. It is given in distinction from gaze following (B) and shared attention (D).

The definition of joint attention in this study is based on that by Butterworth [1991] and Emery [2000].

2.1.2 Staged Development of Joint Attention

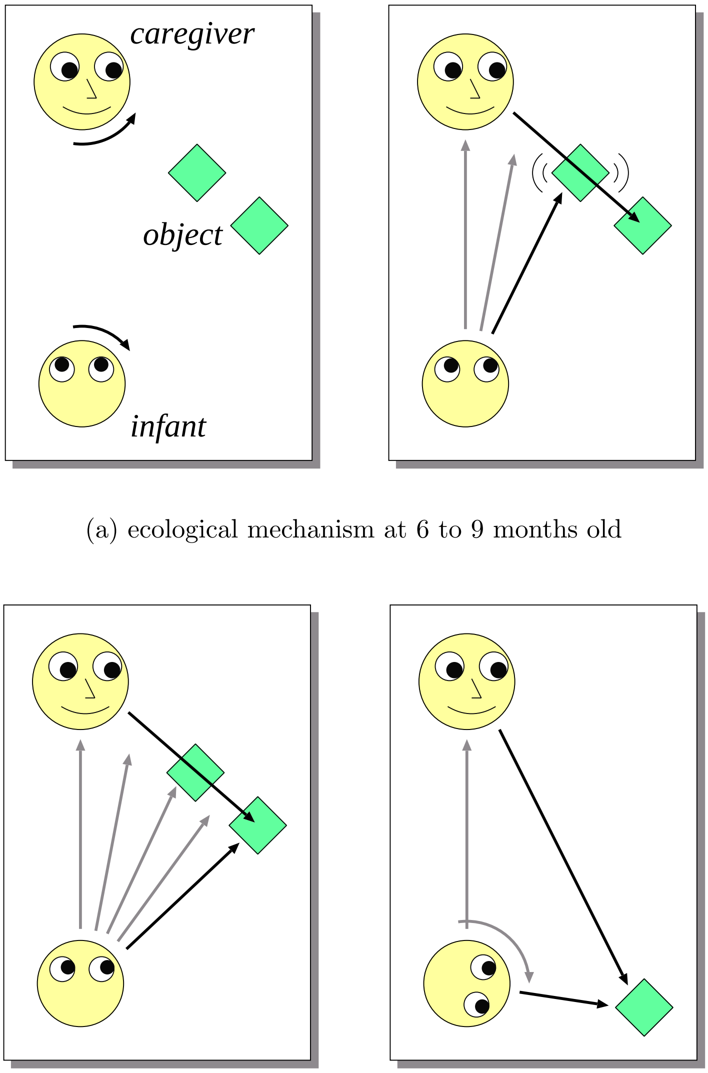
Butterworth and Jarrett [1991] examined how human infants recognized where someone else was looking. In the experiments, an infant and his/her mother were seated face-to-face in a structured environment where several objects, e.g. toys that attract the infant's interest, were placed around them. After interacting with the infant as usual, the mother was asked to look at one toy without pointing it. Butterworth and Jarrett recorded the interactions between the infant and the mother on VTR and analyzed what cues the infant utilized to determine an object to be gazed at and how correctly the infant responded to the shift of the mother's gaze. From the results of the experiments with a number of infants, Butterworth and Jarrett found that infants from 6 to 18 months old develop the ability of joint attention through three stages: ecological, geometric, and representational stages as shown in Figure 2.2. The mechanisms of infants in these stages were explained as follows:

Ecological mechanism at 6 to 9 months old:

An infant at 6 months old is able to look the same side as his/her mother's gaze sifts, that is, the infant can distinguish between that the mother is looking left and that the mother is looking right (see Figure 2.2 (a) left). At the same time, the infant has a tendency to look at an object which has interesting features or is moving in the field of the infant's view. An infant at 9 months old is able to track his/her mother's gaze direction until a salient object is first encountered in the infant's view (see Figure 2.2 (a) right). If the object that the mother is looking at is further along the mother's gaze direction, the infant stops to follow the mother's gaze when he/she first detects an salient object in the field of his/her view. The ability of the infant in this stage is called "ecological mechanism" of joint attention since it is believed that the structure of a natural environment allows the infant to show these behaviors.

Geometric mechanism at 12 months old:

An infant at 12 months old becomes to look at an object that his/her mother



(a) ecological mechanism at 6 to 9 months old

(b) geometric mechanism at 12 months old

(c) representational mechanism at 18 months old

Figure 2.2: The three stages of the development of infants' joint attention. (a) An infant at 6 to 9 months old is able to distinguish that his/her mother is looking left or right. However, the infant at this stage has a tendency to look at a salient object in the field of the infant's view. (b) An infant at 12 months old is able to look at the same object that his/her mother is looking at only when the object is observed in the field of the infant's first view. (c) An infant at 18 months old can realize joint attention even if the object that his/her mother is looking at is behind the infant.

is looking at (see Figure 2.2 (b)). Even if another object is encountered in the field of the infant's view while he/she tracks the mother's gaze, the infant is able to pass the object and looks at the correct one. The infant's ability of joint attention in this stage is called "geometric mechanism" since the infant can determine the direction of the mother's attention. However, the infant in this stage exhibits gaze following only when the object that the mother is looking at is observed in the field of the infant's view when he/she is looking at the mother. The infant at this stage does not turn to look behind himself/herself even if the mother is looking there.

Representational mechanism at 18 months old:

An infant at 18 months old is able to follow his/her mother's gaze and turn his/her head to take a look at the object that the mother is looking at. The infant in this stage realizes joint attention regardless of whether the object that the mother is looking at is inside or outside of the field of the infant's view when he/she is looking at the mother (see Figure 2.2 (c)). This ability of the infant's joint attention is called "representational mechanism" since the infant seems to have a representation of which space is not observed in the infant's view.

As described here, it was found that human infants acquire the ability of joint attention through three developmental stages. However, a question still remains. What internal and external mechanisms of infants enable them to develop the ability? Observational and analytical studies have explained the developmental *phenomena* of infants' joint attention but not revealed the developmental *mechanisms* of that. In Chapter 5, we propose a constructivist model that could be one of the models to explain how infants acquire the ability of joint attention through the staged developmental process by making a robot reproduce a similar developmental process as that of infants.

2.1.3 Innate Preferences

Human infants exhibit behaviors which show that the infants have several preferences and capabilities shortly after birth [Bremner, 1994]. Such preferences and capabilities of infants seem to be inherently prepared or to be acquired in their mothers' wombs. Infants, in any case, are able to have various experiences based on their innate preferences and capabilities; furthermore, the experiences enable the infants to acquire much more preferences and capabilities.

Innate preferences and capabilities of infants have been examined by a preferential looking method and a habituation-dishabituation method [Bremner, 1994]. Preferential looking is a phenomenon that infants have a tendency to look longer at a favorite visual stimulus than others. Infants are shown two visual stimuli in an appropriate position where the stimuli can be recognized. Under this situation, if the infants distinguish two stimuli and have preferences for one rather than the other, the infants are expected to look longer at the favorite one than the other. Observers examine the infants' preferences by measuring the time in which the infants have looked at each visual stimulus. On the other hand, habituation means that infants gradually do not look at a familiar visual stimulus by getting weary as time goes on. In contrast, dishabituation means that infants have interest in a novel stimulus and come to look at the novel one for long again. Observers examines the infants' capabilities to perceive and discriminate visual stimulus depending on whether the infants exhibit dishabituation phenomena or not when a novel visual stimulus is presented after a familiar one has been presented.

The following paragraphs explain the findings about visual preferences of infants revealed in cognitive developmental science. Infants are known to prefer to look at salient visual stimuli, especially face-like stimuli the most.

Preferences for Salient Visual Stimuli

Human infants have preferences for salient visual stimuli, such as bright colors, rich patterns, and motion. Bright colors are easier to be detected than dark ones; therefore, such colors attract infants' interests. Shankle and his research members [Shankle, 2003] have collected the data of the preferences of more than one thousand infants and examined what visual stimuli infants prefer and since when infants start to prefer. They found that infants at 0 month old prefer to look at yellow, orange, red, green, and turquoise, and infants at 3 months old prefer yellow and red rather than blue and green. Preferences for rich patterns are also one of the characteristics of infants. Infants cannot receive as fine images as adults because of the immaturity of their visual accommodation (refer to Section 2.1.5 for more detail). However, infants are known to prefer looking at the richest patterns that they can detect. Banks *et al.* [Banks and Ginsburg, 1985; Banks and Dannemiller, 1987] have investigated the development of the infants' sensitivities to visual stimuli and their preferences for complexity. They suggested that the most interesting visual stimuli for infants are those with maximum complexity that the infants can detect. In addition, it is also known that infants prefer to look at moving objects. When infants are shown moving toys or stationary ones, they exhibit stronger interests in moving ones than stationary ones. It is considered that the sensitivity to the motion is beneficial not only for human infants but also for all animals. The reason is that animals have to always protect themselves from enemies. For the same reason, human infants have also preferences to look at motion.

In Chapter 5, we propose a constructivist model by which a robot learns joint attention through experiences based on the preferences for salient visual stimuli such as bright colors, rich patterns, and motion.

Preferences for Human Faces

Infants are well known to prefer looking at human faces or face-like stimuli. It is observed that infants intently look at the mother's face shortly after birth. Fantz [1961; 1963] investigated what visual stimuli infants prefer to look at by a preferential looking method. Infants from 4 days to 6 months old were presented some patterns shown in Figure 2.3, which includes a pattern diagram of a human face, printed out letters, concentric circles, and colored boards (red, white, and yellow). Fantz measured how long infants looked at each stimulus. The results of the percent of the total fixation time are shown as a bar chart in the figure. The upper bar of each stimulus shows the result of infants at 2 to 3 months old, and the lower one shows that of infants at 3 months old or older. From the result, it was found that infants prefer to look at rich patterns than simple ones; furthermore, they prefer to look at a human face the most. Such a preference of infants for a human face have been confirmed in other studies [Morton and Johnson, 1991; Mondloch *et al.*, 1999; Cassia *et al.*, 2001].

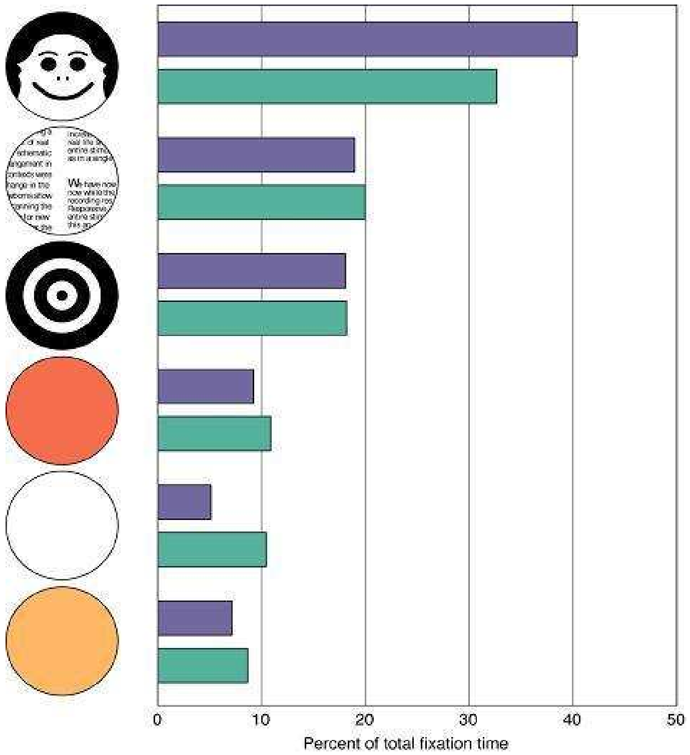


Figure 2.3: Infants' preferences for human faces [Fantz, 1961]. Infants were presented some visual stimuli and examined how long they looked at each stimulus by a preferential looking method. The upper bar of each stimulus shows the percent of total fixation time of infants at 2 to 3 months old, and the lower one shows that of infants at 3 months old or older. This result indicates that infants prefer to look longer at rich patterns than simple ones; furthermore, they prefer human faces the most.

Recent studies in electrophysiology, neuropsychology, and computationally have suggested that infants' preferences for human faces are attributable to both intrinsic and extrinsic factors [Morton and Johnson, 1991; Simion *et al.*, 2001; de Haan *et al.*, 2002]. The interactions of these factors form cortical specialization for the processing of faces. This evidence suggests that infants inherently have a non-specific mechanism to perceive human faces, and then from a specific one through experiences based on the non-specific one.

The both findings described in this section, the preferences for salient visual stimuli and that for human faces, are considered to be related to the development of joint attention of infants. The preferences for salient stimuli enable infants to have experiences to look at various objects, and that for human faces allow them to interact with their caregivers. Through such experiences, infants form triadic interactions among themselves, caregivers, and objects, and consequently acquire the ability of joint attention by finding some sort of relationship among the three. The constructivist models proposed in Chapters 4 and 5 embed the mechanisms of these preferences into a robot and make it interact with its environment.

2.1.4 Contingency Learning

Infants are able to find a contingency and causality through interactions with their environments [Leslie and Keeble, 1987; Hains and Muir, 1996; Nadel *et al.*, 1999] and to learn the meanings of those [Dunham and Dunham, 1995; Dickinson, 2001]. A contingency means a relationship between an action and a change of an environment. If an action of an infant shifts his/her environment from a certain state to another one at high probability, the action and the change of the environment have a contingency between them. Infants detect a contingency when their environment shifts from a certain state to another one by their own action, and learn the relationship between the action and the environmental change. Such ability enables infants to iterate actions by which the infants obtain a reward or positive feedback from their environments. At the same time, the ability also allows the infants to prevent actions by which they receive a punishment or negative feedback. The ability to find and learn a contingency and causality is crucial for all animals to adapt to and survive in environments. In the standpoint of robotics, the scheme of contingency learning is consistent with that of reinforcement learning [Sutton and Barto, 1998]. Hence, it is considered that contingency learning is also suitable for robot learning.

In the learning of joint attention, infants seem to utilize the mechanism of contingency learning. For example, it is assumed that an infant shifts his/her gaze direction to a certain space during interacting with his/her caregiver. If the infant consequently finds an interesting object that the caregiver is looking at and receives positive feedback from the caregiver, the infant can detect a contingency between his/her action and the change of his/her perceptions. In contrast, even if an infant does not receive any feedback from the caregiver, the infant can find some kind of contingency because he/she can evaluate their own action by himself/herself using self-feedback. Such mechanisms based on contingency learning serve as bases for the proposed constructivist models for joint attention. The learning model based on caregiver's feedback and that based on self-feedback are presented in Chapters 4 and 5, respectively.

2.1.5 Development of Visual Accommodation

Infants develop their visual accommodation [Banks, 1980; Bremner, 1994; Currie and Manny, 1997]. Although infants receive blurred images shortly after birth because of the immaturity of their visual accommodation, they gradually become to receive fine images by improving their accommodation. Visual acuity of infants are examined by using a preferential looking method and a habituation-dishabituation method or by measuring their brain activities. For example, the changes of infants' behaviors when they are gazing at a target which gradually changes the fineness of its grid pattern show the limitation of the infants' visual resolution. A grid pattern is observed as a grid one at high resolution while it is observed as a gray color at low resolution. The qualitative change of the target causes the change of the time in which the infants gaze at the target. As described in Section 2.1.3, it is known that infants do not have a preference for dark colors such as gray, but have a strong preference for a rich pattern which the infants can detect. Such visual preferences of infants allow to measure their visual acuity. Through the longitudinal experiments of infants' visual acuity, it has been found that infants develop their visual accommodation in the first few months or in the first year of their lives.

Triggers for the development of visual accommodation have several possibilities. Time could cause the development of visual accommodation, and a computational ability to process input images could also cause that. Our study supports the latter theory, that is, infant's visual acuity improves according to the computational ability in his/her brain, which improves through interactions with environment. The theory that the interactions with environments cause the development of the infant's visual acuity is based on one of the main principles of embodiment.

The development of visual accommodation is related to the next knowledge, that is, development helps learning. Our study suggests that the development of infants' visual accommodation could facilitate the infants themselves learning joint attention. A constructivist model proposed in Chapter 4 examines how the visual development of a robot facilitates the learning of joint attention.

2.1.6 Development Helps Learning

Infants develop physically and functionally. The developments of infants are considered to help themselves to learn various capabilities [Newport, 1990; Elman, 1993]. A difference between *development* and *learning* is directionality of the change. Development means that things change in a positive direction, e.g. more differentiation, more organization, and usually ensuring better outcomes [Elman *et al.*, 1996]. In contrast, learning means that things change in a certain direction that is determined through interactions with environments (refer to Section 4.1 for more detailed definitions).

In the learning of infants through interactions with caregivers, functional changes in a positive direction could happen in the caregivers as well as the infants. Infants develop their functional elements as they grow; at the same time, caregivers develop how to interact with infants. It is believed that the caregivers facilitate the infants' learning by changing how to interact according to the infants' performance. Newport [1990] suggested that both of the developments of infants and caregivers help the infants to learn their first languages. Infants improve their perceptual abilities and memory as they grow; at the same time, their caregivers adapt their responses to the infants according to the infants' performance. Both of these developments become filters to reduce difficulties to acquire languages. As a result, it allows infants to learn languages more efficiently. Newport referred to this idea as "less is more hypothesis." Elman [1993] also empirically showed that development helps learning. He designed an artificial neural network model, in which inputs and memory capacity develop, and made the network learn the structures of language in complex sentences. From the experimental results, he showed that the neural network could learn the structures of language only when the network included the developmental factors. This effect is labeled as “importance of starting small.”

In Chapter 4, we proposes a constructivist model by which a robot learns joint attention in parallel with the developments of the robot and a caregiver. A robot develops its visual function as learning advances; at the same time, a caregiver adapts how to evaluate the robot according to the robot's performance. Some experiments verify the knowledge that a development helps learning.

2.2 Robotics Approaches to Cognitive Developmental Science

A number of findings about human infants in cognitive developmental science have motivated robotics researchers to build human-like intelligent systems [Brooks *et al.*, 1998; Asada *et al.*, 2001]. This section reviews previous robotics approaches to cognitive developmental science. First, several projects to build social and/or developmental robots are described. Then, other studies related to our work are cited.

2.2.1 Social and/or Developmental Robots

Many studies on social and/or developmental robots have been conducted in recent years. Such studies ware surveyed in [Fong *et al.*, 2003; Lungarella and Metta, 2003]. Parts of the studies are based on knowledge about infants from cognitive developmental science. Among those, this section reviews the research projects called *Kismet*, *Cog*, *Infanoid*, *Robovie*, *Babybot*, and *MESA projects*. All of these projects are inspired by the findings in cognitive developmental science. The goal of each project and that of our study are summarized in Table 2.1. The goals are classified broadly into the following three categories:

(a) to build a robot socially communicating with humans,

(b) to construct an artificial model that enables a robot to seem to develop like human infants, or

(c) to understand the development of human infants by constructing a developmental model for a robot.

Table 2.1: The review of the research projects on social and/or developmental robots. The goal of our study is described in contrast with that of previous ones.

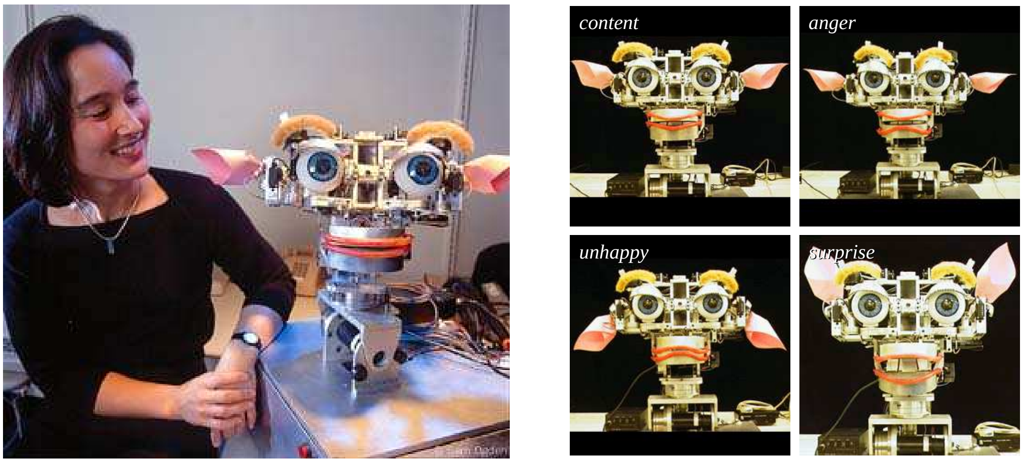
Robot's name or Project's name	Goal of the study
Kismet	To build various skills that enable the robot to enter into natural and intuitive social interactions with human caregivers and to learn from them like infants. ⇒ (a)
Cog	To investigate how to build intelligent robotic systems by following a developmental progression of skills similar to that observed in human developments and to design an artificial model of a theory of mind. ⇒ (b)
Infanoid	To investigate the underlying mechanisms of social intelligence that enable the robot to communicate with humans and to participate in human social activities. ⇒ (b)
Robovie	To develop a robot that communicates with humans and participates in a human society as a partner. ⇒ (a)
Babybot	To uncover the mechanisms of the functioning of the brain by building physical models of the neural control and cognitive structures. ⇒ (c)
MESA project	To understand the mechanisms of the emergence of shared attention in infants through observational, modeling, and robotic studies. ⇒ (c)
Our study	To understand the developmental mechanism of infants' joint attention through constructing artificial models by which a robot develop the ability of joint attention based on knowledge from cognitive developmental science. ⇒ (c)

The aim of (a) is to build a social communication robot, and it is not taken into account whether the mechanisms embedded into the robot copy after the mechanisms of humans. The good appearances of communication between a robot and humans are required in the study. Among the research projects reviewed in this section, the projects of *Kismet* and *Robovie* are considered to belong to this group. The aim of (b) is to construct an artificial developmental model for a robot. The process of the development of the robot seems to equal to that of human infants. However, all modules of the cognitive functions built in the robot are fully-programmed by a designer. It means that it is not taken into account whether the robot can develop the cognitive functions by itself or not. The research projects of *Cog* and *Infanoid* have this kind of the goal. The aim of (c) is to understand the developmental mechanisms of infants from constructivist approaches. Through constructing a robot that develops and learns like infants, it is expected to reveal the developmental mechanisms of infants. The knowledge from cognitive developmental science become significantly essential in this kind of study. The rest of the projects, *Babybot*, *MESA*, and *our study*, fall into this group. The following paragraphs explain the features of these projects and describe the aims of the projects in contrast with our study.

Kismet

A social interactive robot, called *Kismet*, has been developed at the Artificial Intelligence Laboratory at Massachusetts Institute of Technology [MIT-AI-Lab, 2003b]. Figure 2.4 shows Kismet, which has an anthropomorphic face with a large set of expressive features: eyelids, eyebrows, ears, jaw, lips, neck, and eye orientation. The purpose of this project is to build various skills that enable Kismet to enter into natural and intuitive social interactions with human caregivers and to learn from them like infants [Breazeal, 2000].

Breazeal and her colleagues have developed a system by which Kismet emotionally interact with humans. They build a mechanism to provide emotional feedback to humans through the facial expressions based on perceptions, attention, drives, and emotions [Breazeal and Scassellati, 2000]. The attention of Kismet is determined by negotiating the robot's physical constraints, the perceptual needs of the robot's behavioral and motivational systems, and the social implications of the motor acts



(a) Kismet and the developer, Breazeal, who has investigated infant-like social interactions between the robot and a human caregiver.

(b) Examples of Kismet's facial expressions: content, anger, unhappy, and surprise.

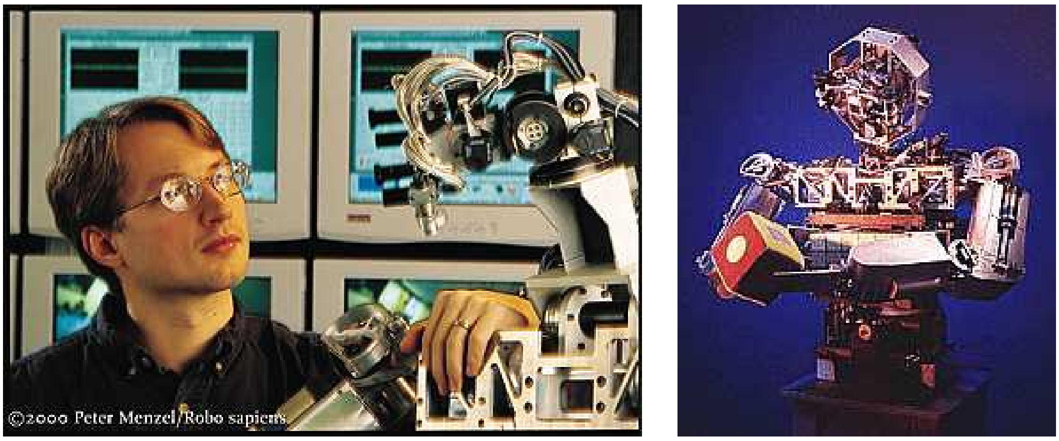
Figure 2.4: Kismet, an expressive robot which has been developed at the Artificial Intelligence Lab at MIT [MIT-AI-Lab, 2003b]. Kismet has a large set of expressive features: eyelids, eyebrows, ears, jaw, lips, neck and eye orientation, and emotionally communicates with humans through facial expressions and speech.

[Breazeal and Scassellati, 1999; Breazeal *et al.*, 2001]. Kismet can recognize four distinct prosodic patterns of speech directed from humans and shift its emotions according to the speech patterns [Breazeal and Aryananda, 2002]. The emotion of Kismet is delivered to humans not only through the facial expressions but also through the affective speech, in which the emotion is enhanced by synchronizing the motion of lips [Breazeal, 2003b]. Their psychological experiments with Kismet have suggested that all of these mechanisms: facial expression, attention, and speech with emotions, should have regulation and entrainment in human-robot interactions [Breazeal, 2002; 2003a; 2003c]. Furthermore, they have discussed imitations and social interactions in robots, and suggested that robots can find when to imitate, what to imitate, and how to imitate from some cues of humans [Breazeal and Scassellati, 2001; 2002]. This project has been heavily inspired by infants' developments, psychology, ethology, and so on. However, the system embedded in Kismet does not model infants'.

Cog

An upper-torso humanoid robot, called *Cog*, has been developed at the Artificial Intelligence Laboratory at Massachusetts Institute of Technology [MIT-AI-Lab, 2003a]. Figure 2.5 shows the appearance of Cog, which has twenty-one degrees of freedom and a variety of sensors, e.g. vision, auditory, and tactile, to approximate a human motions and senses. The aim of this project is to investigate how to build intelligent robotic systems by following a developmental progression of skills similar to that observed in human developments [Brooks *et al.*, 1999; Adams *et al.*, 2000; Scassellati, 2001c].

Scassellati, who is one of the developers of Cog, has investigated social interactions between the robot and humans and proposed a model of a theory of mind for the robot [Scassellati, 2000; 2001b; 2002]. His model is based on two theories: Leslie's model [Leslie, 1994] and Baron-Cohen's model [Baron-Cohen, 1995]. The former treats causality as a central principle to represent theories of objects' mechanics and others' minds, and the latter equips an intentionality detector, an



(a) Cog and the developer, Scassellati, who has proposed an artificial model of a theory of mind for the robot.

(b) Cog has twenty-one degrees of freedom and various sensory systems.

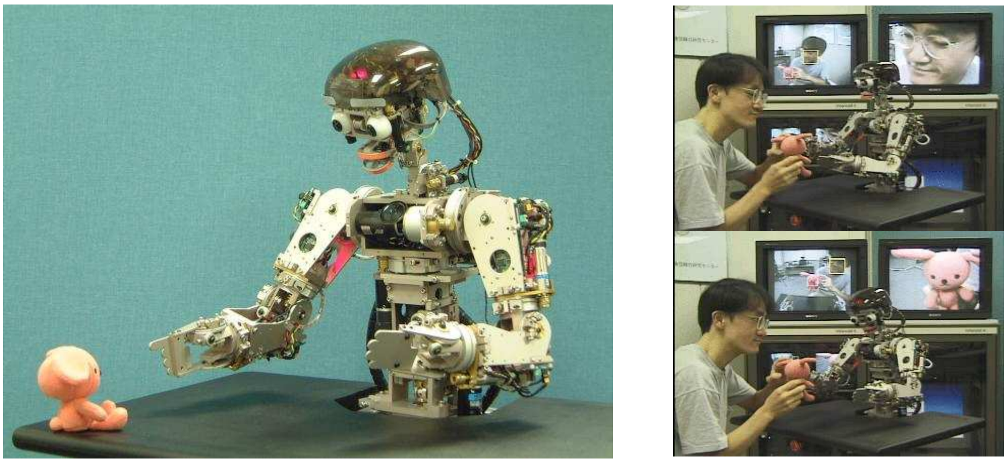
Figure 2.5: Cog, an upper-torso humanoid robot which has been developed at the Artificial Intelligence Lab at MIT [MIT-AI-Lab, 2003a]. Cog has some basic skills such as attention, face detection and gaze following which serve the basis for a theory of mind.

eye direction detector, and a shared attention mechanism as precursors to a theory of mind mechanism. To put it plainly, the theory of mind model proposed by Scassellati first distinguishes animate stimuli from inanimate ones based on Leslie's model and then makes the robot socially interact with the animate beings through shared attention based on Baron-Cohen's model. For realization of the model, Scassellati has implemented some basic skills into Cog, e.g. the same attention system as Kismet [Breazeal and Scassellati, 1999], a mechanism to discriminate animate from inanimate stimuli by spatio-temporal properties [Scassellati, 2001a], a face and eye finding mechanism with ratio template algorithm [Scassellati, 1998], and so on. These are key components in his theory of mind model. Furthermore, he has proposed idea to develop a gaze following mechanism [Scassellati, 1996; 1999], which serves as a basis for joint attention. Each of these mechanisms has been constructed inspired by human beings' mechanisms that are suggested in cognitive developmental science. However, all of the mechanisms have been implemented into the robot as accomplished ones by the designer; therefore, this project cannot explain how infants acquire the social abilities.

Infanoid

An infant-like robot, named *Infanoid*, has been developed by Kozima at Communications Research Laboratory in Japan [Kozima, 2003]. Infanoid is shown in Figure 2.6, in which the robot achieves joint attention with a human caregiver. The robot has an upper torso that is approximately the same kinematic structure and size as a human infant at three years old, and equips vision and auditory sensors. The objective of this project is to investigate the underlying mechanisms of social intelligence that enable the robot to communicate with humans and to participate in human social activities [Kozima, 2000; 2002].

Kozima has discussed social intelligence of robots from a viewpoint of epigenetic robotics [Kozima and Zlatev, 2000]. Epigenesis of communication means that a robot which has a body and minimum innate abilities explores how to socially communicate with humans through simple interactions based on the innate abilities. Kozima has advocated that an attention mechanism and an imitation one enable a robot to acquire social intelligence through indirect experiences of what others experience by sharing



(a) Infanoid, an upper torso humanoid robot that is approximately the same kinematic structure and size of an infant at three years old.

(b) Infanoid realizes joint attention with a human caregiver, Kozima, based on the preprogrammed mechanism.

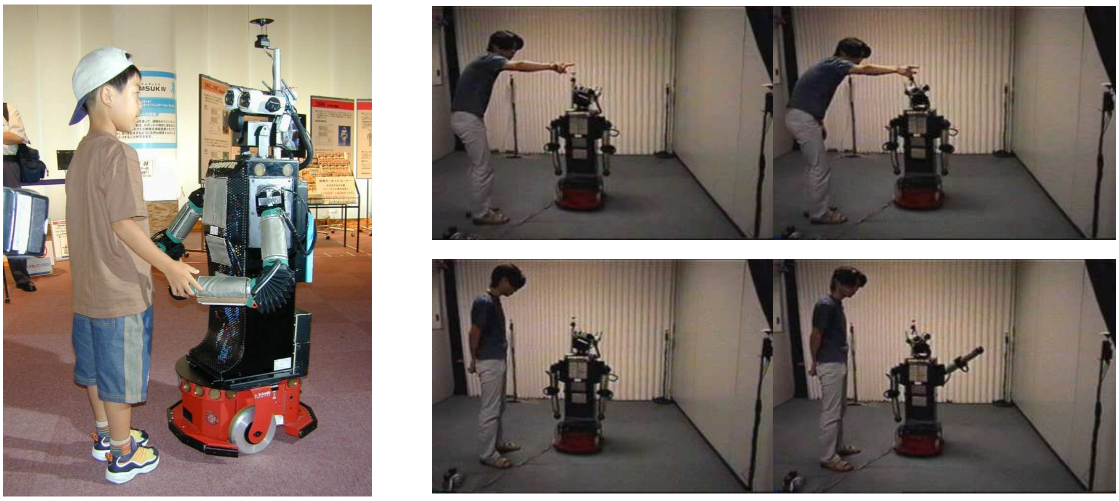
Figure 2.6: Infanoid, an infant-like robot which has been developed at Communications Research Laboratory in Japan [Kozima, 2003]. Infanoid has the ability of joint attention fully-programmed by a designer.

an attention target and an action to the target [Kozima, 1998]. For the purpose of this, Kozima has equipped Infanoid with basic social abilities of infants from 6 to 9 months old as an initial stage for social and communicative development [Kozima, 2002]. For example, Infanoid has the abilities to track a human face and salient objects, to alternately look at the face and the object, to point to and reach out for the face or the object, and to vocalize babbling with lip-synching. In addition, the ability to roughly determine the direction of the human's attention, which leads to gaze following and joint attention, has also been implemented as a basic skill [Kozima, 1998; Kozima and Yano, 2001]. Based on these mechanisms, Infanoid is able to realize primary joint attention with a human caregiver as shown in Figure 2.6 (b). However, it has been supposed in this project that the ability of joint attention is innate or already acquired, and it has not been discussed how the robot or an infant acquire the ability of joint attention through interactions with environments.

Robovie

An interaction-oriented robot, named *Robovie*, has been developed at Intelligent Robotics and Communication Laboratories at ATR (Advanced Telecommunications Research Institute International) [ATR-IRC, 2003]. Robovie has a human-like appearance shown in Figure 2.7 and has a mobile platform, two arms, and a head with various sensors, e.g. vision, sense of touch, audition, and so on. The aim of this project is to develop a robot that communicates with humans and participates in a human society as a partner [Kanda *et al.*, 2002b].

Robovie has the capabilities to generate human-like behaviors to communicate with humans by using the human-like actuators and sensors [Ishiguro *et al.*, 2001; Kanda *et al.*, 2002a; Miyashita and Ishiguro, 2003]. The behaviors are determined based on over 100 behavior modules and 800 episode rules, which connect the behaviors. All of these were fully-programmed by designers based on knowledge obtained



(a) Robovie, an human-like mobile robot that is 120 [cm] tall and has several actuators and sensors.

(b) Robovie realizes joint attention by following the pointing of a human (top). Robovie makes a human perform joint attention by generating eye-contact, gaze shift, and pointing (bottom).

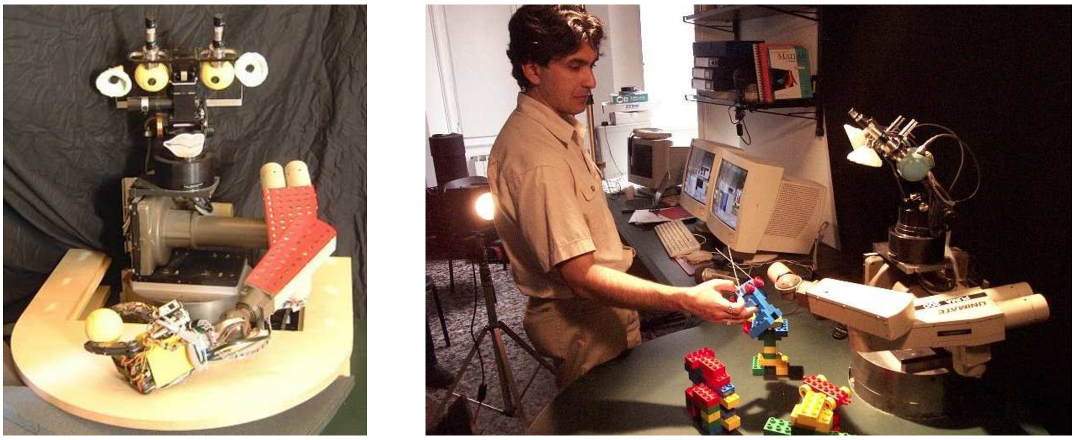
Figure 2.7: Robovie, an interaction-oriented robot which has been developed at Intelligent Robotics and Communication Laboratories at ATR [ATR-IRC, 2003]. Robovie has a number of behavior modules and episode rules to communicate with humans.

through cognitive experiments. Furthermore, Robovie is able to entrain humans into communications by generating such behaviors. For example, Robovie makes humans perform joint attention by showing attention expressions such as gaze shift, pointing, and eye-contact as shown in Figure 2.7 (b) [Imai *et al.*, 2001]. An utterance model that takes advantage of the theory of mind of humans enables Robovie to gain the cooperation of humans [Ono *et al.*, 2000]. The interactions between Robovie and humans have been evaluated by using a psychological method and/or a motion capturing system, which allow the researchers to measure the movements of Robovie's and humans' bodies in detail [Kanda *et al.*, 2003]. This project aims at realizing human-robot communication but not at understanding how the robot acquires such communication abilities.

Babybot

An artificial newborn, called *Babybot*, has been developed at Laboratorio Integrato di Robotica Avanzata (LIRA) Laboratory at the University of Genova [LIRA-Lab, 2003]. Babybot, which is shown in Figure 2.8, has eighteen degrees of freedom distributed along the head, arm, torso, and hand, and various sensors, e.g. a pair of cameras with space-variant resolution, two microphones, tactile sensors, and so on. The goal of this project is to uncover the mechanisms of the functioning of the brain by building physical models of the neural control and cognitive structures.

Sandini and Metta *et al.* [Sandini *et al.*, 1997; Metta, 2000; Metta *et al.*, 2000; 2001] proposed an approach aimed at the design and the comprehension of complex adaptive systems like humans. The idea of their approach is analogous to the principle of cognitive developmental robotics. For the purpose of designing and uncovering complex adaptive systems, Capurro *et al.* [1997] first developed a camera system that can acquire log-polar images like humans and implemented several behaviors of eye movements such as vergence and saccade into Babybot. Based on these primary behaviors as what newborns have, they have investigated how to integrate different sensory modalities. Metta *et al.* [1999] proposed a developmental model of visually-guided reaching. Natale *et al.* [2002] developed a functional model of the acquisition of visual, acoustic, and multi-modal motor responses. These models have enabled Babybot to develop its sensorimotor coordination like newborns. However, it has not



(a) Babybot, an artificial newborn that has eighteen degrees of freedom and sensors, e.g. vision, auditory, and so on.

(b) Babybot and the developer, Metta, who has proposed an approach aimed at the design and comprehension of complex adaptive systems and investigated a developmental learning model for a reaching behavior.

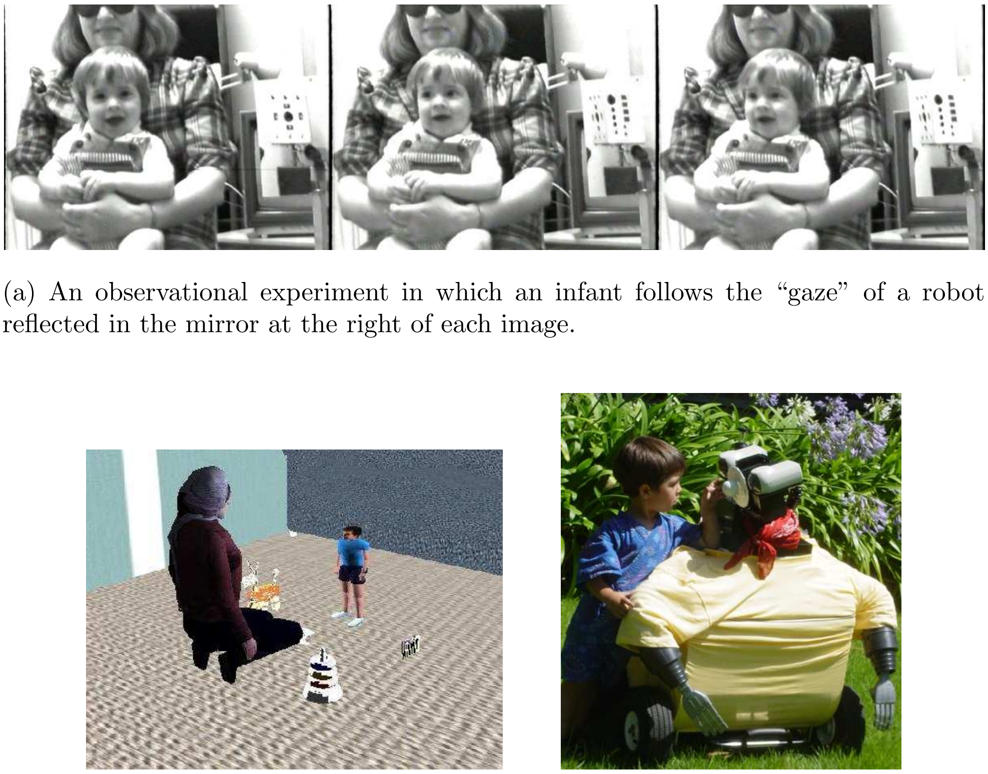
Figure 2.8: Babybot, an artificial newborn that has been developed at LIRA-Lab at the University of Genova [LIRA-Lab, 2003]. Babybot has several eye movements and learns how to integrate different sensory modalities.

been taken into account sociality yet. Social interaction is one of the essence of human intelligence and enables robots to acquire more advanced cognitive functions. Therefore, it should have respect to social interaction in studying cognitive development from a robotics viewpoint.

MESA project

Movellan, Triesch, Deák and their research members in University of California, San Diego are working on *MESA (Modeling the Emergence of Shared Attention) project* [UCSD, 2003]. This project combines three approaches: observational, modeling, and robotic approaches for understanding the mechanisms of the emergence of shared attention in infants. Examples of their experiments are shown in Figure 2.9.

The observational approach aims at collecting a database of everyday infant-caregiver interactions and at generating more refined theories of processes of social learning and development in infants based on the collected data [Deák *et al.*, 2000;



(a) An observational experiment in which an infant follows the "gaze" of a robot reflected in the mirror at the right of each image.

(b) A simulated environment which will be used in the modeling study.

(c) An interactive robot used in the robotic approach.

Figure 2.9: MESA project at University of California, San Diego [UCSD, 2003]. The project combines three approaches: (a) observational, (b) modeling, and (c) robotic approaches to understand how infants develop the ability of shared attention.

Movellan and Watson, 2002]. Infant-caregiver interactions, which include both undirected and scripted episodes, have been recorded on high-quality digital video and coded with high temporal and spatial precision. The data are utilized not only to construct a developmental theory of infants but also to control virtual agents in the modeling studies. The modeling approach intend to develop computational models by which shared attention emerges in infant-caregiver interactions [Deák *et al.*, 2001; Fasel *et al.*, 2002; Triesch *et al.*, 2003]. They have proposed a basic set of mechanisms that are sufficient for shared attention to emerge [Fasel *et al.*, 2002]. The basic set comprises perceptual and motivational biases, habituation mechanisms that drive infants to look at and shift their attention between interesting visual stimuli, a reinforcement learning mechanism, and a structured environment that provides a strong correlation between where a caregiver is looking and where interesting stimuli are. This basic set for the emergence of shared attention is similar to our idea presented in Chapter 5 except a structured environment. Based on the computational model, they have found that which parameter settings of the learning mechanism facilitate or inhibit the emergence of shared attention [Triesch *et al.*, 2003]. However, their model has been examined only in a computational simulation but not in a virtual or a real environment. The theories developed in the observational and the modeling studies then serve as the basis for human-robot interactions in the robotic approach. In the robotic approach, the interactions between a robot and a child who is normal or has autism or Down's syndrome have been examined. The robot has been designed to respond to the child with/without contingency. Through such experiments, they have attempted to determine whether a certain contingency can facilitate autistic children or Down's syndrome ones responding to social agents. The three approaches: observational, modeling, and robotic ones, will be highly integrated so that each approach become a help to modify others.

Research Map

Figure 2.10 shows a research map of the studies on social and/or developmental robots which were described here. The research projects, *Kismet*, *Cog*, *Infanoid*, *Robovie*, *Babybot*, *MESA project*, and our study are allocated in the space of "*social interaction based on embodiment* vs. *development and learning*." Social interaction, embodiment, development, and learning were discussed in Section 1.1 as essence of human intelligence, and it has been suggested that the essence should be considered in studying human intelligence from a viewpoint of robotics. Each axis in the map means the degree that each project takes into account "social interaction based on embodiment" and "development and learning." This map shows that our study adequately considers the essence of human intelligence compared to the others.

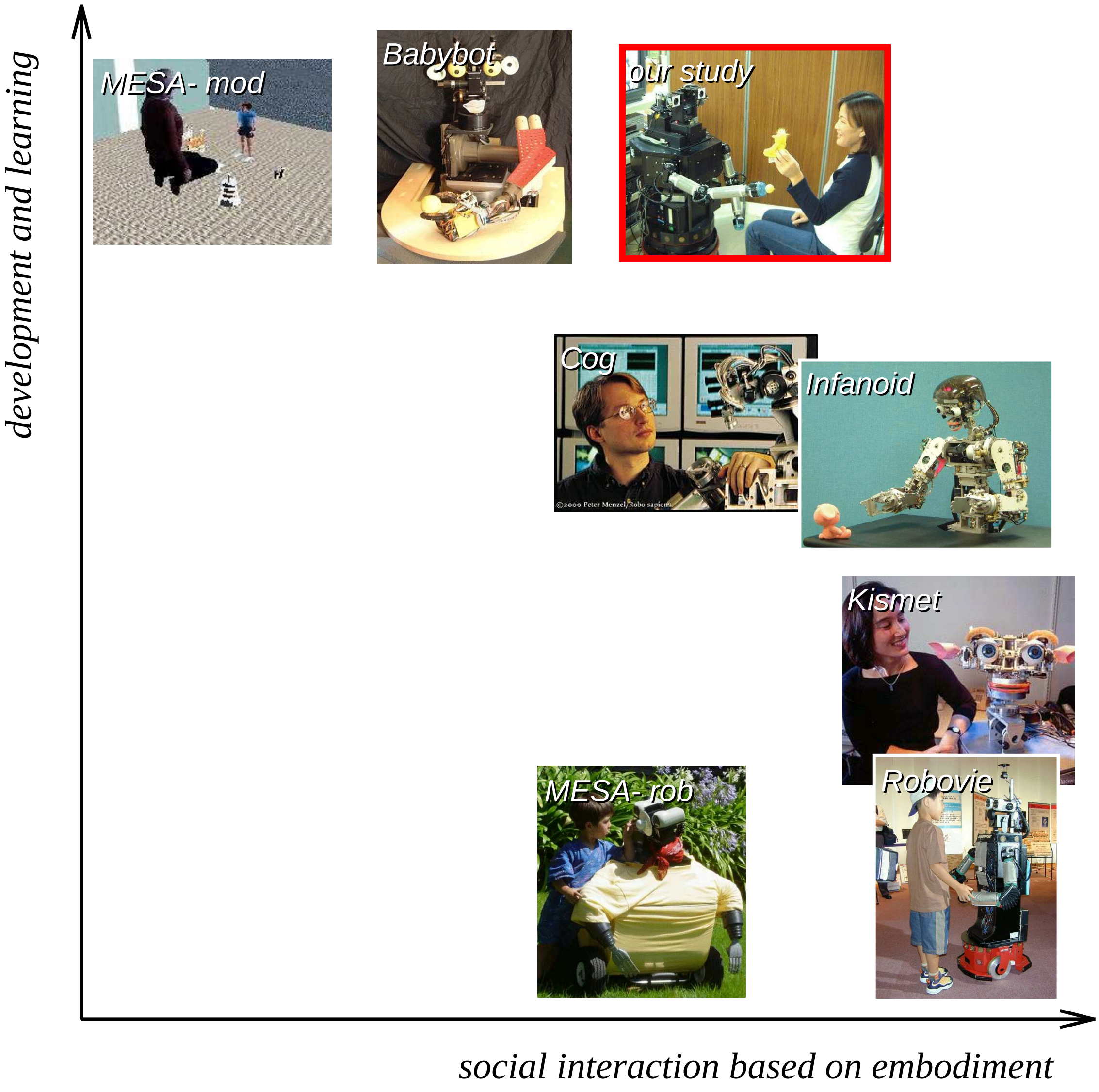


Figure 2.10: A research map of the studies on social and/or developmental robots. The research projects, *Kismet*, *Cog*, *Infanoid*, *Robovie*, *Babybot*, *MESA project*, and our study are allocated in the space of "*social interaction based on embodiment* vs. *development and learning*." This shows that our study adequately considers the essence of human intelligence compared to the others.

2.2.2 Other Approaches

Dominguez and Jacobs [Dominguez, 2003; Dominguez and Jacobs, 2003; Jacobs and Dominguez, 2003] have investigated the hypothesis that system learning with visual perception may benefit from the use of suitably designed developmental progressions during the training. To examine this hypothesis in the acquisition of binocular disparity sensitivities [Dominguez and Jacobs, 2003] and motion velocity sensitivities [Jacobs and Dominguez, 2003], they designed learning mechanisms in which visual inputs changed coarse-to-multiscale, fine-to-multiscale, or randomly in stages, or did not change. Their simulation results showed that the coarse-to-multiscale developmental model improved the performance best. Through these studies, they suggested that suitably designed visual development can aid visual learning.

Uchibe *et al.* [Uchibe *et al.*, 1998] proposed a method to control the complexity of an environment and state vectors of a robot for robot learning. For a soccer task of mobile robots, they designed a learning scheduling in which an opponent robot increased its motion speed in phase; at the same time, a learning robot increased the dimension of the state vector by taking a trade-off between the size of the state space and the learning time. From the experimental results, it was shown that their proposed method enabled the robot to acquire almost the same performance faster than the method in which maximum dimension of the state vector was used from the beginning of learning. They concluded that the acceleration owed the small dimension of the state vector in the early stage of learning and the initial value of the action value function acquired previously. Their work also suggested that the developments in an environment and a robot can facilitate the robot learning a soccer task.

2.3 Summary

This chapter has described knowledge about the cognitive developments of human infants and reviewed robotics approaches to cognitive developmental science. Studies in cognitive developmental science have made a number of findings about the developments of infants, especially the development of joint attention, Furthermore, these findings have inspired robotics researchers to build infant-like robots that can develop and learn through interactions with their environments. To investigate such robots has potentials to reveal the developmental mechanisms of human infants and to realize more adaptive and intelligent robots than fully-programmed ones. Our study has the same objective to understand the developmental mechanisms of infants' joint attention by constructing developmental models for a robot.

The next chapter provides the task definition of joint attention between a robot and a human caregiver and presents the basic idea of the proposed constructivist models for joint attention. The models are constructed based on the findings described in this chapter.

Chapter 3

Joint Attention between a Robot and a Human Caregiver

This chapter first provides the definition of joint attention between a robot and a human caregiver. A robot shifts its gaze direction according to the direction of a caregiver's gaze and tries to identify the object that the caregiver is looking at. The definition of joint attention stands on Butterworth's [Butterworth, 1991], in which joint attention is realized not based on a theory of *mind* but based on a theory of *body*. In our definition, a robot is able to realize joint attention without understanding a caregiver's intention. This chapter first defines the task of joint attention using an environmental setup which includes a robot, a human caregiver, and multiple objects.

Next, the concepts of two approaches for a robot to acquire the ability of joint attention are described. The approaches are based on knowledge about infants' developments. The findings in cognitive developmental science have suggested that caregiver's evaluation makes a significant difference in infant's learning. If an infant learns *with* caregiver's evaluation, the infant could be facilitated his/her learning owing to the evaluation. At the same time, the infant's learning could be made easier owing to the development of himself/herself. On the other hand, even if an infant learns *without* caregiver's evaluation, the infant could acquire new abilities based on his/her innate or pre-acquired capabilities. This chapter presents the basic idea of two learning models for robot's joint attention, each of which concerns the scheme of learning *with* caregiver's evaluation or that of learning *without* any external evaluation.

3.1 Task Definition of Joint Attention

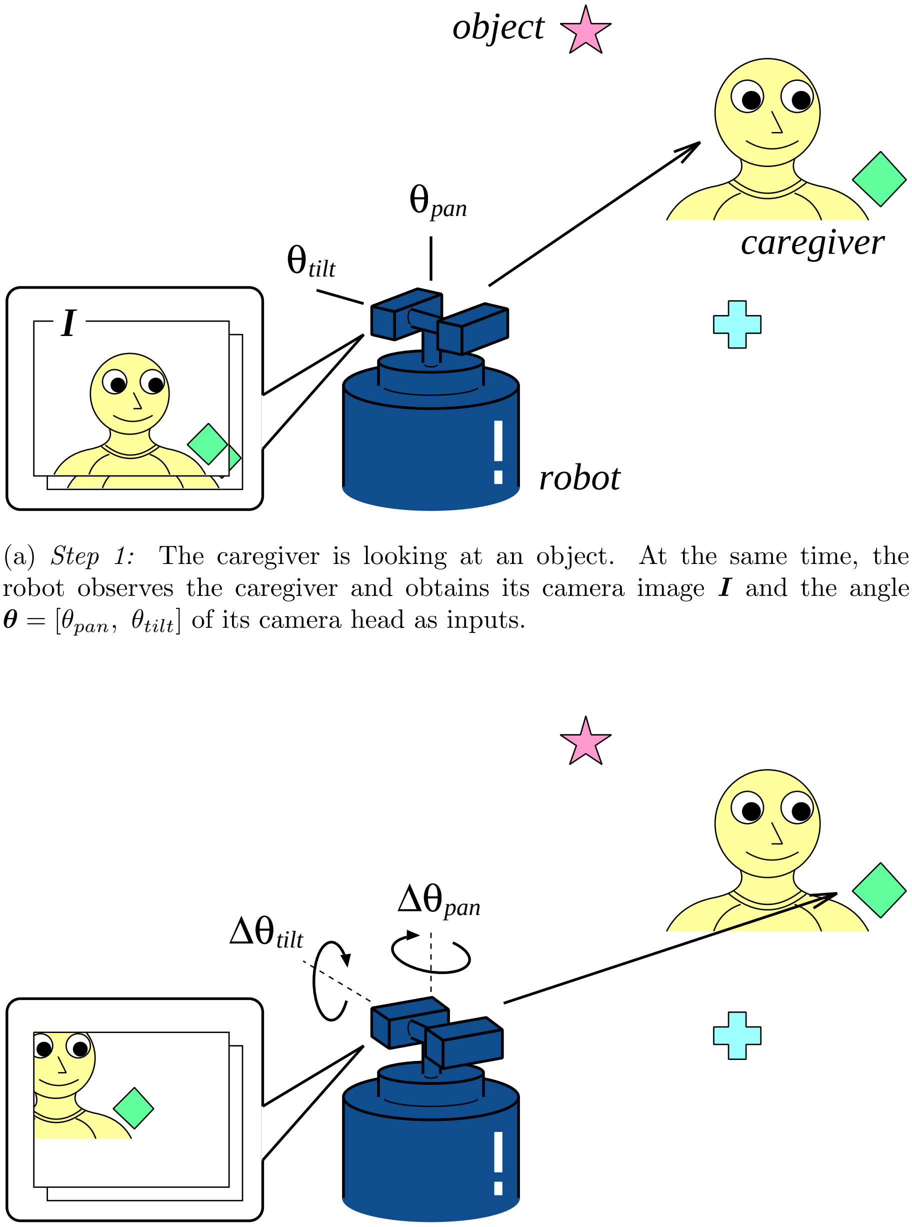
Joint attention between a robot and a human caregiver is defined as a process by which a robot looks at the same object that a caregiver is looking at. Figure 3.1 shows a two-step process through which a robot achieves joint attention with a caregiver in an environment including multiple objects. A caregiver and a robot with two cameras are seated face-to-face in an environment. The environment includes multiple salient objects of which positions and the degrees of saliency, e.g. the brightness of colors, the complexity of patterns, and the amount of motion, change randomly in every trial. This means that the environment is not structured for joint attention.

In each trial, the caregiver looks at one of the objects at random. In Figure 3.1, the caregiver is looking at the square object. At the same time, the robot has the capability to obtain its camera image $\boldsymbol{I}$ and the angle $\boldsymbol{\theta} = [\theta_{pan},\ \theta_{tilt}]$ of its camera head as sensor inputs, and to output a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ to rotate the camera head. The robot performs joint attention with the caregiver through the following two steps.

Step 1: The robot observes the caregiver who is looking at an object. The reason why the robot first attends to the caregiver is that social agents are known to have a preference for faces of other agents as mentioned in Section 2.1.3. Then, the robot obtains its camera image $\boldsymbol{I}$ and the angle $\boldsymbol{\theta}$ of its camera head as sensor inputs (see Figure 3.1 (a)).

Step 2: The robot outputs a motor command $\boldsymbol{\Delta\theta}$ for the camera head to rotate based on the sensor inputs $\boldsymbol{I}$ and $\boldsymbol{\theta}$ acquired in Step 1. As a result, if the robot has looked at the same object that the caregiver is looking at, joint attention succeeds (see Figure 3.1 (b)).

Note that a cue by which the robot shifts from Step 1 to Step 2 is explicitly provided by the caregiver since our definition of joint attention does not assume the intention



(a) *Step 1:* The caregiver is looking at an object. At the same time, the robot observes the caregiver and obtains its camera image $\boldsymbol{I}$ and the angle $\boldsymbol{\theta} = [\theta_{pan},\ \theta_{tilt}]$ of its camera head as inputs.

(b) *Step 2:* The robot outputs a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ to the camera head and shifts its gaze direction. If it follows that the robot has looked at the same object that the caregiver is looking at, joint attention succeeds.

Figure 3.1: A two-step process of joint attention between a robot and a human caregiver in an environment including multiple objects.

of a robot. For the realization of the above process of joint attention, the robot is required to have the abilities (1) to estimate the direction of the caregiver's gaze based on the sensor inputs and (2) to change the gaze direction of the robot's camera so that it tracks the estimated direction of the caregiver's gaze. Our study considers these problems as a direct mapping between the sensor inputs and the motor output. In other words, the proposed learning models do not utilize the intermediate representation of the direction of the caregiver's gaze. The robot has to learn/acquire the sensorimotor coordination between the inputs $\boldsymbol{I}$, $\boldsymbol{\theta}$ and the output $\boldsymbol{\Delta\theta}$ to realize joint attention.

3.2 Two Approaches for Learning of Joint Attention

To acquire the sensorimotor coordination for joint attention, several approaches are possible. This section presents the concepts of two learning models for joint attention based on knowledge about infants' developments.

The studies in cognitive developmental science have made a number of findings about infant's development and learning as described in Section 2.1. It is known that one of the significant factors in the process of infant's learning is caregiver's evaluation. The caregiver's evaluation makes a great difference in the infant's learning. If an infant learns *with* caregiver's evaluation, the infant could be facilitated his/her learning owing to the evaluation. In this case, it should be discussed how the caregiver evaluates the infant to facilitate his/her learning more. On the other hand, if an infant learns *without* any external evaluation, the infant might have a difficult learning time compared to the former case with caregiver's evaluation. However, it is conjectured that the infant has some potentials to acquire new abilities by himself/herself. In this case, it becomes an issue what capabilities the infant should have for acquiring new abilities. Focusing on these points, we propose the following two models for a robot to learn joint attention.

Developmental Learning Model with Caregiver's Evaluation

The first approach is *a developmental learning model with caregiver's evaluation*. This model is based on the knowledge that a caregiver can facilitate an infant learning by adjusting the criterion for the evaluation according to the performance of the infant. At the same time, it is known that an infant matures his/her internal mechanisms so that it makes his/her own learning easier. Based on the knowledge, the developmental learning model is structured as shown in Figure 3.2. This shows the concept of the model. A robot learns joint attention based on evaluation from a human caregiver. As learning advances, the robot matures its visual accommodation, and the caregiver adjusts the criterion for the task evaluation according to the performance of the robot. These changes are called a robot's development and a caregiver's development. The development learning model examines how the robot's and the caregiver's developments facilitate the learning of robot's joint attention. This model is discussed in Chapter 4.

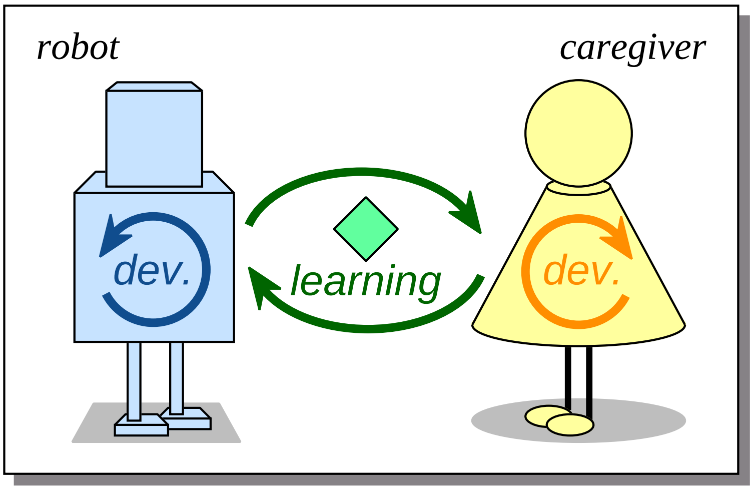


Figure 3.2: The concept of the developmental learning model. A robot learns joint attention based on caregiver's evaluation. As learning advances, both the robot and the caregiver develop their internal mechanisms. This model examines how these developments facilitate the learning of robot's joint attention.

Bootstrap Learning Model based on Robot's Embedded Mechanisms

The second approach is *a bootstrap learning model based on robot's embedded mechanisms*. This model is based on the knowledge that an infant inherently has various capabilities, e.g. preferences for salient visual stimuli and contingency learning, and such capabilities enable the infant to acquire new abilities. The scheme of learning based on only innate or pre-acquired abilities without any external evaluation is called bootstrap learning. The concept of the bootstrap learning model is shown in Figure 3.3. A robot has the embedded mechanisms of visual attention and learning with self-evaluation. The former is to look at a salient object in the robot's view, and the latter is to learn its sensorimotor coordination when visual attention has succeeded. Based on the mechanisms, the robot interacts with the environment and learns joint attention. Note that a caregiver just looks at one object and does not provide any evaluation to the robot. The bootstrap learning model examines how the robot acquires the ability of joint attention based on only its embedded mechanisms. This model is discussed in Chapter 5.

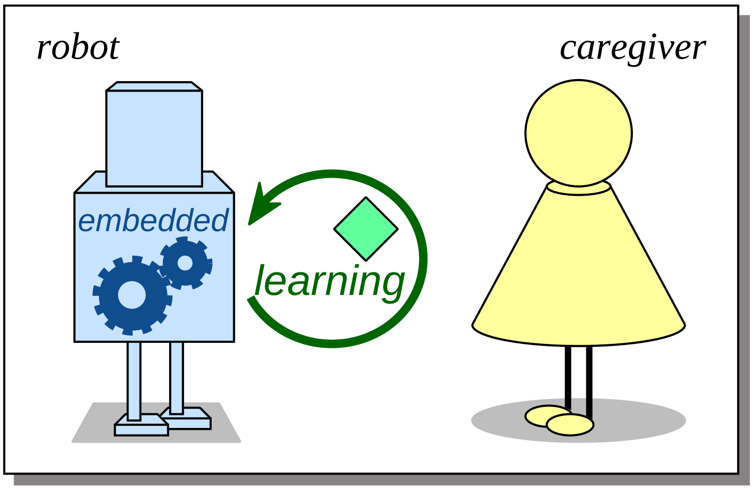


Figure 3.3: The concept of the bootstrap learning model. A robot learns joint attention based on only its embedded mechanisms: visual attention and learning with self-evaluation. The caregiver just looks at one object and does not provide any evaluation to the robot. This model examines how the robot acquires the ability of joint attention based on the embedded mechanisms.

3.3 Summary

This chapter has provided the task definition of joint attention between a robot and a human caregiver and described the concepts of the two learning models for joint attention. In the joint attention task, a robot is required to have the sensorimotor coordination to shift its gaze direction from a caregiver to the object that the caregiver is looking at. To acquire the sensorimotor coordination for joint attention, the concepts of two learning models have been presented. One is a developmental learning model with caregiver's evaluation, and the other is a bootstrap learning model based on robot's embedded mechanisms.

The following Chapters 4 and 5 give the detailed descriptions of the proposed models and empirically show how a robot acquires the ability of joint attention based on the models.

Chapter 4

Developmental Learning with Caregiver's Evaluation

This chapter describes a developmental learning model for joint attention with caregiver's evaluation. The model demonstrates that the developments of a robot's and a caregiver's internal mechanisms help the robot to learn joint attention through explicit interactions between the robot and the caregiver. This is based on the knowledge that a caregivers can facilitate an infant learning cognitive functions by evaluating him/her and adjusting the evaluation criterion according to the performance of the infant. At the same time, it is known that an infant matures his/her internal mechanisms so that it makes his/her own learning easier.

In this chapter, the definition of development is given first. Development is defined in contrast to learning. Next, this chapter introduces two kinds of developments: a robot's development and a caregiver's development, which play significant roles in the learning of joint attention. Then, a developmental learning model for joint attention, which consists of a learning mechanism based on caregiver's evaluation and the above two developmental mechanisms, is described. Some experiments show the validity of the proposed model, especially the effectivity of the robot's and the caregiver's developments. Finally, discussion and future work are given.

4.1 Development and Learning

Development is defined as “the act of growing by degrees into a more advanced or mature state” [Bartleby.com, 2003]. This is a phenomenon that emerges in every animal. For example, human infants develop their visual accommodation so that they can gradually acquire clear images and mature their motor mechanisms so that they can correctly control their actuators. It is known in advance which direction the development proceeds to. To clarify the definition of development, it is compared with learning, which has a somewhat similar meaning. *Learning* is defined as “behavioral modification especially through experience or conditioning” [Bartleby.com, 2003]. This is a phenomenon caused in some animals. Through experiences, human infants learn how they should communicate with others and what meanings environments have. While the direction to which development proceeds is defined in advance, learning changes its process case by case depending on experiences.

Table 4.1 describes differences between *development* and *learning*. The table focuses on how development and learning change the state of an agent, that is, (1) the direction of a state's change, (2) the reversibility of it, and (3) its dependency on experiences in an environment. In terms of these points, the differences between development and learning are summarized as follows.

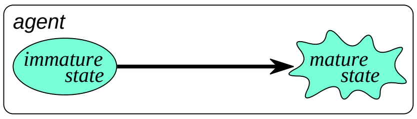
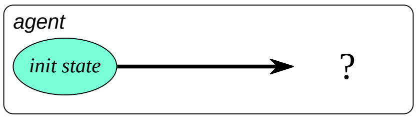
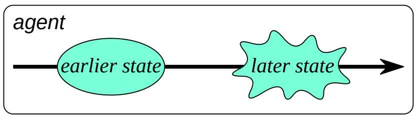
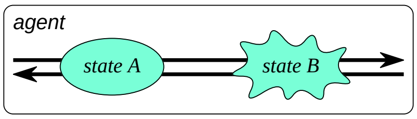
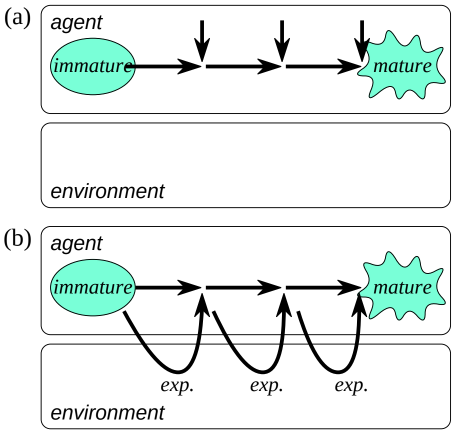
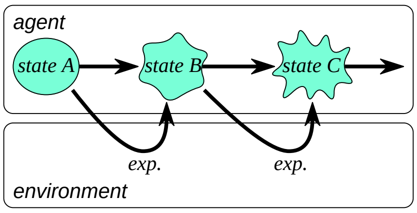
(1) The direction of a state's change in *development* is defined in advance while that of *learning* is determined through experiences.

(2) A state in *development* cannot return to an earlier state while a state in *learning* could go back.

(3) The change of a state in *development* does not depend on experiences or depends on only the timing to change. On the other hand, the change of a state in *learning* strongly depends on experiences.

The followings are discussed based on these differences.

Table 4.1: Differences between development and learning.

	Development	*Learning*
(1) the direction of a state's change	*Defined:* The direction of an agent's development is defined in advance.  	*Undefined:* The direction of an agent's learning changes depending on experiences. 
(2) the reversibility of a state's change	*Nonreversible:* The developmental state of an agent never return to an earlier state.  	*Reversible:* The learning state of an agent could go back to an earlier state. 
(3) the dependency of a state's change on experiences in an environment	*None or Semi-strong:* The development proceeds in either of two ways: (a) it is defined in advance how and when the development proceeds, or (b) "how" is defined while "when" depends on experiences.  (a) agent immature mature environment (b) agent immature mature exp. exp. exp. environment	*Strong:* The learning is never caused without experiences in an environment. The experiences shape how and when the learning progresses.  agent state A state B state C exp. exp. environment

4.1.1 Robot's Development and Caregiver's Development

In the learning of joint attention between an infant and a caregiver, it is conjectured that developments proceed in both the infant and the caregiver. An infant matures his/her internal mechanisms in parallel with learning, and a caregiver also adapts his/her functions according to the performance of the infant. In this chapter, the former is called *a robot's development* as a substitute for an infant's development, and the latter *a caregiver's development*. The developmental learning model presented in this chapter includes both of these developments.

Robot's Development
The robot's development means that the sensing and the actuating capabilities of a robot change from immaturity to maturity. In the proposed model, a robot develops its visual mechanism so that it gradually changes the sharpness of input camera images from coarse to fine states.

Caregiver's Development
The caregiver's development is defined as a process that a caregiver changes the criterion for task evaluation from an easy level to a difficult one. In the proposed model, the area in which a caregiver provides a good evaluation to a robot is gradually changed from a wide area to a narrow one by the caregiver.

4.1.2 Trigger for Development

A trigger for a development has two possibilities [Bremner, 1994]. One is that a development is triggered by a given clock, and the other is that a development is caused depending on experiences. The former means that an agent advances its development along a time schedule which is defined in advance. In this case, the development is certainly realized. On the other hand, the latter means that an agent advances its development through experiences in an environment. In other words, the progress of an agent's performance which is realized through experiences drives the development.

The developmental learning model presented in this chapter adopts a latter trigger,

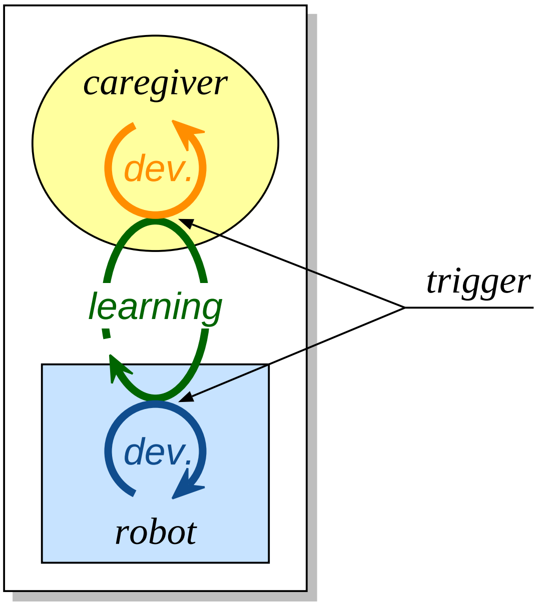


Figure 4.1: The concept of the relationship between the developments and learning in the proposed developmental learning model. Both a robot and a caregiver have their own developmental cycles in their internal mechanisms, and the learning cycle triggers the two developments. It means that both a robot and a caregiver develop synchronously as learning advances.

i.e. experiences. Both a robot and a caregiver develop their mechanisms depending on the progress of learning. A robot matures its visual mechanism according to its own performance; at the same time, a caregiver adjusts the criterion for the task evaluation according to the performance of the robot. The concept of the relationship between the developments and learning is shown in Figure 4.1. Both a robot and a caregiver have their own developmental cycles in their internal mechanisms, and the learning cycle triggers the two developments. It means that a robot's development and a caregiver's progress as learning advances. The proposed model takes an advantage of the explicit interactions between a robot and a caregiver. Furthermore, it is expected that the developments make learning more efficient by adjusting the difficulty of the joint attention task according to the performance of a robot.

4.2 Developmental Learning Model for Joint Attention

The mechanism of the proposed developmental learning model for joint attention is shown in Figure 4.2. This model consists of two modules:

(a) *a neural network as a learning module for a robot* which consists of four layers (an input layer, a retina one, a visual cortex one, and an output one) and has a developmental mechanism between the input and the retina layers, and

(b) *a task evaluator for a caregiver* which consists of two layers (an error layer and an evaluation one) and has a developmental mechanism between them.

The developmental mechanisms of a robot and a caregiver are called a robot's development and a caregiver's development, respectively.

Based on the model, a robot learns joint attention in an environment shown in Figure 3.1 according to the following procedure.

1. A robot first looks at a caregiver who is looking at an object (the situation shown in Figure 3.1 (a)) and obtains its camera image $\boldsymbol{I}$ as a sensor input. The angle $\boldsymbol{\theta} = [\theta_{pan},\ \theta_{tilt}]$ of the robot's camera head is not considered here because it is always fixed to zero when the robot is looking at the caregiver.

2. The robot inputs the image $\boldsymbol{I}$ to the neural network and computes a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ for the camera head. Then, the robot rotates its camera head by $\boldsymbol{\Delta\theta}$ (the situation shown in Figure 3.1 (b)).

3. The caregiver measures the gaze direction of the robot's camera and the position of the object that the caregiver is looking at, and then calculates the output error between them.

4. The caregiver determines task evaluation $V_k = 1$ or 0, each of which indicates the success of joint attention or the failure, based on the output error, and provides it to the robot.

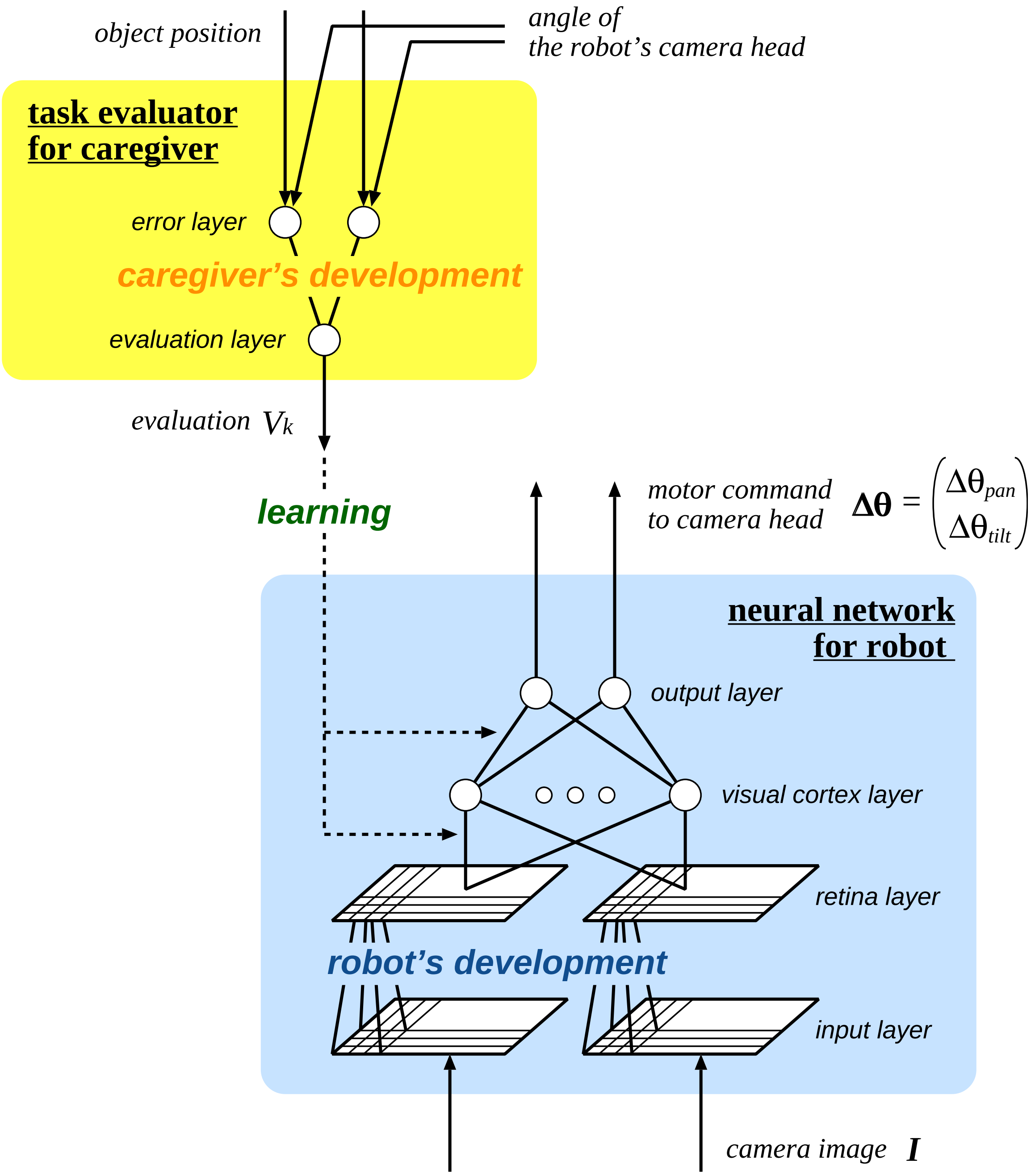


Figure 4.2: The developmental learning model for joint attention. The model consists of two modules: a neural network for a robot and a task evaluator for a caregiver. Both modules include a developmental mechanism: a robot's development and a caregiver's development, respectively. A robot learns joint attention under the task evaluation from a caregiver based on this model.

5. The robot modifies the connecting weights in the neural network based on the task evaluation V_k.

6. Return to 1.

In parallel with the learning process, the robot's development and the caregiver's development proceed in each module. The robot matures its visual mechanism in synchronization with the learning progress; at the same time, the caregiver advances the criterion for the task evaluation according to the task performance of the robot.

The following sections describe the learning mechanism based on task evaluation and two developmental mechanisms of the caregiver and the robot in order.

4.2.1 Learning Mechanism based on Task Evaluation

The robot learns joint attention by modifying the connecting weights in the neural network based on the task evaluation from the caregiver. The connecting weights W_k^{rc} between the retina layer and the visual cortex one and W_k^{co} between the visual cortex layer and the output one are modified based on the task evaluation V_k, where k indicates the learning time step. The task evaluation V_k has two possibilities, 1 or 0, in which $V_k = 1$ means that joint attention has succeeded while $V_k = 0$ means failure.

Based on the task evaluation V_k, the robot adjusts the connecting weights $W_k^{rc,co}$ for the next learning time step $k + 1$ as

$$W_{k+1}^{rc,co} = \begin{cases} W_k^{rc,co}, & \text{when } V_k = 1 \\ W_k^{rc,co} \pm \Delta W, & \text{when } V_k = 0, \end{cases} \tag{4.1}$$

where ΔW denotes a small random value. This adjusting method means that the robot keeps the current connecting weights if it has received an evaluation of the success of joint attention. In contrast, if the robot has received an evaluation of the failure, it slightly changes the connecting weights. The reason why the robot can adjust the connecting weights only to a random direction is that the caregiver is not able to tell the robot how the connecting weights should be adjusted since the caregiver does not know how the internal mechanism of the robot works. In this way, the robot gradually learns the connecting weights in the neural network to achieve joint attention based on task evaluation from the caregiver.

4.2.2 Mechanism of Caregiver's Development

In parallel with the learning progress, the caregiver's development proceeds in the task evaluator. The caregiver advances the criterion for the task evaluation according to the task performance of the robot.

At the learning time step k, the caregiver first observes the gaze direction of the robot's camera and the position of the object that the caregiver is looking at, and then calculates the absolute value of the output error e_k between them. Next, the task evaluation for joint attention V_k is determined as

$$V_k = \begin{cases} 1, & \text{when } e_k \le t_k \\ 0, & \text{when } e_k > t_k, \end{cases} \tag{4.2}$$

where t_k denotes a tolerance for the output error e_k. This means that the task of joint attention is judged as success if the output error e_k is less than or equal to the tolerance t_k, and it is judged as failure otherwise. The tolerance t_k is defined as

$$t_k = E_{k-1} - \epsilon \quad (\epsilon\text{: a small value}), \tag{4.3}$$

where E_{k-1} is a mean value of the output error e_{k-1} in various situations at the learning time step $k-1$. In other words, E_{k-1} represents the task performance of the robot. This evaluation method makes the caregiver

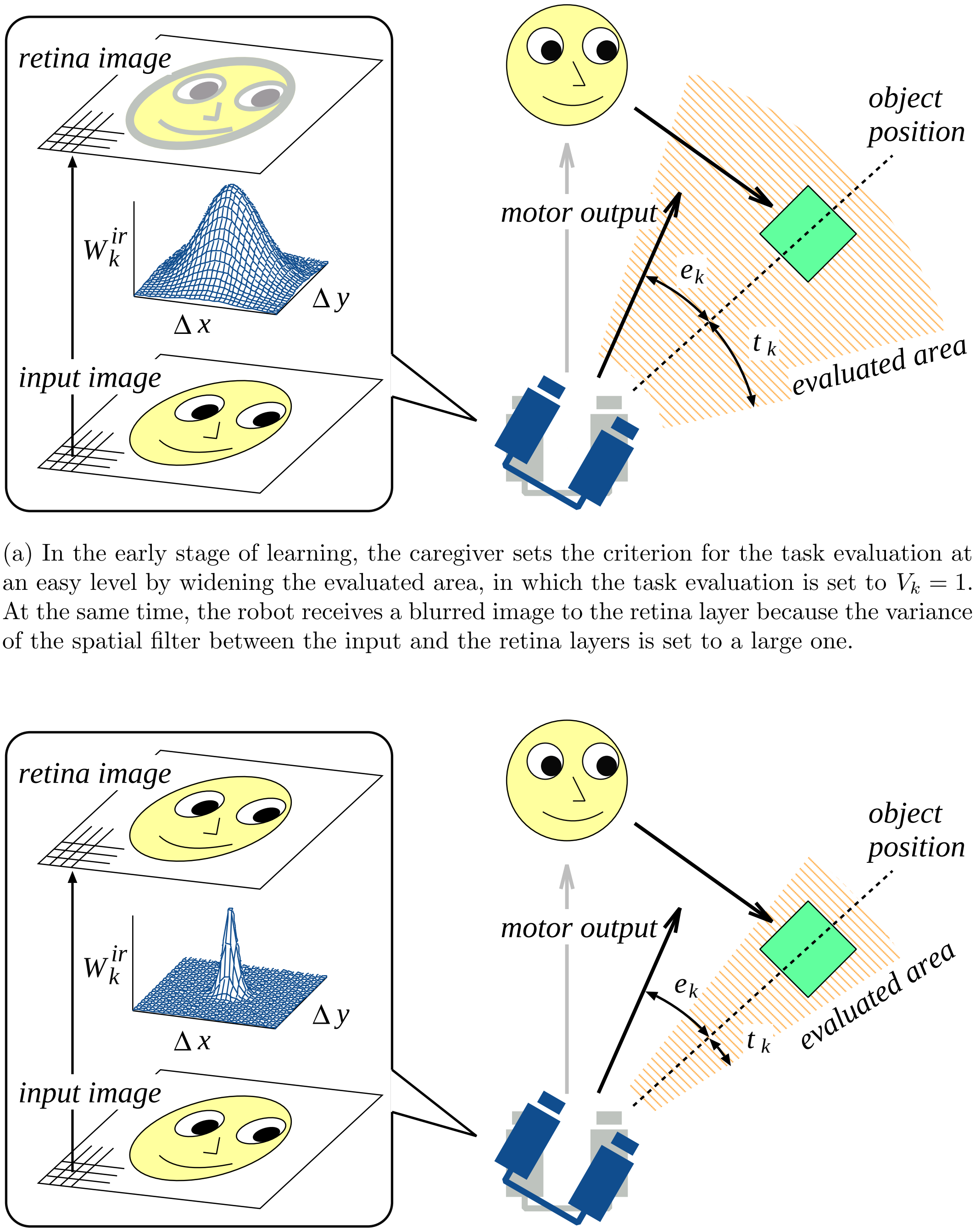
- provide an evaluation $V_k = 1$ if the output error e_k is better than the criterion that is slightly advanced than the last task performance of the robot, and
- provide $V_k = 0$ if the error e_k is worse than that.

Note that the tolerance t_k in Eq. (4.3) is updated only when

$$E_{k-1} < \min E_j \quad (0 \le j < k-1). \tag{4.4}$$

In other words, the advance of the caregiver's criterion for the task evaluation, i.e. the caregiver's development, is caused only when the task performance of the robot is improved.

The appearance of the caregiver's development is shown in the right side of Figure 4.3, in which (a) and (b) present the early stage of learning and the later one,



(a) In the early stage of learning, the caregiver sets the criterion for the task evaluation at an easy level by widening the evaluated area, in which the task evaluation is set to $V_k = 1$. At the same time, the robot receives a blurred image to the retina layer because the variance of the spatial filter between the input and the retina layers is set to a large one.

(b) In the later stage of learning, the caregiver sets the criterion for the task evaluation at a difficult level by narrowing the evaluated area. The robot becomes to receive a clear image to the retina layer because it makes the variance of the spatial filter small.

Figure 4.3: The robot's development (left) and the caregiver's development (right). The robot develops its visual mechanism, and the caregiver develops its criterion for the task evaluation.

respectively. The left side of Figure 4.3 shows the robot's development, which is explained in the next section. In the figure, the sector formed area with slant lines, namely an evaluated area, indicates the criterion for the task evaluation that is defined by the tolerance t_k. If the motor output of the robot shown as an arrow starting from the robot's camera lies in the evaluated area, the task evaluation $V_k = 1$ is provided to the robot, otherwise $V_k = 0$. Figure 4.3 shows the caregiver's development as the followings.

(a) In the early stage of learning, the caregiver sets the evaluated area as a wide one because the mean value of the output error E_k of the robot is large value, so that the robot can easily learn the task of joint attention.

(b) In the later stage of learning, the caregiver changes the evaluated area to a narrow one and makes the robot accurately learn joint attention because the output error E_k becomes small.

Owing to the caregiver's development, the caregiver can facilitate the robot learning joint attention by evaluating the robot based on the appropriate criterion to the task performance of the robot. This mechanism is expected to make the robot's learning more efficient.

4.2.3 Mechanism of Robot's Development

In parallel with the caregiver's development, the robot's development proceeds between the input layer and the retina one in the neural network. The robot matures its visual mechanism by changing the connecting weight W_k^{ir} between the two layers according to the improvement of the robot's task performance.

At the learning time step k, the robot first obtains its camera image $\boldsymbol{I}$ in which the caregiver's face is extracted, and sends it to the input layer. Then, the image is forwarded to the retina layer through the connecting weight W_k^{ir} between the two layers. The connecting weight W_k^{ir} is defined as a Gaussian spatial filter

$$W_k^{ir} = \exp\left(-\frac{(x-x_c)^2 + (y-y_c)^2}{2{\sigma_k}^2}\right), \tag{4.5}$$

where $(x,\ y)$, $(x_c,\ y_c)$, and σ_k denote a position in the input image, a position of the target pixel of the spatial filter, and the variance of the filter at the learning time step k, respectively. This means that the spatial filter works as a smoothing filter, which blurs the image sent from the input to the retina layers. The variance of the spatial filter σ_k, which defines how the image is blurred, is determined as

$$\sigma_k = \sigma_{init} \left(\frac{E_{k-1} - E_{fin}}{E_{init} - E_{fin}} \right), \tag{4.6}$$

where E_{init} and E_{k-1} indicate the mean of the output error at the beginning of learning and that at the learning time step $k-1$, and E_{fin} indicates a tolerance at the end of learning. The coefficient σ_{init}, which is the initial value of σ_k, and E_{fin} are given as constant values which define the initial condition and the end one of the visual development of the robot, respectively. In other words,

- if σ_{init} has a large value, the visual development starts from a more immature level, that is, the retina image is filtered to a blurrier one than that when σ_{init} is small.
- If E_{fin} has a small value, the visual development continues for a long time until E_{k-1} equals to E_{fin}.

Note that the update of σ_k in Eq. (4.6) is triggered only when

$$E_{k-1} < \min E_j \quad (0 \leq j < k-1) \tag{4.7}$$

as well as the update of the tolerance t_k in the caregiver's development. It means that the maturation of the robot's visual mechanism, i.e. the robot's development, is also triggered by the improvement of the robot's task performance.

The appearance of the robot's development is shown in the left side of Figure 4.3. The figure shows that the input image is blurred through the Gaussian spatial filter W^{ir}_k, which is applied to all small areas $(\Delta x \times \Delta y)$ in the input image, and then the filtered image is sent to the retina layer. The robot's development shown in Figure 4.3 represents the followings.

(a) In the early stage of learning, the variance of the spatial filter σ_k is large because the output error of the robot $E_{k-1} \approx\ E_{init}$ in Eq. (4.6). Therefore, the robot receives a blurred image on the retina layer.

(b) In the later stage of learning, the variance of the spatial filter σ_k becomes small because $E_{k-1} \approx E_{fin}$. Therefore, the robot receives a clear image on the retina layer.

Owing to the robot's development, it is expected that only characteristic features in the input image are extracted through the spatial filter in the early stage of learning. As a result, this mechanism could allow the robot to detect some significant features for joint attention and to acquire the ability in well-organized internal representation.

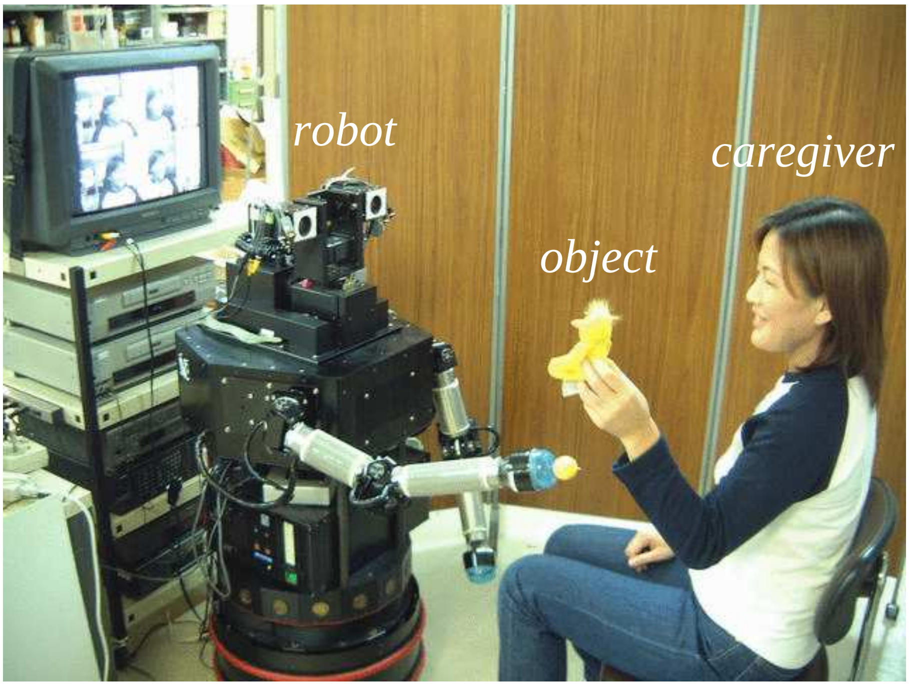
4.3 Experimental Setup

The validity of the proposed model is examined in some experiments using an actual robot. An experimental environment which includes a robot with two cameras, a human caregiver, and an object is shown in Figure 4.4 (a), and the camera head of the robot is shown in (b). The robot and the caregiver are seated face-to-face at the fixed positions. The caregiver has a salient object in her hand, which is moved to various positions. Note that this experiment uses only one object since the number of objects does not matter in the case that the caregiver explicitly evaluates the joint attention task of the robot. At the same time, the robot observes the caregiver through its cameras and obtains its camera images as inputs for the neural network. The robot's camera head is rotated on the pan and the tilt axes according to a motor output from the neural network. The robot's right arm with touch sensors is used to obtain learning data or to shift the step from (1) looking at the caregiver to (2) turning its camera head as described in Section 3.1.

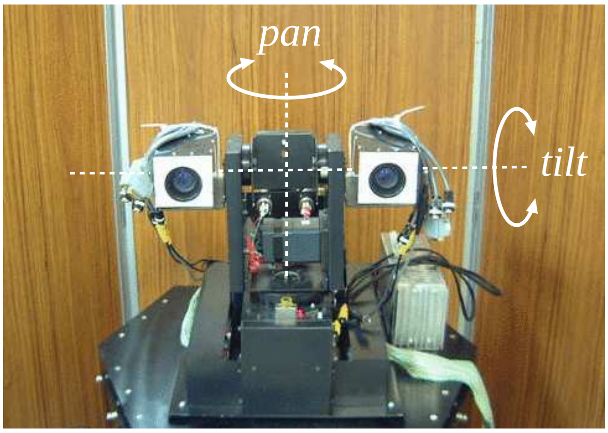
Under this experimental setup, learning datasets are acquired in the real environment in advance, and then off-line learning is conducted. The learning data include 75 datasets of

- a left camera image $\boldsymbol{I}$ in which the caregiver's face is extracted as a window of which size is 30 × 25 [pixel] and

- a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ to shift the gaze direction of the robot's camera from the caregiver's face to the object that the caregiver is looking at.



(a) An experimental environment for joint attention in which a robot with two cameras, a human caregiver, and an object are indicated. The robot first looks at the caregiver who is looking at the object and captures its camera image as an input.



(b) The robot's camera head, which rotates on the pan and the tilt axes.

Figure 4.4: An experimental setup for developmental learning of joint attention.

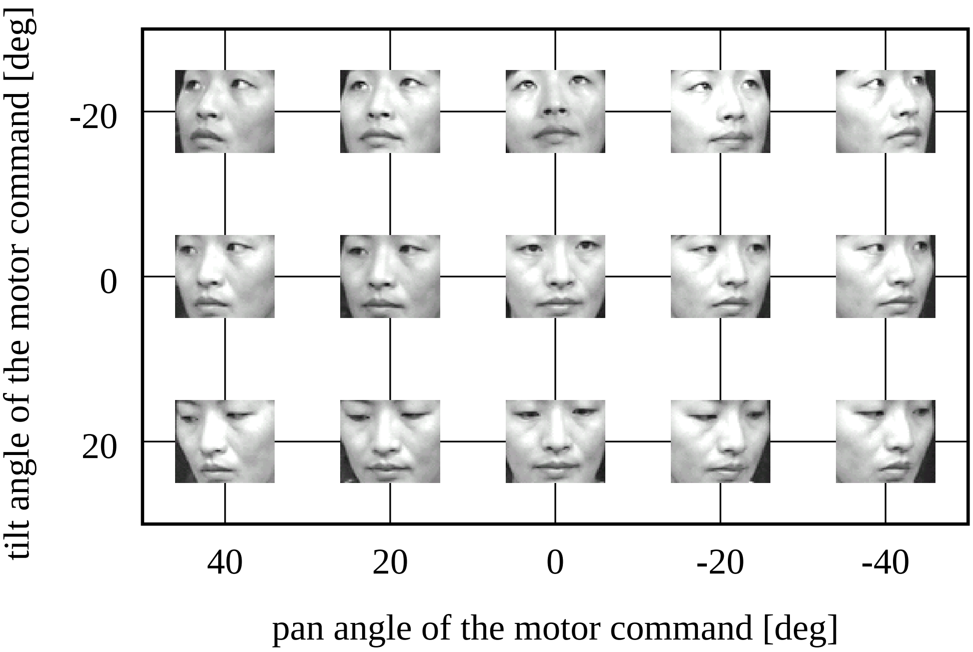

tilt angle of the motor command [deg]
-20
0
20
40
20
0
-20
-40
pan angle of the motor command [deg]

Figure 4.5: Examples of the input image distributed in the robot's motor output space. The images of the caregiver's face are extracted by template matching. The position at which each image is placed denotes the robot's motor command to look at the object that the caregiver is looking at.

The caregiver's face is extracted from the camera image using template matching. The object is set at 15 positions every 20 degrees in the pan range from -40 to 40 [deg] and in the tilt range from -20 to 20 [deg] as shown in Figure 4.5, in which examples of the input image are presented. The horizontal and the vertical axes indicate the pan and the tilt angles of the robot's motor command to shift its gaze direction from the caregiver's face to the object that the caregiver is looking at. The images show the input images when the robot observes the caregiver who is looking at an object at various positions. The values of the brightness at all pixels in the image are used as inputs to the neural network. In other words, any characteristic feature does not abstracted in advance. The robot captures five datasets at each position. The acquired 75 datasets are repeatedly applied to the proposed model in the experiments of off-line learning.

The experiments described in this section uses a neural network that has 750 neurons for an input image of which size is 30 × 25 [pixel] on the input layer, 750 neurons on the retina one, seven neurons on the visual cortex one, and two neurons for the pan and the tilt angles on the output one. The number of the neurons on the visual cortex layer was determined based on preliminary experiments which had shown the output error under 5 [%].

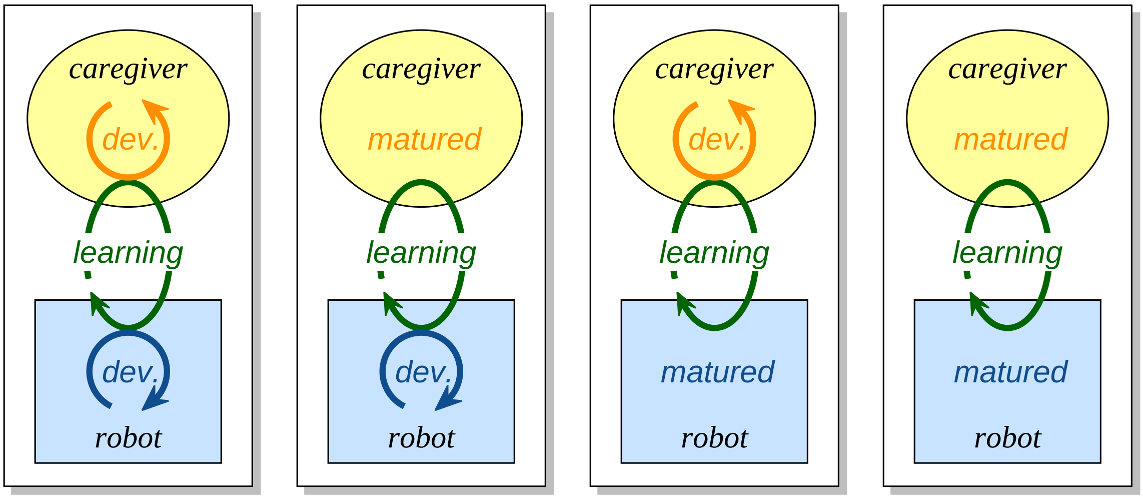
4.4 Experimental Results

Several learning experiments are conducted based on the proposed model. The evaluated performances are

- the learning speed,
- the effect of the trigger for the development on the learning speed,
- the output error after the learning,
- the effect of the internal representation of the acquired neural network on the output error, and
- the final task performance of joint attention in a real environment.

These performances of the proposed model are evaluated by comparing to three other models. Figure 4.6 shows (a) the proposed model, namely *RC-dev. model*, and the compared models: (b) *R-dev. model*, (c) *C-dev. model*, and (d) *Matured model*. *RC-dev. model* obviously includes both the mechanism of the robot's development and that of the caregiver's development. On the other hand, *R-dev. model* and *C-dev. model* include either of the robot's development or the caregiver's development, respectively. The caregiver in *R-dev. model* and the robot in *C-dev. model* have a matured mechanism for the task evaluation and for the vision, respectively. In *Matured model*, both the robot and the caregiver have the matured mechanisms. The matured mechanism of the caregiver means that the criterion for the task evaluation described in Eq. (4.3) is defined as

$$t_k = \epsilon' \quad (\epsilon'\text{: a small value}). \tag{4.8}$$



(a) *RC-dev. model* (the proposed model)

(b) *R-dev. model*

(c) *C-dev. model*

(d) *Matured model*

Figure 4.6: The learning models which are compared in the evaluation of the learning performance of the proposed model. (a) *RC-dev. model* is the proposed model, which has both the robot's development and the caregiver's development. (b) *R-dev. model* includes the robot's development while the caregiver has a matured mechanism. In contrast, (c) *C-dev. model* includes the caregiver's development while the robot has a matured one. (d) *Matured model* has neither of the developments and both the robot and the caregiver have a matured mechanism.

It indicates that the caregiver sets the criterion for the task evaluation at a high level from the beginning of learning. On the other hand, the matured mechanism of the robot's vision means that the Gaussian spatial filter between the input and the retina layers presented in Eq. (4.5) is defined as

$$W_k^{ir} = \begin{cases} 1, & x = x_c, \ y = y_c \\ 0, & x \neq x_c, \ y \neq y_c. \end{cases} \tag{4.9}$$

It means that the robot can receive a clear image on the retina layer from the beginning of learning since the spatial filter works as a transmission one. The experiments using the above four models allow us to understand the effectiveness of each developmental mechanism.

The learning performances of the robot were evaluated with various parameters. In this section, an experimental result of which parameters were set $\Delta W_{max} = 0.007$, $\epsilon = 0.02$, $\sigma_{init} = 3.0$, $E_{fin} = 0.05$, and $\epsilon' = 0.02$ in Eq. (4.1) – (4.8) is presented since all results seemed to show similar performances.

4.4.1 Evaluation of Learning Speed

The learning speed of the proposed model, namely *RC-dev. model*, is evaluated by comparing to the three models: *R-dev. model*, *C-dev. model*, and *Matured model.* Figure 4.7 shows the changes of the output errors of the four models over the learning. The horizontal axis indicates the learning time step k, and the vertical one denotes the mean value of the normalized output error E_k of the robot. In the graph, the solid curve, the long dashed one, the short dashed one, and the dotted one respectively present the error changes of *RC-dev. model*, *R-dev. model*, *C-dev. model*, and *Matured model.*

From the result, we can see that the mechanism of the caregiver's development accelerates the learning of joint attention. The learning speed of *RC-dev. model* including the caregiver's development is improved compared to that of *R-dev. model* without the development. It should be noted that the mechanisms other than the caregiver's development in these models are exactly same. Moreover, the same is true of *C-dev. model* and *Matured model.* It is considered that the reason why the caregiver's development accelerates the learning of joint attention is that the caregiver can provide appropriate task evaluation for the robot according to its task performance by changing the criterion for the evaluation. While the output error of the robot E_k has a large value, the caregiver facilitates the robot learning joint attention in easy situations by setting the tolerance t_k for the task evaluation as a large one based on Eq. (4.3). Furthermore, when the robot's output error E_k becomes small, the caregiver can make the robot accurately learn the task by decreasing the tolerance t_k.

In contrast to the caregiver's development, the robot's development is confirmed to have an effect to delay the learning of joint attention. The learning speed of *RC-dev. model*, which includes the robot's development, is later than that of *C-dev. model* without the development. Moreover, the same is true of *R-dev. model* and

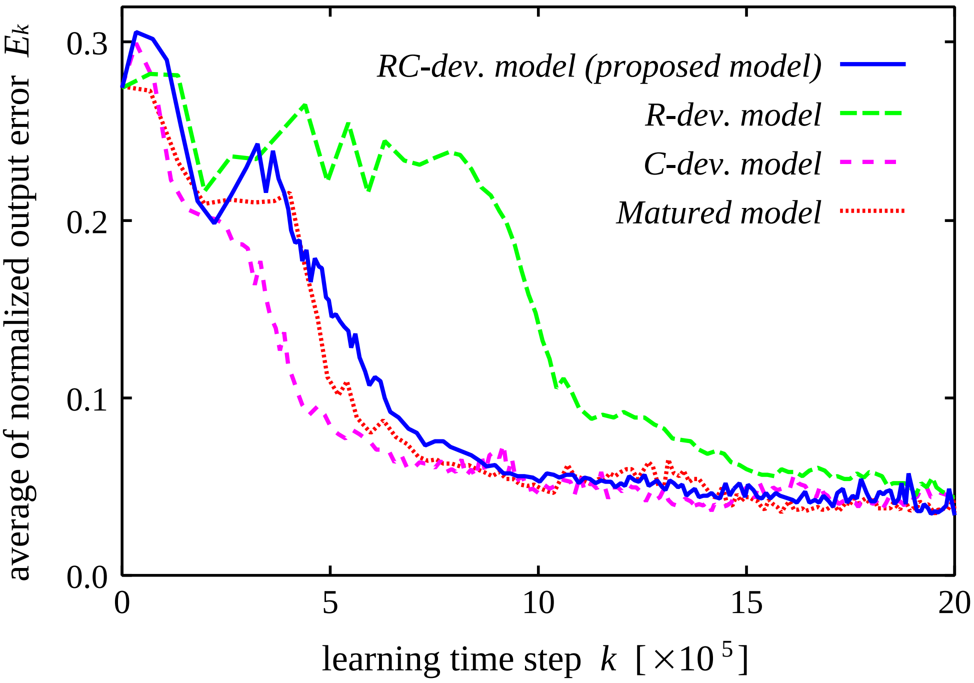


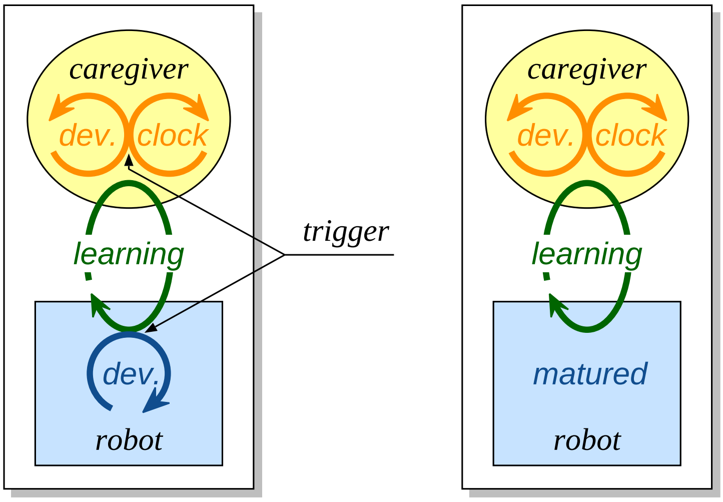
Figure 4.7: The changes of the mean values of the normalized output errors E_k over the learning. The solid curve, the long dashed one, the short dashed one, and the dotted one represent the error changes of *RC-dev. model*, *R-dev. model*, *C-dev. model*, and *Matured model*, respectively. This result shows that the caregiver's development included in *RC-dev. model* and *C-dev. model* can accelerate the learning speed of joint attention while the robot's delays that.

Matured model. Note that the mechanisms other than the robot's development in these two pairs are exactly the same. It is considered that the reason why the robot's development delays the learning speed of joint attention is that the robot cannot have enough information to achieve joint attention in the early stage of the learning because of the immaturity of its visual mechanism. While the output error E_k of the robot has a large value, the robot receives a blurred image on the retina layer since the robot sets the variance of the spatial filter between the input and the retina layers as a large one under Eqs. (4.5) and (4.6). The delaying of the learning speed of joint attention by the robot's development seems like a disadvantage of the proposed model; however, an advantage of the robot's development is shown in Section 4.4.3.

4.4.2 Relationship between Learning Speed and Trigger for Development

It is examined how the trigger for the caregiver's development has an effect to accelerate the learning of joint attention. *RC-dev. model* and *C-dev. model*, in which the caregiver's development is triggered by the learning progress, are compared with *RC'-dev. model* and *C'-dev. model*, in which the caregiver's development is caused by a given clock. Figure 4.8 shows the conceptual images of *RC'-dev. model* and *C'-dev. model*, which respectively correspond to the models in Figures 4.6 (a) and (c). The clock trigger for the development means that the development proceeds along a time schedule that is defined by a designer in advance. Note that the robot's development in *RC'-dev. model* is caused by the learning progress as well as that in *RC-dev. model*.

The changes of the output errors of the four models and that of the tolerance for the task evaluation are shown in Figure 4.9, in which (a) indicates the results



(a) *RC'-dev. model* (b) *C'-dev. model*

Figure 4.8: The compared learning models in which the caregiver's development is triggered by a given clock. The robot's development in *RC'-dev. model* is caused by the learning progress as well as that in *RC-dev. model*.

of *RC-dev. model* and *C-dev. model* and (b) indicates that of *RC'-dev. model* and *C'-dev. model*. The horizontal axis of each graph denotes the learning time step k, and the vertical one denotes the mean value of the normalized output error E_k and the tolerance for the task evaluation t_k. A solid curve shows the change of E_k, and a broken one shows that of t_k, which means how the caregiver's development proceeds. The results shown in Figure 4.9 (a) are the same as that in Figure 4.7, and those shown in (b) present the best performance in some experiments using various clock triggers.

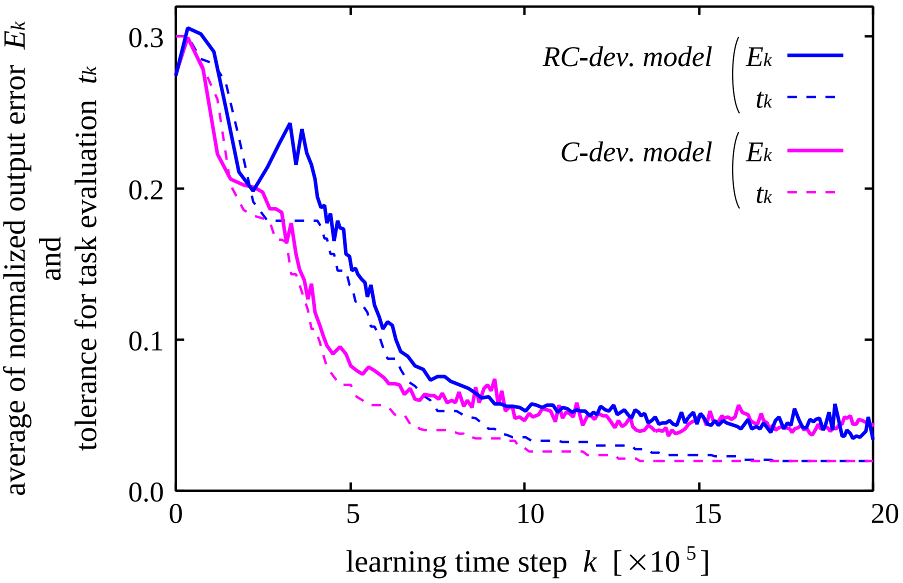
From the comparison of Figures 4.9 (a), (b) and Figure 4.7, we may conclude as follows.

- The caregiver's development has an effect to accelerate the learning of joint attention regardless of whether the development is triggered by the learning progress or a given clock.

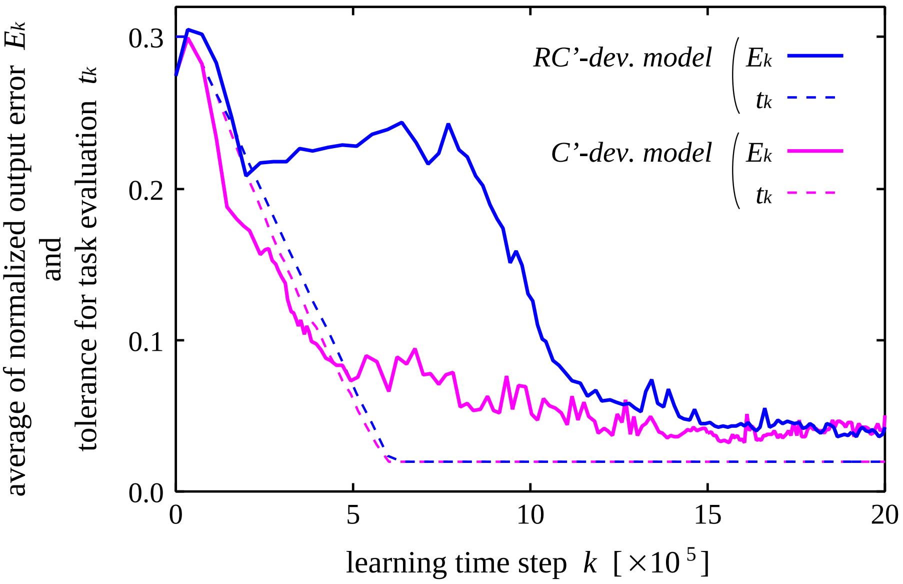
- In the case that the caregiver's development is triggered by a given clock, the designer has to test many kinds of triggers because the effect of the acceleration of learning strongly depends on the schedule of the clock trigger. Furthermore, the best performance of the acceleration by the caregiver's development using a clock trigger is almost the same as that using a trigger based on the learning progress.

- In contrast, in the case that the caregiver's development is caused by the learning progress, the learning speed is certainly accelerated. The reason is that the caregiver can facilitate the robot learning joint attention at an appropriate task level according to the robot's performance by changing the criterion for the task evaluation. It is verified that the above is true when $-0.05 \leq \epsilon \leq 0.05$ in Eq. (4.3).

It is concluded that the caregiver's development triggered by the learning progress has an advantage to accelerate the learning of joint attention because the caregiver evaluates the robot's task performance under the appropriate criterion for the robot.



(a) The learning curves of *RC-dev. model* and *C-dev. model*, in which the caregiver's development is triggered by the learning progress. The caregiver's development always accelerates the learning of joint attention.



(b) The learning curves of *RC'-dev. model* and *C'-dev. model*, in which the caregiver's development is triggered by a given clock. The caregiver's development caused by a clock cannot always accelerate the learning of joint attention.

Figure 4.9: The relationship between the learning speed and the trigger for the caregiver's development. The caregiver's development triggered by the learning progress always accelerates the learning while that triggered by a given clock cannot always.

4.4.3 Evaluation of Final Task Performance

The acquired performances of the robot are evaluated. About the four models: *RC-dev. model*, *R-dev. model*, *C-dev. model*, and *Matured model*, the output error E_k to unknown data are measured after learning. The unknown data mean that the position of the object and the lighting condition are changed compared to those of the trained data while the caregiver is the same person. Figure 4.10 shows the mean value of the normalized output error E_k to 45 unknown inputs and its standard deviation of each model. Starting from the left, the results of *RC-dev. model*, which is the proposed model, *C-dev. model*, *R-dev. model*, and *Matured model* are presented.

From this result, it is confirmed that the robot's development improves the final task performance. The output error of *RC-dev. model*, which includes the robot's development, is less than that of *C-dev. model* without the development. At the same time, the output error of *R-dev. model* including the robot's development is also less

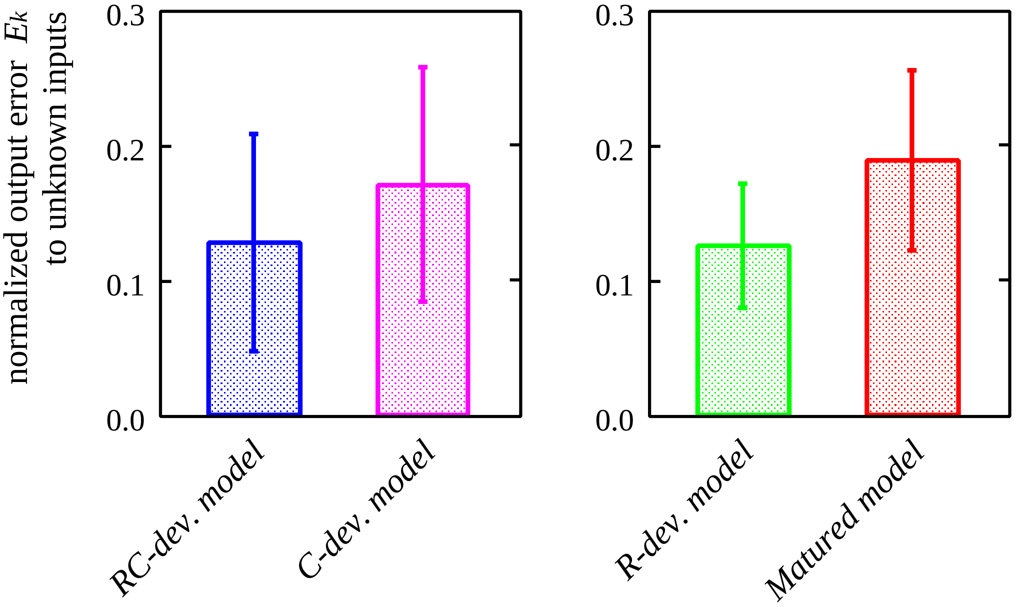


Figure 4.10: The mean value of the normalized output error to unknown inputs. Each bar shows the output error to 45 unknown inputs and its standard deviation. This result shows that the robot's development included in *RC-dev. model* and *R-dev. model* can improve the final task performance of joint attention.

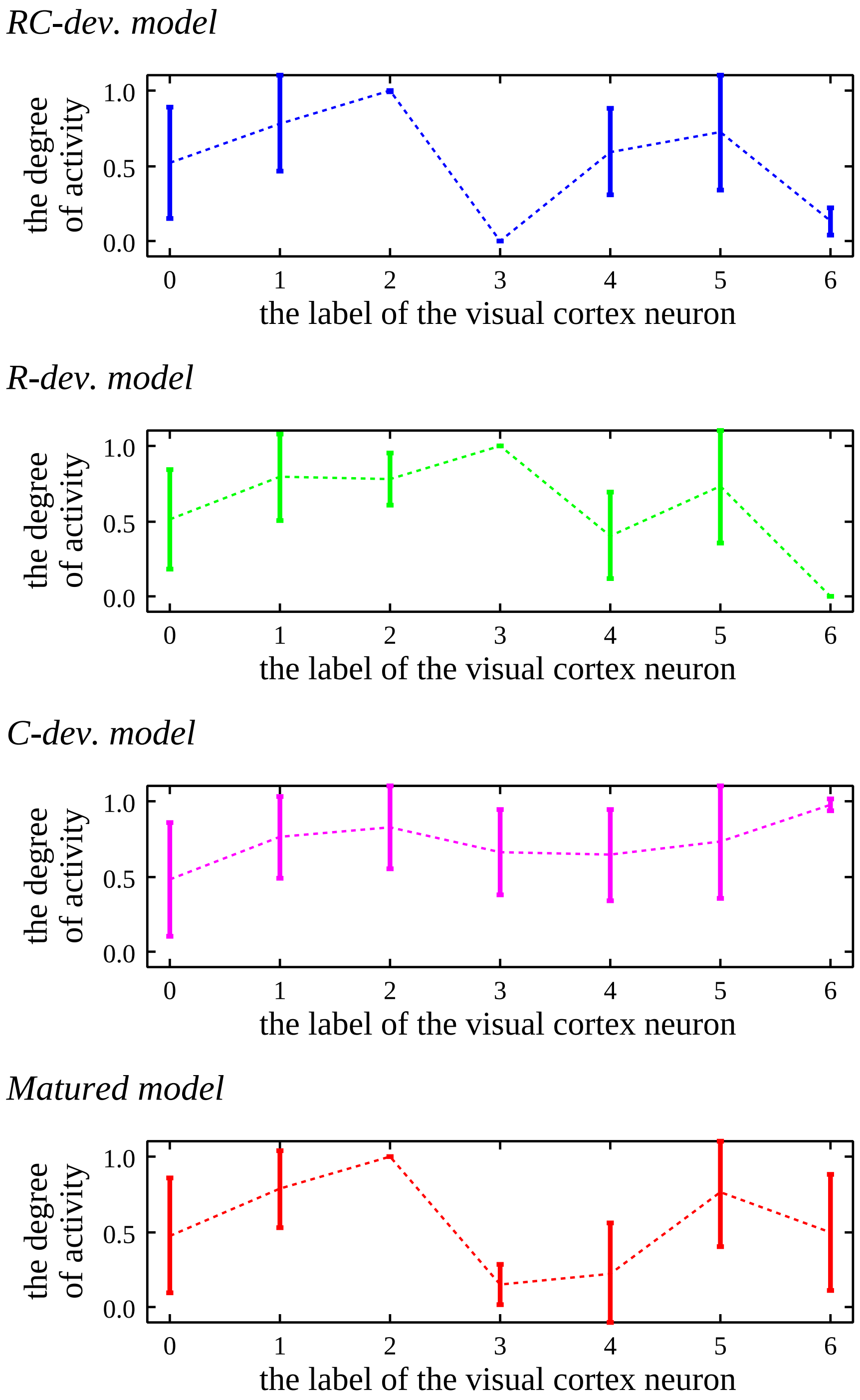
than that of *Matured model* without the development. It should be noted that the mechanisms other than the robot's development in each pair are exactly the same. The standard deviations of the output errors have large values; however, the validity of the results has been verified using a statistical test by Tukey's method of multiple comparison [Ishimura, 1992]. It has proved that the experimental result shown in Figure 4.10 has significant differences between

- *RC-dev. model* and *C-dev. model*,
- *RC-dev. model* and *Matured model*,
- *R-dev. model* and *C-dev. model*, and
- *R-dev. model* and *Matured model*

when the level of significance is 5 [%]. In other words, all the models that include the robot's development have significant differences against all the models without the robot's development. About the statistical test by Tukey's method, refer to Appendix A for more detailed description. The reason why the robot's development improves the final task performance is considered that the immaturity of the robot's visual mechanism in the early stage of learning realizes the abstraction of the input image. The abstracted input image is expected to allow the robot to acquire well-organized internal representation to achieve joint attention.

4.4.4 Relationship between Final Task Performance and Internal Representation of Neural Network

It is evaluated how the internal representation of the acquired neural network has an effect to improve the final task performance owing to the robot's development. The activities of the visual cortex neurons in the acquired neural network are examined when 45 unknown data are applied to the neural network. The unknown data are the same ones that used in Section 4.4.3. The mean value of the activity of each neuron and its standard deviation are shown in Figure 4.11. Starting from the top, the results of *RC-dev. model*, *R-dev. model*, *C-dev. model*, and *Matured model* are presented. The horizontal axis in each graph denotes the label of the visual cortex



RC-dev. model

R-dev. model

C-dev. model

Matured model

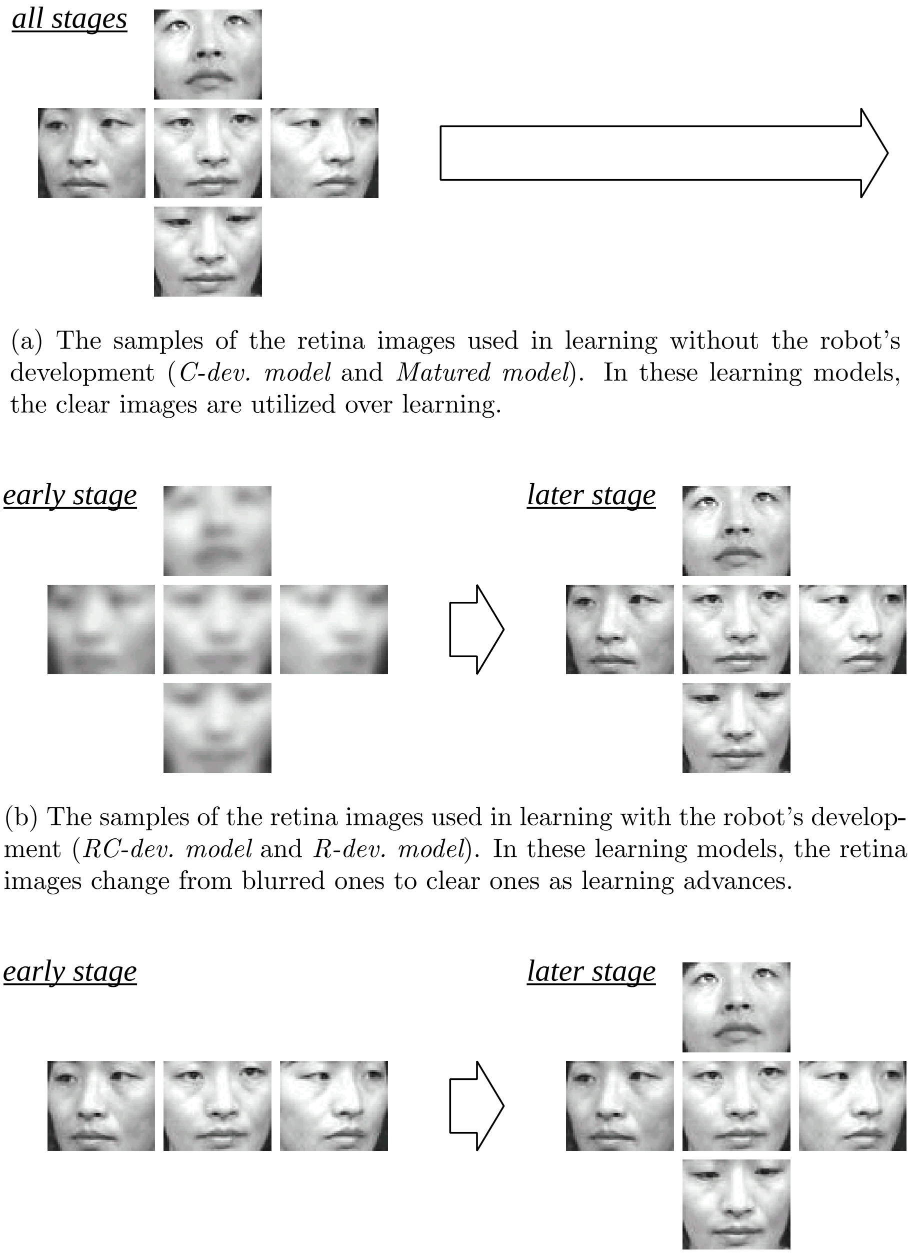
Figure 4.11: The mean value of the activities of the visual cortex neurons and its standard deviation when 45 unknown inputs are applied to the models. The neurons of which standard deviations equal to zeros are not utilized at all to realize joint attention while the neurons that have large standard deviations are leveraged effectively.

Table 4.2: The mean number of the unutilized neurons in the visual cortex layer when 50 patterns of parameters are applied to the neural network.

learning model	the mean number of the unutilized neurons [/7 neurons]
RC-dev. model	1.2
R-dev. model	1.2
C-dev. model	0.7
Matured model	0.6

neurons, and the vertical one denotes the degree of the activity of each neuron. From this result, it is confirmed that the neural networks include some neurons of which standard deviations equal to zero. It means that the neurons are not utilized to achieve joint attention. The neurons of #2 and #3 in *RC-dev. model*, #3 and #6 in *R-dev. model*, and #2 in *Matured model* are unutilized neurons. Furthermore, it should be noted that the number of the unutilized neurons is increased by adding the mechanism of the robot's development into the learning models. Table 4.2 shows the mean numbers of the unutilized neurons in the visual cortex layer when 50 patterns of parameters are applied to the networks. The results of Figure 4.11 and Table 4.2 clearly show that the mechanism of the robot's development downsizes the internal representation of the neural network by reducing the neurons that are utilized to achieve joint attention.

Then, the reason why the internal representation of the neural network is downsized through learning is analyzed. It is conjectured that the downsizing is attributable to the change of the complexity of the input images, which is caused by the robot's visual development. Figures 4.12 (a) and (b) show the samples of the retina images utilized through learning without/with the robot's development. The retina images are projected through the spatial filter, which presents the robot's visual development, from the input layer. In the case that the learning model does not include the robot's development (*C-dev. model* and *Matured model*), the retina image remains clear over learning as shown in Figure 4.12 (a). By contrast, in the case that the learning model includes the robot's development (*RC-dev. model* and *R-dev. model*), the retina image changes from a blurred one to a clear one as learning



(a) The samples of the retina images used in learning without the robot's development (*C-dev. model* and *Matured model*). In these learning models, the clear images are utilized over learning.

(b) The samples of the retina images used in learning with the robot's development (*RC-dev. model* and *R-dev. model*). In these learning models, the retina images change from blurred ones to clear ones as learning advances.

(c) The samples of retina images used in staged learning. The images including only the horizontal variance are applied in the early stage of learning, and the images including the vertical one are added in the later stage of learning.

Figure 4.12: The changes of the retina images over learning (a) when the robot learns without the visual development, (b) when the robot learns with the development, or (c) when the robot learns in stages.

advances as shown in (b). From these images, it can be found that the blurred images in (b) keep the horizontal component of the image variance compared to the vertical one. The variance in the vertical direction is almost lost because of the immature mechanism of the robot's vision. From the difference of the kept variance, it is conjectured that the robot with the developmental mechanism learns only the principal component with large variance in the early stage of learning and then learns the other components with small variance in the later stage. Such staged learning process is expected to simplify the task of joint attention and to enable the robot to acquire the ability by using small number of the visual cortex neurons. This conjecture is verified by conducting staged learning in which the input images with horizontal variance and those with vertical one are applied in incremental steps. Figure 4.12 (c) show the samples of the retina images used in the staged learning. In the early stage of learning, the images including only the horizontal variance are utilized, and those including the vertical one are added in the later stage of learning. As the result of the staged learning, the mean number of the unutilized neurons in the acquired neural network was 1.0 [/7 neurons]. Compared to the results shown in Table 4.2, it is confirmed that the neural network acquired through the staged learning includes more unutilized neurons as well as the network acquired through learning with the robot's development. This result supports the conjecture that the robot's visual development enables the robot to learn the joint attention task in incremental steps and to acquire the ability in a downsized and well-organized way.

From the above, the relationship between the final task performance and the internal representation of the neural network is summarized as follows.

- Owing to the robot's development, which is a visual development, an input image is blurred in the early stage of learning. The blurred input image allows the neural network to learn joint attention in an abstracted input space in which only principal components of the image survives.
- As a result, the neural network including the robot's development can acquire well-organized internal representation to achieve joint attention, and this leads the robot to acquire a high task performance.

It is concluded that the robot's development can improve the final task performance of joint attention since it makes the neural network acquire downsized and well-organized internal representation by abstracting the input image.

4.4.5 Experiments in Real Environment

Finally, the validity of the acquired mechanism based on the proposed developmental learning model is evaluated using an actual robot. The experimental setup was shown in Figure 4.4. The neural network acquired through learning based on the proposed model, i.e. *RC-dev. model*, is implemented in the actual robot. The caregiver is the same person as that in the learning. The other conditions, e.g. lighting conditions, the object position, and so on, are changed from the trained ones. The experiment is performed along the following procedure.

1. The robot observes the caregiver who is looking at an object.

2. The robot extracts the caregiver's face from its camera image $\boldsymbol{I}$ by template matching and inputs the image to the neural network.

3. The robot generates a motor output $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ based on the neural network and then rotates its camera head according to the motor command.

4. As a result, the task of joint attention is judged as success by the caregiver if the robot has detected the object that the caregiver is looking at under the condition
$$\sqrt{(x - cx)^2 + (y - cy)^2} < \frac{W_x}{6}, \tag{4.10}$$
where $(x,\ y)$, $(cx,\ cy)$, and W_x are the position of the object in the robot's camera image, the center position of the image, and the width of the image, respectively (see Figure 4.13).

Examples of the robot's camera image when it observes the caregiver are shown in Figure 4.14. In each image, a rectangle denotes the detected position of the caregiver's face, and the image enclosed in it is input to the neural network. The motor output of the neural network is drawn as a line, in which the horizontal component and the vertical one show the pan and the tilt angles of the motor command, respectively. Note that the line does not mean the gaze of the caregiver but means the motor

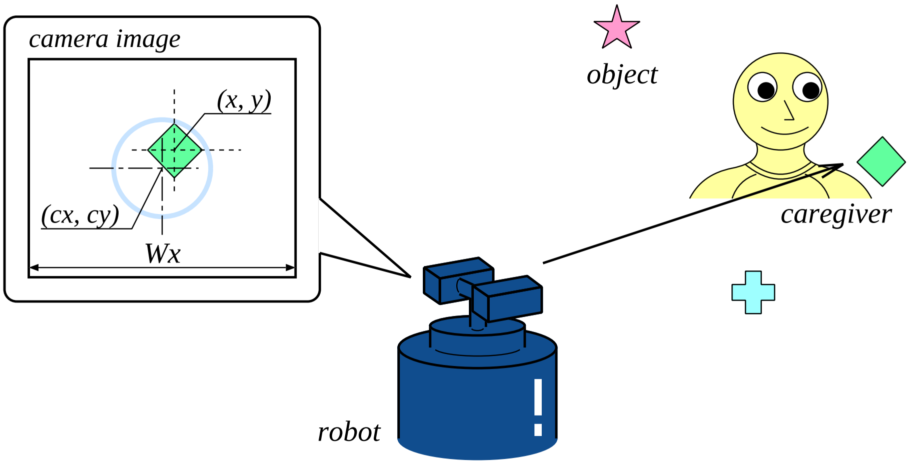


Figure 4.13: The criterion for the success of joint attention. If the object that the caregiver is looking at is detected in the center of the robot's camera image, joint attention succeeds.

command of the robot. From the results shown in Figure 4.14, we can see that the robot's motor command to perform joint attention is well estimated by the acquired neural network since the direction of the line corresponds to the direction of the caregiver's gaze, which is conjectured from the image. In the experiments, the robot had 20 trials in which the object's position was changed randomly and tried to realize joint attention by rotating its camera head based on the generated motor commands. As a result, it was confirmed that the robot can achieve joint attention at the success rate 95 [%] (=19/20 [trials]). This result shows that the proposed model has enough capability to make the robot acquire the ability of joint attention based on the task evaluation from the caregiver.

4.5 Discussion and Future Work

This chapter has presented a developmental learning model by which a robot learns joint attention based on caregiver's evaluation. It has been suggested in cognitive developmental science that caregivers strongly facilitate infants acquiring new abilities. At the same time, the developmental mechanisms of infants and caregivers are also

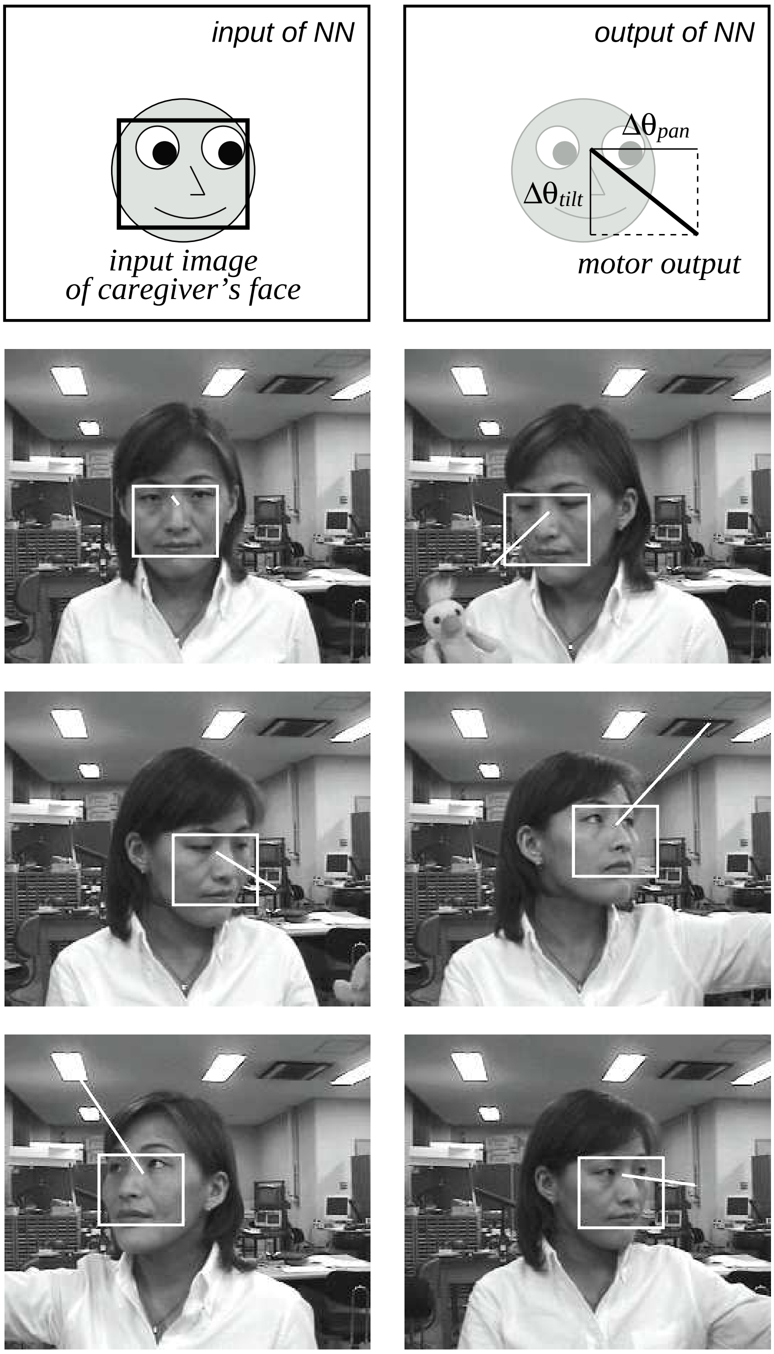


Figure 4.14: Experiments of joint attention in a real environment. The robot performs joint attention by the neural network acquired based on the proposed developmental learning model. An image enclosed in a rectangle indicates the input to the neural network, and the line shows the motor output from the network. The robot rotates its camera head based on the motor output and tries to realize joint attention.

aids for the learning of the infants. On the basis of the insights, the proposed model consists of the learning mechanism based on caregiver's evaluation, the mechanism of the robot's development, and that of the caregiver's development. As developmental mechanisms, the robot matures its visual mechanism in parallel with the learning progress; at the same time, the caregiver adjusts the criterion for the task evaluation according to the performance of the robot. From the experimental results which evaluated the learning performance of the proposed model, the followings have been drawn.

- The proposed developmental learning model enables a robot to acquire the ability of joint attention, which is adequate to achieve the task in a real environment.

- The caregiver's development can accelerate the learning of joint attention by changing the criterion for the task evaluation according to the learning performance of the robot. In addition, the caregiver's development triggered by the learning progress has a larger effect on the acceleration than the development triggered by a given clock.

- The robot's development caused by the learning progress can improve the final task performance of the robot's joint attention since it makes the robot acquire downsized and well-organized internal representation by abstracting the inputs by itself.

It is summarized that both of the developmental mechanisms, the robot's development and the caregiver's development, in the proposed model can facilitate the learning of the robot. This result demonstrates the knowledge that developments can help learning, and gives interesting suggestions that the development should be triggered by the learning progress and that the development can downsize the internal representation of the neural network. It is expected that the proposed model could be a help to understand human infants' development and learning.

As future work, the following issues should be addressed.

- *How does the robot represent the ability of joint attention in the neural network?*
  The internal representation in the acquired neural network has been already examined. As a result, it was revealed that the neural network included several unutilized neurons. However, it is not clear how utilized neurons represent the ability of joint attention. The neurons utilized in the task are conjectured to have some kind of response selectivity. The response selectivity means that a neuron responses only a certain feature in input data. In other experiments related to this work, it has been confirmed that the visual cortex neurons show the response selectivity to each of the horizontal change of the caregiver's face in input images or the vertical one. It should be investigated how the acquired neural network represents the ability of joint attention in its internal mechanisms.

- *How is the internal representation in the neural network acquired through learning?*
  It should be analyzed how the internal representation such as response selectivity in the neural network is acquired through learning. Several preliminary experiments have indicated that the response selectivity in the neural network is related to the robot's development. At the beginning of learning, blurred input images make the robot enhance the horizontal change of the caregiver's face in the input images and consequently learn the horizontal change in first. Then, as the visual mechanism develops, the robot becomes to learn the vertical change because the input images become clear. This process is conjectured to enable the robot to acquire response selectivity in the neural network. The learning and the developmental change in the neural network should be examined in more detail.

- *Why do the learning models including the robot's development show an U-shape change?*
  The experimental result shown in Figure 4.7 indicates that only the models with the robot's development, i.e. *RC-dev. model* and *R-dev. model*, show temporary retrogression in the learning curves. Such retrogression is called U-shaped change [Taga, 2002] in the research fields of cognitive developmental science.

> It is well known in cognitive developmental science that human infants show U-shaped changes in their motion and their recognition through their developmental processes. It should be analyzed why an U-shaped change is generated by the robot's development and what qualitative changes occur in the neural network. This is also considered to be related to the process of the acquisition of response selectivity; therefore, these issues described here should be discussed interactively.

Through addressing the above issues, it is expected to understand the mechanisms of developmental and learning in human infants more clearly.

The next chapter describes a bootstrap learning model for joint attention, which examines the knowledge that human infants have potentials to acquire new abilities based on only their innate capabilities without any evaluation from their caregivers.

Chapter 5

Bootstrap Learning based on Robot's Embedded Mechanisms

This chapter describes a bootstrap learning model for joint attention. Bootstrap learning means that an agent independently learns and acquires new abilities through interactions with an environment based on its innate capabilities or pre-acquired ones without any teaching, any external evaluation, or any structured environment. The proposed model is based on the insight that human infants do not always require caregivers' teaching or evaluation to learn joint attention. Furthermore, infants seem to inherently have various capabilities that allow them to learn joint attention.

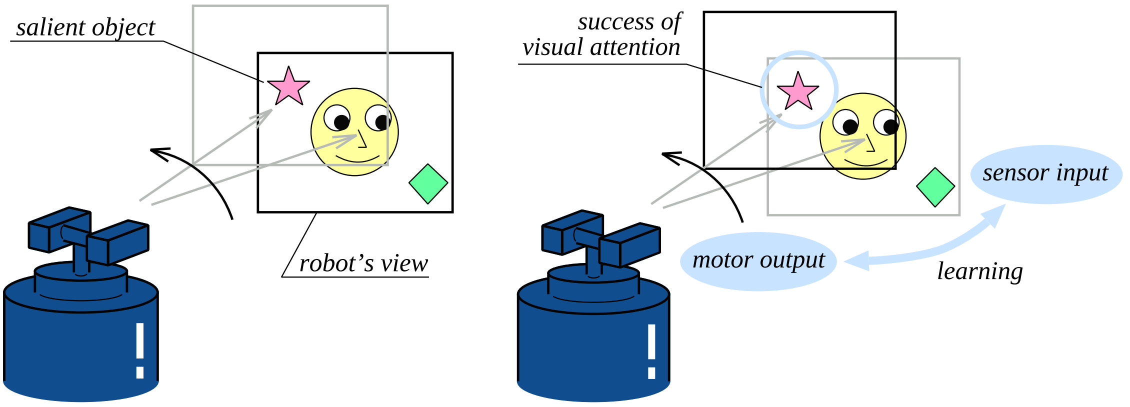
In this chapter, the mechanism of the proposed model is described first. The model is based on robot's embedded capabilities: visual attention that means to look at a salient object in the robot's view and learning with self-evaluation that means to learn its sensorimotor coordination when visual attention has succeeded. These two capabilities enable a robot to acquire the ability of joint attention through bootstrap learning. Next, the mechanism of bootstrap learning is explained. A mathematical proof shows how a robot acquires the sensorimotor coordination for joint attention based on experiences of visual attention. Then, the staged learning process of robot's joint attention, which is produced by the proposed model, is described in contrast to the staged development of infants' joint attention. Finally, some experimental results show the validity of the proposed model, and discussion is given.

5.1 Bootstrap Learning Model for Joint Attention

Human infants seem to inherently have various capabilities which allow them to interact with environments and to acquire new abilities, such as the ability of joint attention. The innate capabilities include preferences for salient visual stimuli, exploring environments, contingency learning, and so on [Bremner, 1994]. We suggest that infants can acquire the ability of joint attention with their caregivers through bootstrap learning based on their innate capabilities.

The mechanism of the proposed bootstrap learning model for joint attention is shown in Figure 5.2. A robot obtains its camera image $\boldsymbol{I}$ and the angle $\boldsymbol{\theta} = [\theta_{pan},\ \theta_{tilt}]$ of its camera head as inputs, and outputs a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ to rotate the camera head under the experimental setup shown in Figure 3.1. In the model, a robot has the following mechanisms:

(a) *visual attention*, which consists of *a salient feature detector* and *a visual feedback controller*, and has the capabilities to detect and gaze at a salient object in the robot's current view (see Figure 5.1 (a)), and



(a) Visual attention: to detect and gaze at a salient object in the robot's view.

(b) Learning with self-evaluation: to judge the success of visual attention and then learn the sensorimotor coordination.

Figure 5.1: The robot's embedded capabilities: visual attention and learning with self-evaluation, for the learning of joint attention.

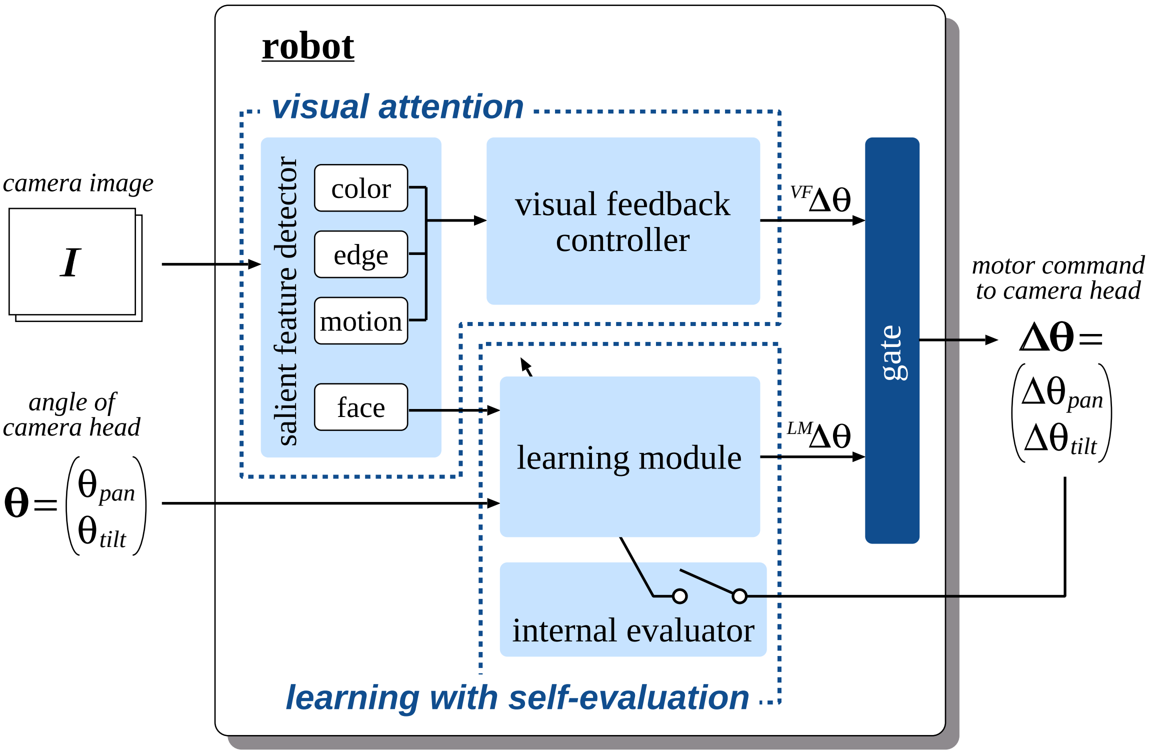

robot
visual attention
salient feature detector
color
edge
motion
face
visual feedback controller
$^{VF}\Delta\theta$
gate
learning module
$^{LM}\Delta\theta$
internal evaluator
learning with self-evaluation
camera image
I
angle of camera head
$\theta = \begin{pmatrix}\theta_{pan} \\ \theta_{tilt}\end{pmatrix}$
motor command to camera head
$\Delta\theta = \begin{pmatrix}\Delta\theta_{pan} \\ \Delta\theta_{tilt}\end{pmatrix}$

Figure 5.2: The proposed bootstrap learning model for joint attention. A robot obtains its camera image $\boldsymbol{I}$ and the angle of its camera head $\boldsymbol{\theta}$ as inputs, and outputs a motor command $\boldsymbol{\Delta\theta}$ to rotate the camera head. The model includes the mechanisms of visual attention, which consists of a salient feature detector and a visual feedback controller, and learning with self-evaluation, which consists of a learning module and an internal evaluator. The former generates an output to look at a salient object in the robot's view, and the latter learns sensorimotor coordination when visual attention has succeeded. The gate module makes a choice between the output $^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller and the output $^{LM}\boldsymbol{\Delta\theta}$ from the learning module.

(b) *learning with self-evaluation*, which consists of *a learning module* and *an internal evaluator*, and has the capabilities to judge the success of visual attention and then to learn the sensorimotor coordination (see Figure 5.1 (b)).

Both of the mechanisms of visual attention and learning with self-evaluation generate a motor command ${}^{VF}\boldsymbol{\Delta\theta}$ or ${}^{LM}\boldsymbol{\Delta\theta}$, respectively. The generated two are arbitrated in the following module:

(c) *a gate module*, which makes a choice between an output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller and an output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module according to a selecting rate.

The capability of visual attention means that a robot has preferences for visual salient stimuli, e.g. bright colors, rich patterns, and motions, and is able to gaze at them by rotating its camera head. On the other hand, the capability of learning with self-evaluation means that a robot has an ability to detect and learn an experience of visual attention since it provides the robot some sort of good feedback. In other words, a robot has the capability of contingency learning. The capability to arbitrate motor outputs in the gate module means that a robot can appropriately utilize its own mechanisms based on the validity of them.

Based on the above mechanisms, a robot learns the sensorimotor coordination for joint attention through the following procedure.

1. A robot first looks at a caregiver who is looking at an object (the situation shown in Figure 3.1 (a)) and obtains its camera image $\boldsymbol{I}$ and the angle $\boldsymbol{\theta}$ of its camera head as sensor inputs.

2. If a salient object is observed in $\boldsymbol{I}$, the robot detects the object by the salient feature detector and then generates a motor command ${}^{VF}\boldsymbol{\Delta\theta}$ to look at the object by the visual feedback controller.

3. At the same time, the robot generates another motor command ${}^{LM}\boldsymbol{\Delta\theta}$ by the learning module based on the inputs of the caregiver's face image, which has been detected by the salient feature detector, and the angle $\boldsymbol{\theta}$ of the camera head.

4. The gate makes a choice between ${}^{VF}\boldsymbol{\Delta\theta}$ and ${}^{LM}\boldsymbol{\Delta\theta}$ according to a selecting rate that is designed to mainly select the former one in the early stage of learning and to gradually become to select the latter one as learning advances. The robot outputs the selected motor command as $\boldsymbol{\Delta\theta}$ (= ${}^{VF}\boldsymbol{\Delta\theta}$ or ${}^{LM}\boldsymbol{\Delta\theta}$).

5. After the motor output, if an object is observed in the center of the robot's camera image (the situation shown in Figure 3.1 (b)), the robot judges the success of visual attention by the internal evaluator and then triggers the processing of back-propagation learning (BP processing) in the learning module.

6. The robot learns the sensorimotor coordination in the learning module by back-propagation using the output $\boldsymbol{\Delta\theta}$ as a reference.

7. Return to 1.

The following sections explain the modules in the proposed model: the salient feature detector, the visual feedback controller, the internal evaluator, the learning module, and the gate. In the section of the learning module, the mechanism that enables a robot to acquire the ability of joint attention based on visual attention is explained.

5.1.1 Salient Feature Detector

The salient feature detector extracts distinguishing image areas from $\boldsymbol{I}$ by color, edge, motion, and face detectors. Figure 5.3 shows the processing flow of the salient feature detector. First, salient objects which are observed in the field of the robot's view are detected by color, edge, and motion detectors. The color detector extracts image areas with a bright color, e.g. red, yellow, pink, and so on, which are defined as values in certain areas in YUV color space in advance. The edge detector extracts image areas with rich patterns by a Laplacian spatial filter. The motion detector extracts image areas with time differences between contiguous image frames. Each of the detected image areas is labeled as a salient object i $(= 1, \ldots, n)$. Then, the salient feature detector selects the most interesting object i_{trg} as a target to be gazed at by comparing the sum of the interests of all features:

$$i_{trg} = \arg\max_i (\alpha_c f_i^{col} + \alpha_e f_i^{edg} + \alpha_m f_i^{mot}), \tag{5.1}$$

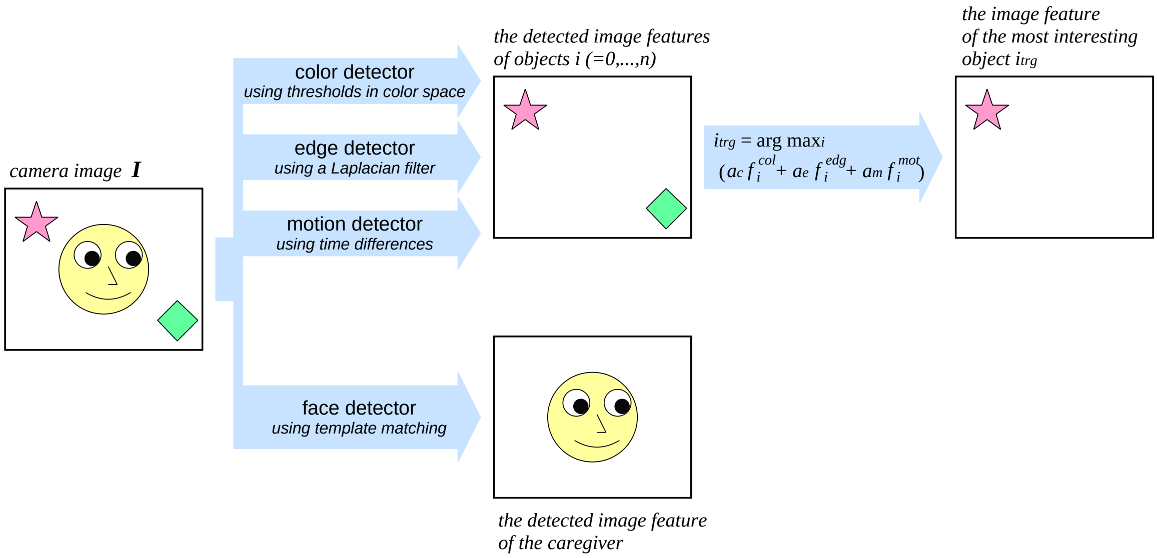


Figure 5.3: The salient feature detector. Primitive features of objects are detected from a camera image $\boldsymbol{I}$ by color, edge, and motion detectors, and then the most interesting object i_{trg} is selected by comparing the sum of the interests of all features. At the same time, a face-like stimulus of the caregiver is extracted by face detector using template matching. The detected primitive feature of the object i_{trg} and the face-like one of the caregiver are sent to the visual feedback controller and the learning module, respectively.

where f_i^{col}, f_i^{edg}, and f_i^{mot} indicate the size of the colored area, the complexity of the pattern, and the amount of the motion of the object i, respectively. The coefficients $(\alpha_c,\ \alpha_e,\ \alpha_m)$ denote the degrees of the interests in three features, which are determined according to the context or the preferences of the robot. This mechanism makes the robot randomly change the object to be gazed at in every trial.

At the same time as the detection of salient objects, the face detector extracts a face-like stimulus of the caregiver by template matching. A template image of the caregiver is given to a robot in advance. The detection of face-like stimuli is a fundamental ability for social agents; therefore, it should be treated in the same manner as the detection of the primitive features of objects. Finally, the detected feature of the object i_{trg} and the face-like one of the caregiver are sent to the visual feedback controller and the learning module, respectively.

5.1.2 Visual Feedback Controller

The visual feedback controller receives the detected image feature of the object i_{trg} and then generates a motor command ${}^{VF}\boldsymbol{\Delta\theta}$ for the camera head to gaze at the object. Figure 5.4 shows the mechanism of the visual feedback controller. First, the controller calculates the position $(x_{i_{trg}},\ y_{i_{trg}})$ of the object i_{trg} in the camera image. Then, it generates a motor command ${}^{VF}\boldsymbol{\Delta\theta}$ as

$$ {}^{VF}\boldsymbol{\Delta\theta} = \begin{pmatrix} {}^{VF}\Delta\theta_{pan} \\ {}^{VF}\Delta\theta_{tilt} \end{pmatrix} = g \begin{pmatrix} x_{i_{trg}} - cx \\ y_{i_{trg}} - cy \end{pmatrix}, \tag{5.2} $$

where g and $(cx,\ cy)$ denote a scalar gain and the center position of the camera image. The scalar gain g is defined by a designer in advance. The generated motor command ${}^{VF}\boldsymbol{\Delta\theta}$ is sent to the gate module as a possible of the robot's motor command.

As described here, visual attention, which is one of the robot's embedded mechanisms, is realized by the salient feature detector and the visual feedback controller.

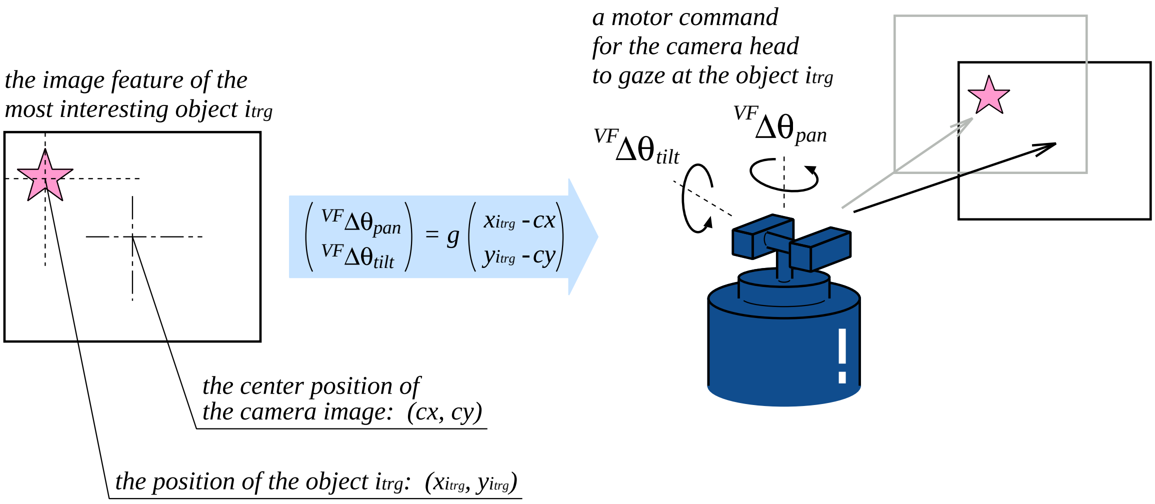

the image feature of the
most interesting object itrg
a motor command
for the camera head
to gaze at the object itrg
VFΔθpan
VFΔθtilt
$\begin{pmatrix} {}^{VF}\Delta\theta_{pan} \\ {}^{VF}\Delta\theta_{tilt} \end{pmatrix} = g \begin{pmatrix} x_{itrg} - cx \\ y_{itrg} - cy \end{pmatrix}$
the center position of
the camera image: (cx, cy)
the position of the object itrg: (xitrg, yitrg)

Figure 5.4: The visual feedback controller. The controller generates a motor command ${}^{VF}\boldsymbol{\Delta\theta}$ to gaze at the most interesting object i_{trg} based on the detected image feature of the object. The generated motor command is sent to the gate module.

5.1.3 Internal Evaluator

The other embedded mechanism, learning with self-evaluation, is realized by the internal evaluator and the learning module.

The internal evaluator triggers the BP processing in the learning module when visual attention has succeeded. Figure 5.5 shows the mechanism of the internal evaluator. First, the evaluator judges the success of visual attention when

$$\sqrt{(x_i - cx)^2 + (y_i - cy)^2} < d_{th}, \tag{5.3}$$

where $(x_i,\ y_i)$ and $(cx,\ cy)$ denote the position of an object i and the center position of the camera image, and d_{th} is a threshold to judge whether the robot is looking at the object in the center of the camera image or not. If an object is observed in the center of the camera image, that is, when visual attention has succeeded, the internal evaluator triggers the BP processing in the learning module. Note that the success of *visual attention* does not always mean the success of *joint attention*. In other words, the object that the robot has looked at does not always coincide with the object that the caregiver is looking at in an environment including multiple salient objects. Furthermore, the robot cannot recognize whether the object that the robot is looking

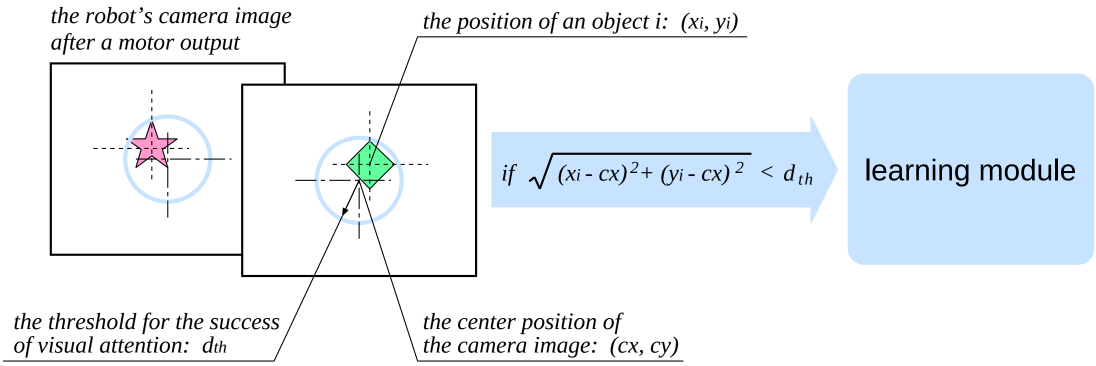
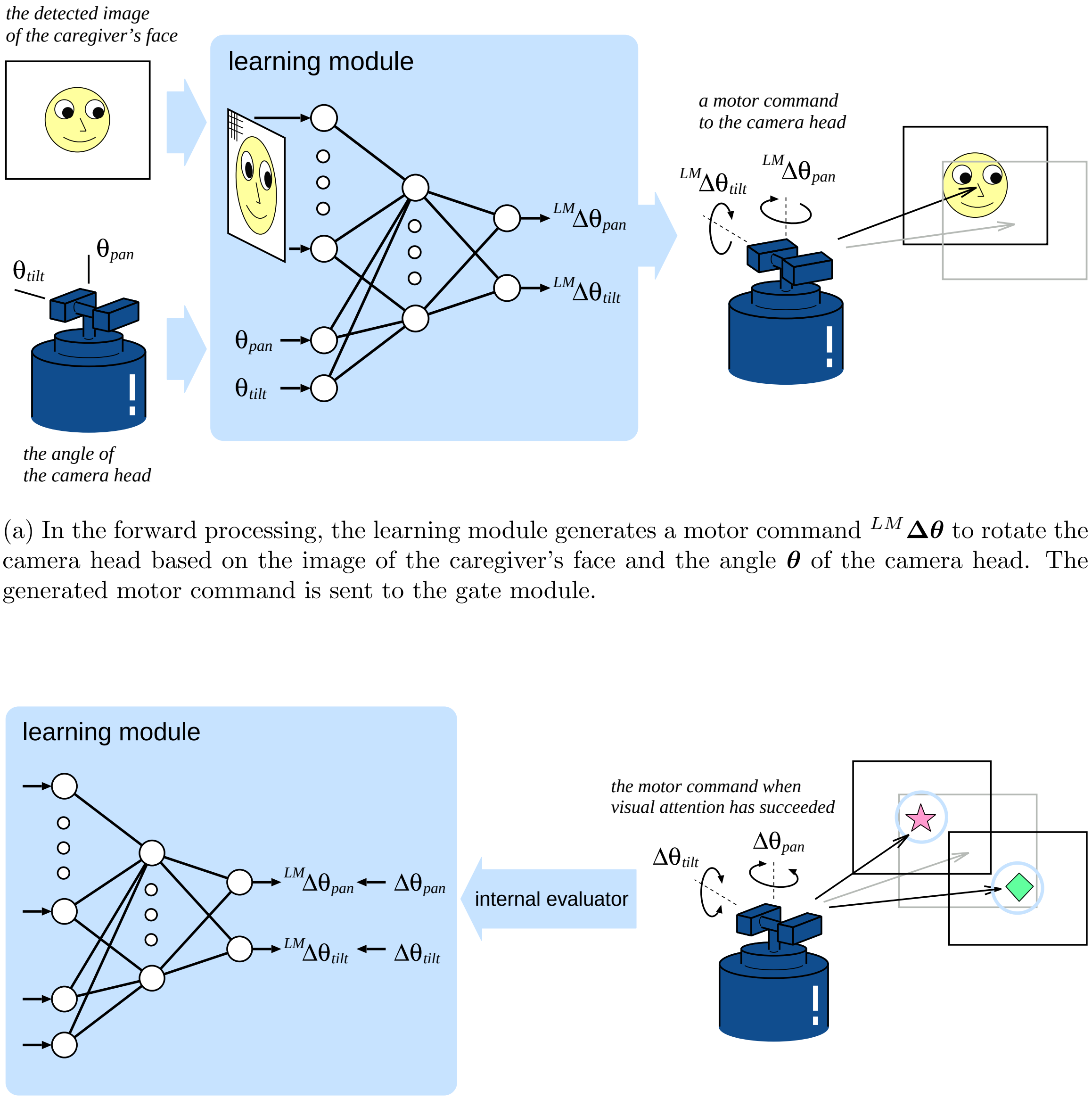

the robot's camera image
after a motor output
the position of an object i: (xi, yi)
if $\sqrt{(x_i - cx)^2 + (y_i - cx)^2} < d_{th}$
learning module
the threshold for the success
of visual attention: dth
the center position of
the camera image: (cx, cy)

Figure 5.5: The internal evaluator. First, the evaluator judges the success of visual attention by measuring the distance between the position of an object i and the center position of the camera image. If an object is observed in the center of the camera image, which means the success of visual attention, the evaluator triggers the BP processing in the learning module.

at coincides with the object that the caregiver is looking at. Therefore, the robot learns not only *correct* data for joint attention but also *incorrect* ones in the learning module.

5.1.4 Learning Module

The learning module consists of a three-layered neural network. Figure 5.6 shows the mechanism of the learning module, in which (a) and (b) present the forward processing and the BP processing. In the forward processing, the module receives the image of the caregiver's face detected by the salient feature detector and the angle $\boldsymbol{\theta}$ of the robot's camera head as inputs, and outputs a motor command ${}^{LM}\boldsymbol{\Delta\theta}$ to rotate the camera head. The caregiver's face image, which is input to the neural network as the values of the brightness of all pixels, is utilized to estimate the motor command ${}^{LM}\boldsymbol{\Delta\theta}$ to follow the direction of the caregiver's gaze. On the other hand, the angle $\boldsymbol{\theta}$ of the robot's camera head is used to generate a motor command ${}^{LM}\boldsymbol{\Delta\theta}$ incrementally and to enable it to rotate nonlinearly. The reasons for the increment and the nonlinearity are illustrated in Figure 5.7. In the figure, the caregiver is looking one direction on which three objects are placed. In this case, the robot cannot discriminate which object (i), (ii), or (iii) the caregiver is looking at because the images of the caregiver's face when he/she is looking at each object are almost the same. Therefore, the robot generates its motor commands step by step, and identifies the object that is first detected in the camera image as the object that the caregiver is looking at. At the same time, the motor command to follow the direction of the caregiver's gaze, which is linear in the image space, could be nonlinear in the motor space. The reason is that the rotational center of the camera head does not coincide with the optical center of the each camera. In such case, the robot is expected to gradually shift its motor command by using the angle $\boldsymbol{\theta}$. For these purposes, the robot has the capability to generate its motor command incrementally and nonlinearly by using $\boldsymbol{\theta}$. In the forward processing, the generated motor command ${}^{LM}\boldsymbol{\Delta\theta}$ is sent to the gate module as another possible of the output.



(a) In the forward processing, the learning module generates a motor command $^{LM}\boldsymbol{\Delta\theta}$ to rotate the camera head based on the image of the caregiver's face and the angle $\boldsymbol{\theta}$ of the camera head. The generated motor command is sent to the gate module.

(b) In the BP processing, the module learns the sensorimotor coordination between the inputs, the image of the caregiver's face and the angle $\boldsymbol{\theta}$, and the output $\boldsymbol{\Delta\theta}$ by back-propagation. The internal evaluator triggers this processing when visual attention has succeeded and passes the output $\boldsymbol{\Delta\theta}$ at the time as a reference for learning.

Figure 5.6: The learning module, which consists of a three-layered neural network and conducts two processing: the forward processing to generate a motor command $^{LM}\boldsymbol{\Delta\theta}$ and the BP processing to learn its sensorimotor coordination.

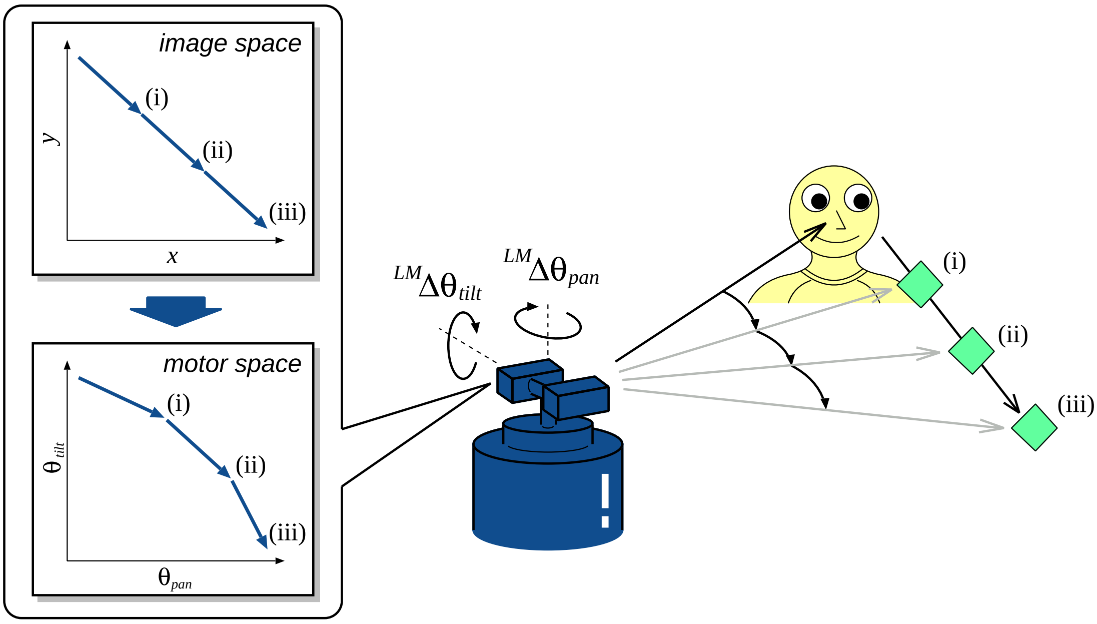


Figure 5.7: The reasons why the motor command $^{LM}\boldsymbol{\Delta\theta}$ is generated incrementally and nonlinearly. The reason for the increment is that the caregiver's attention cannot be narrowed down to a particular point along the line of the caregiver's gaze because of the resolution of the robot's camera image. On the other hand, the reason for the nonlinearity is that the motor command to follow the line of the caregiver's gaze becomes a curve in the motor space because the rotational center of the robot's camera head does not coincide with the optical center of the each camera.

In the BP processing, when the learning module is triggered by the internal evaluator, the module learns the sensorimotor coordination between the inputs, the caregiver's face image and the angle $\boldsymbol{\theta}$ of the camera head, and the output $^{VF}\boldsymbol{\Delta\theta}$ by back-propagation. The learning module utilizes the motor command $\boldsymbol{\Delta\theta}$ at that time as a reference for back-propagation. As mentioned above, the internal evaluator triggers the BP processing when visual attention has succeeded; hence, the learning module receives not only correct learning data for joint attention but also incorrect ones. However, an unstructured environment, in which the positions of objects change randomly, enables the module to acquire the sensorimotor coordination for joint attention based on the following mechanism.

- In the case that the robot has learned *incorrect* data, that is, when joint attention has failed while visual attention has succeeded, the acquired sensorimotor coordination is expected to be relatively weakened compared to that when joint attention has succeeded. The reason is that the position of the object that the robot has looked at does not uniquely correspond to the image of the caregiver's face. In other words, the sensorimotor coordination when joint attention has failed does not have any correlation between the inputs and the output.

- In the case that the robot has learned *correct* data, that is, when joint attention has succeeded, the acquired sensorimotor coordination is expected to be relatively enhanced compared to that when joint attention has failed. The reason is that the position of the object that the robot as well as the caregiver are looking at is uniquely determined by the image of the caregiver's face. In other words, the sensorimotor coordination when joint attention has succeeded has a certain correlation between the sensor inputs and the motor output.

As a result, the enhanced sensorimotor coordination with a correlation, which has been learned when joint attention has succeeded, allows the robot to acquire the ability of joint attention. This mechanism of bootstrap learning is explained in Appendix B with a mathematical proof and an example using a simple environmental setup. The mathematical proof gives some conditions which are required to bootstrap learning. The example shows the process to acquire the sensorimotor coordination for joint attention step by step.

5.1.5 Gate

The gate module arbitrates a motor command $\boldsymbol{\Delta\theta}$ between ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller and ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module. Figure 5.8 shows the mechanism of the gate module. This module sets a gating function to define the selecting rate of the outputs. In the early stage of learning, the selecting rate of ${}^{VF}\boldsymbol{\Delta\theta}$ is set to a higher probability than that of ${}^{LM}\boldsymbol{\Delta\theta}$ because the learning module has not acquired the appropriate sensorimotor coordination for joint attention yet. On the other hand, in the later stage of learning, the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module, which has

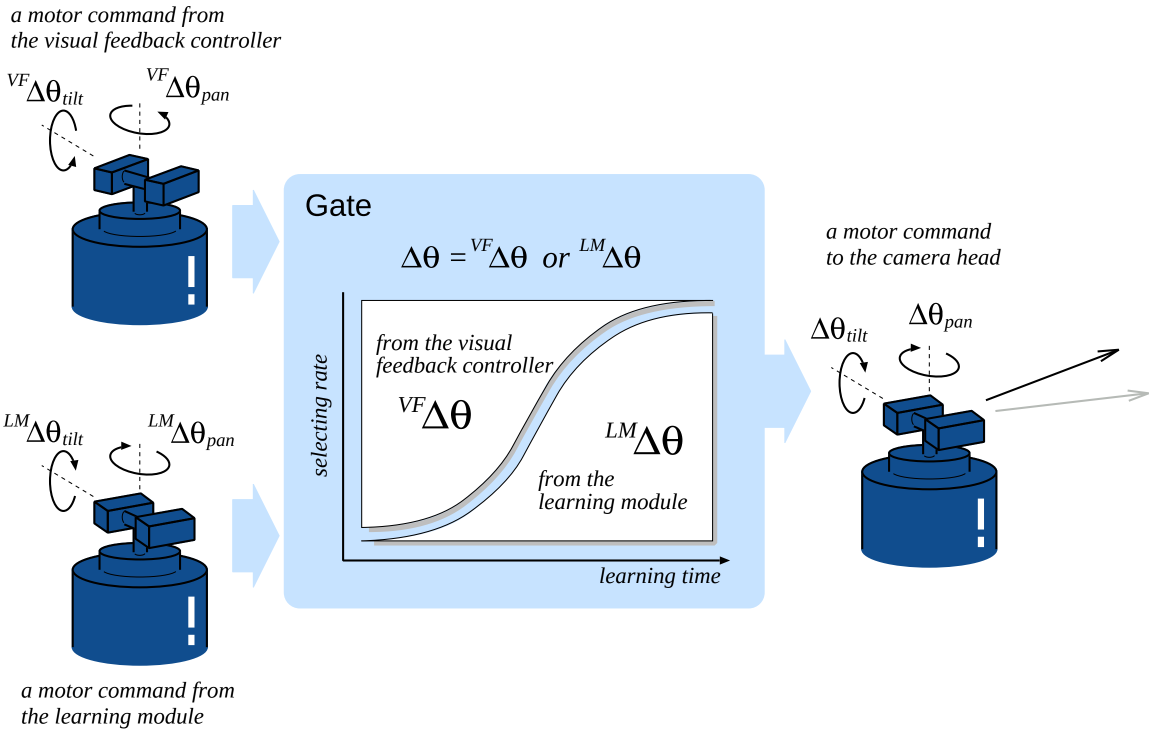


Figure 5.8: The gate module. The module makes a choice between $^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller and $^{LM}\boldsymbol{\Delta\theta}$ from the learning module according to a selecting rate, and outputs the selected motor command as $\boldsymbol{\Delta\theta}$. The selecting rate is designed so that $^{VF}\boldsymbol{\Delta\theta}$ is mainly selected in the early stage of learning, and $^{VF}\boldsymbol{\Delta\theta}$ becomes more probable to be selected as learning advances.

acquired the sensorimotor coordination for joint attention, becomes more probable to be selected. The gate module makes a choice between $^{VF}\boldsymbol{\Delta\theta}$ and $^{LM}\boldsymbol{\Delta\theta}$ according to the selecting rate, and outputs the selected one as the robot's motor command $\boldsymbol{\Delta\theta}$ ($= {}^{VF}\boldsymbol{\Delta\theta}$ or $^{LM}\boldsymbol{\Delta\theta}$).

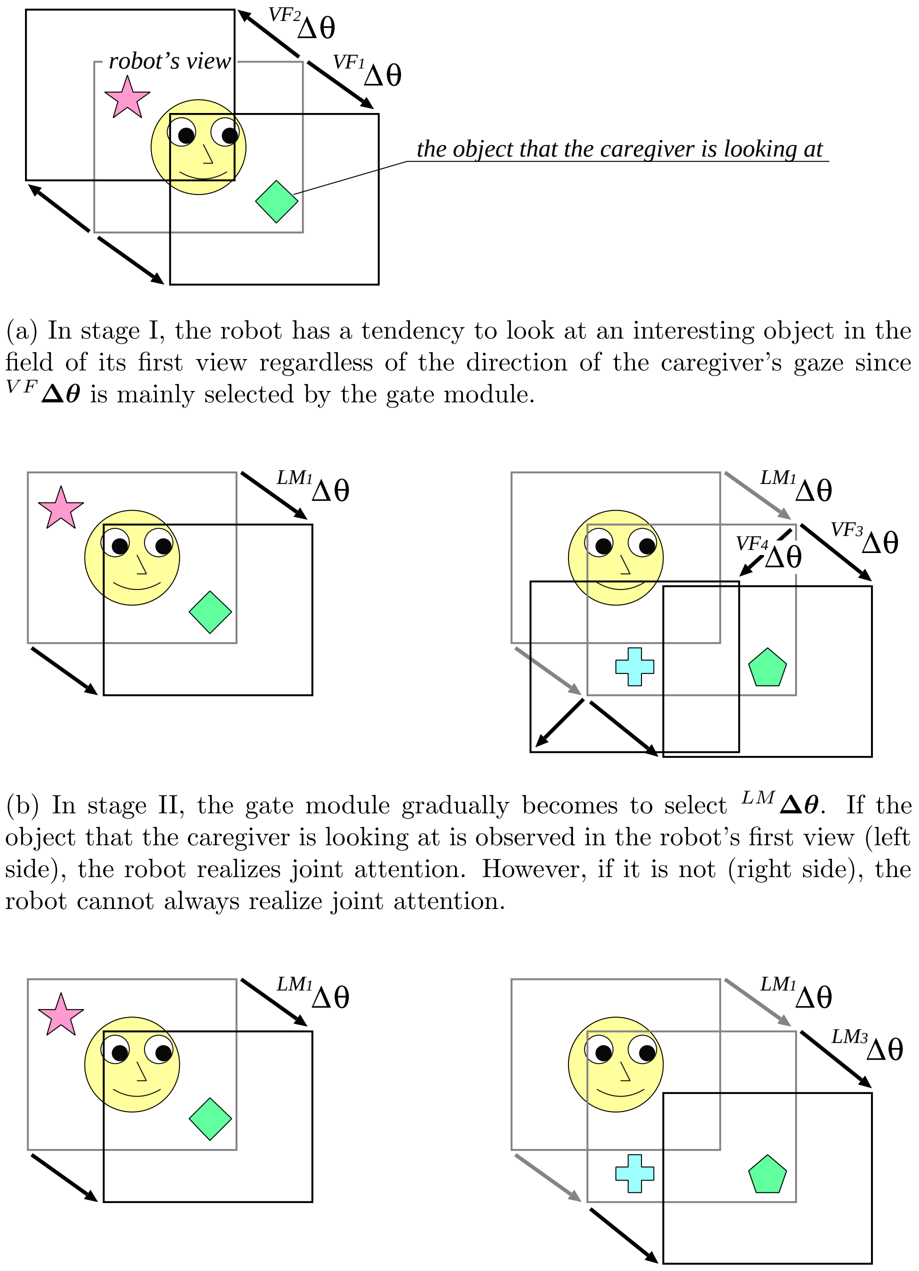
This gate module enables the robot to increase the proportion of correct learning data as learning advances and consequently to acquire more appropriate sensorimotor coordination for joint attention in the learning module. At the same time, the shift of the attention mechanism from visual attention based on $^{VF}\boldsymbol{\Delta\theta}$ to joint attention based on $^{LM}\boldsymbol{\Delta\theta}$ is expected to make the robot reproduce the staged learning process of joint attention, which is similar to the developmental process of infants'. The experiments use a sigmoid function for the selecting rate that is defined by a designer in advance.

5.2 Staged Learning of Joint Attention

It is expected that the proposed model makes the robot acquire the ability of joint attention through a staged learning process. Figure 5.9 represents the transition of the robot's behavior through three stages. In each stage, the behavior of the robot is represented as the change of its camera image when it shifts the gaze direction from the caregiver's face to an object based on the output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller or the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module. In the figure, a rectangle indicates a camera image of the robot, and arrows which connect the corners of two rectangles show a motor output of the robot.

stage I: In the first stage of learning, the robot has a tendency to look at an interesting object in the field of the robot's view based on the embedded mechanism of visual attention. As shown in Figure 5.9 (a), if two objects are observed in the robot's first view, the robot outputs ${}^{VF_1}\boldsymbol{\Delta\theta}$ or ${}^{VF_2}\boldsymbol{\Delta\theta}$ case by case since the gate module mainly selects the output from the visual feedback controller. Thus, the robot in this stage looks at an interesting object in its view regardless of the direction of the caregiver's gaze. It means that joint attention succeeds only at a chance level. At the same time, the robot starts to learn its sensorimotor coordination based on self-evaluation on visual attention.

stage II: In the middle stage of learning, the robot becomes to realize joint attention only when the object that the caregiver is looking at is observed in the field of the robot's first view. As shown in the left of Figure 5.9 (b), the robot is able to look at the same object that the caregiver is looking at by generating the output ${}^{LM_1}\boldsymbol{\Delta\theta}$ from the learning module. The gate module gradually becomes to select the output from the learning module as learning advances. The sensorimotor coordination of ${}^{LM_1}\boldsymbol{\Delta\theta}$ has been acquired through learning in stage I because that of ${}^{VF_1}\boldsymbol{\Delta\theta}$ had a correlation.



(a) In stage I, the robot has a tendency to look at an interesting object in the field of its first view regardless of the direction of the caregiver's gaze since ${}^{VF}\boldsymbol{\Delta\theta}$ is mainly selected by the gate module.

(b) In stage II, the gate module gradually becomes to select ${}^{LM}\boldsymbol{\Delta\theta}$. If the object that the caregiver is looking at is observed in the robot's first view (left side), the robot realizes joint attention. However, if it is not (right side), the robot cannot always realize joint attention.

(c) In stage III, the robot is able to realize joint attention by ${}^{LM}\boldsymbol{\Delta\theta}$ regardless of whether the object that the caregiver is looking at is observed in the field of the robot's first view or not.

Figure 5.9: The staged learning process of the robot's joint attention, in which the robot's behaviors are shown as the change of its camera image. The robot looks at an object by the output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller or the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module. This staged learning process can be regarded as equivalent to the stage developmental process of infants' joint attention shown in Figure 2.2.

By contrast, if the object that the caregiver is looking at is outside the field of the robot's first view, the robot cannot always realize joint attention. As shown in the right of Figure 5.9 (b), the robot finds the object not at the center but at the periphery of its view by generating the output ${}^{LM_1}\boldsymbol{\Delta\theta}$. Then, if several objects are observed in the camera image, the robot outputs ${}^{VF_3}\boldsymbol{\Delta\theta}$ or ${}^{VF_4}\boldsymbol{\Delta\theta}$ from the visual feedback controller and looks at an interesting object case by case. The success rate of joint attention in this stage becomes better than that in stage I; however, joint attention is realized mainly in the field of the robot's first view. As well as in stage I, the robot learns its sensorimotor coordination when visual attention has succeeded.

stage III: In the final stage of learning, the robot has acquired the complete ability of joint attention owing to the learning in stages I and II. As shown in Figure 5.9 (c), the robot can identify the object that the caregiver is looking at by incrementally generating the outputs ${}^{LM_1}\boldsymbol{\Delta\theta}$ and ${}^{LM_3}\boldsymbol{\Delta\theta}$ from the learning module even if the object is not observed in the field of the robot's first view. The gate module in this stage mainly selects the output from the learning module. The sensorimotor coordinations of ${}^{LM_1}\boldsymbol{\Delta\theta}$ and ${}^{LM_3}\boldsymbol{\Delta\theta}$ has been acquired through the learning in stages I and II because ${}^{VF_1}\boldsymbol{\Delta\theta}$ and ${}^{VF_3}\boldsymbol{\Delta\theta}$ had a correlation in their sensorimotor coordination. The robot in this stage realizes joint attention at high performance.

Through the above learning process, the robot acquires the ability of joint attention based on the bootstrap learning model. It is considered that the staged learning process of the robot's joint attention can be regarded as equivalent to the staged developmental process of infants' shown in Figure 2.2. The stages I, II, and III of the robot respectively correspond to infants at 6-9, 12, and 18 months old. In addition, it is supposed in cognitive developmental science that the embedded mechanisms of the robot, visual attention and learning with self-evaluation, are also inherent in infants [Bremner, 1994]. The similarities of the developmental phenomena and the embedded mechanisms between the robot and infants make it possible to understand the developmental mechanisms of infants' joint attention.

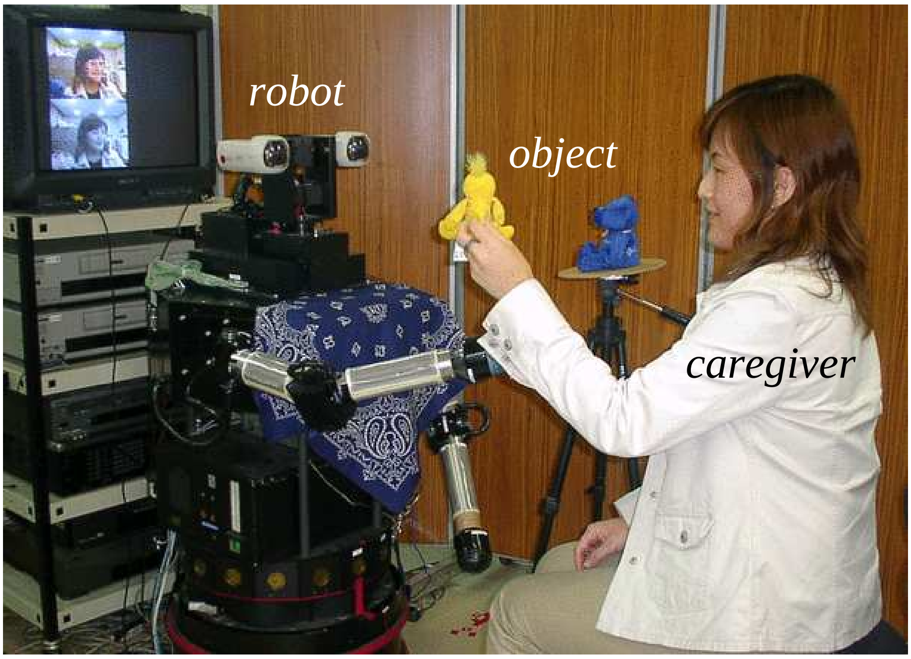
5.3 Experimental Setup

It is examined whether an actual robot can acquire the ability of joint attention based on the proposed bootstrap learning model without any task evaluation from a human caregiver in an unstructured environment including multiple objects. An experimental environment is shown in Figure 5.10 (a), and the robot's camera image in this situation is shown in (b). In the environment, several objects with a bright color are randomly placed around the robot and the caregiver. The caregiver is looking at one object that is randomly selected in each trial. In Figure 5.10, she is looking at the object in her hand. The robot captures an input image through its cameras and detects the caregiver's face and the objects by the salient feature detector as shown in Figure 5.10 (b). In the experiments presented in this chapter, the robot uses only the left camera image. In the left image in (b), the rectangle shows the position of the caregiver's face detected by template matching. The highlighted areas in the right image show the objects with a bright color extracted by using color definitions given in advance. The detected image of the caregiver's face is input to the learning module, and that of the objects is sent to the visual feedback controller.

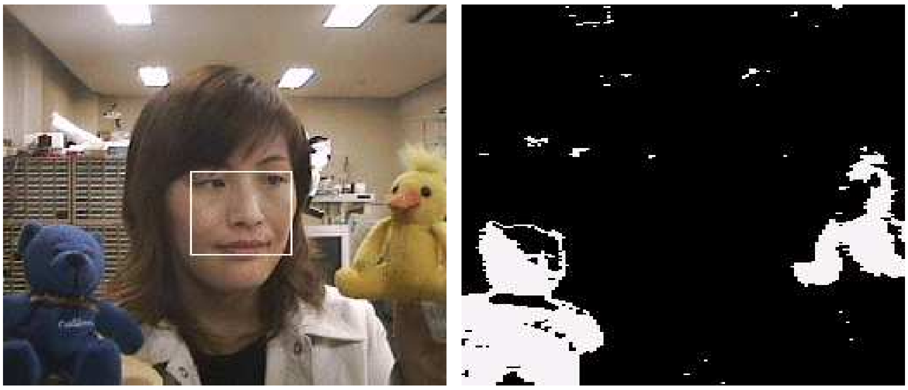
The experiment uses the following parameters. The degrees of the robot's interests in image features in Eq. (5.1) are defined as $(\alpha_c,\ \alpha_e,\ \alpha_m) = (1,\ 0,\ 0)$; in other words, the robot is designed to prefer to look at an object which has a bright color and a larger size in its camera image. Note that the caregiver does not know the preference of the robot. The threshold d_{th} in Eq. (5.3) for the evaluation of visual attention is defined as $d_{th} = W_x/6$, where W_x denotes the width of the robot's camera image. Under these conditions, 125 learning data sets are acquired in the real environment in advance, and then off-line learning is conducted. Each data set includes

- *input data:* a left camera image $\boldsymbol{I}$, in which the caregiver's face is extracted as a window of which size is 30×25 [pixel], and the angle of the camera head $\boldsymbol{\theta} = [\theta_{pan},\ \theta_{tilt}]$ when the robot is looking at the caregiver's face,

- *output data when joint attention has succeeded:* a motor command $\boldsymbol{\Delta\theta} = [\Delta\theta_{pan},\ \Delta\theta_{tilt}]$ for the camera head to shift the robot's gaze direction from the caregiver's face to the object that the caregiver is looking at, and



(a) An experimental environment for joint attention in which a robot with two cameras, a human caregiver, and multiple salient objects are shown. The objects are randomly placed in every trial, and the caregiver looks at one object that has been selected at random. The robot first looks at the caregiver and captures its camera image as shown in (b).



(b) The robot's camera image acquired in the situation (a). The rectangle in the left image shows the position of the caregiver's face detected by template matching, and the highlighted areas in the right show the objects with a bright color extracted by using thresholds in a color space. This processing is conducted by the salient feature detector in the proposed model.

Figure 5.10: An experimental setup for bootstrap learning of joint attention.

- *output data when joint attention has failed while visual attention has succeeded:* motor commands $\boldsymbol{\Delta\theta}$ to look at different objects from that the caregiver is looking at, only these data are obtained in a simulation.

The object that the caregiver looks at is placed at random positions in the pan range from -45 to 45 [deg] and in the tilt range from -30 to 30 [deg]. The robot learns its sensorimotor coordination in the learning module by back-propagation using the above input data and either of two kinds of output data. The number of units in the learning module are set 752 ($30 \times 25 + 2$) for the input units, five for the hidden units, and two for the output units. The number of the hidden units is determined as a minimum one that showed better performance in preliminary experiments.

5.4 Experimental Results

We evaluate the validity of the bootstrap learning model about the following performances:

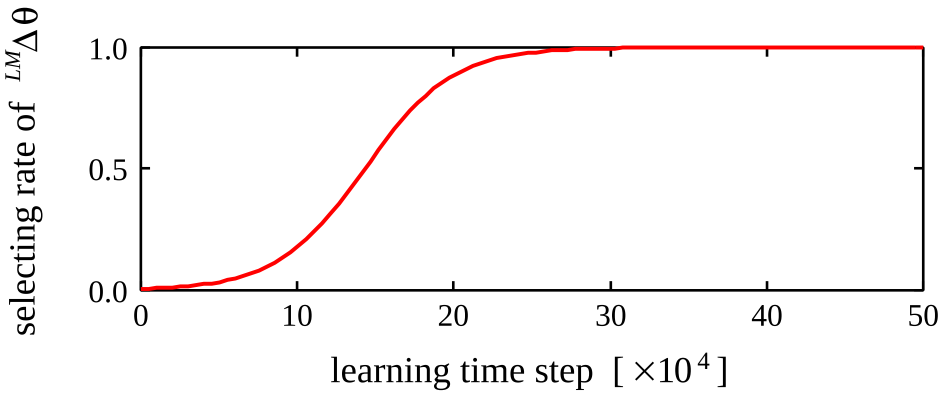
- the change of the success rate of joint attention depending on the number of objects,
- the similarity between the staged learning process of the robot's joint attention and that of infants',
- the effect of the gate module on learning,
- the acquisition of a correlation in the sensorimotor coordination of the learning module,
- the organization of the outputs from the learning module, and
- the final task performance of joint attention in a real environment.

The following sections describe the results of these experiments in order.

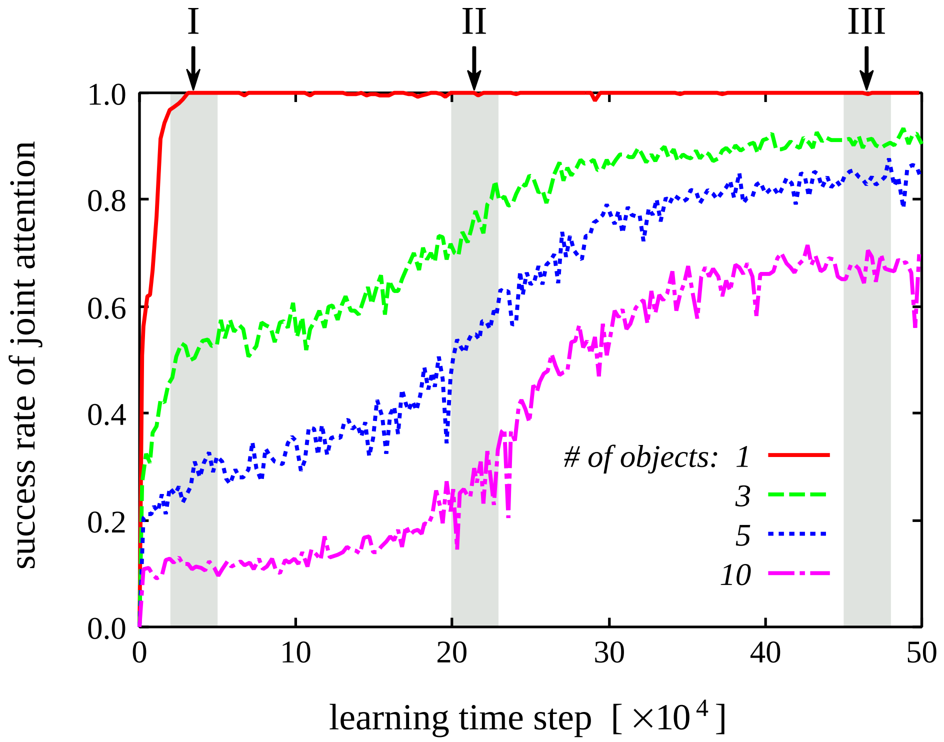
5.4.1 Evaluation of Task Performance

It is verified how the task performance of joint attention changes depending on the number of objects over learning. The gating function to select a motor output in the gate module is defined as a sigmoid one shown in Figure 5.11 (a). The horizontal axis denotes the learning time step, and the vertical one denotes the selecting rate of the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module. The output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller is selected at the residual rate. This gating function is designed based on preliminary experiments. Figure 5.11 (b) shows the changes of the success rates of joint attention over learning, in which the horizontal axis and the vertical one denote the learning time step and the success rate of joint attention, respectively. The four curves show the success rates when the number of objects are set to one, three, five, and ten. The case in which the number of the object equals to one means that the robot always looks at the same object that the caregiver is looking at based on visual attention and learns only correct sensorimotor coordination for joint attention. In contrast, the case of ten means that the robot receives correct learning data only at 1/10 proportion at the beginning of learning. However, the robot is expected to increase the proportion of correct data by adapting the output from the learning module that has already acquired sensorimotor coordination for joint attention.

From the result of Figure 5.11 (b), it can be found that the success rates of joint attention are at chance levels at the beginning of learning; however, they increase to high performance at the end although the environment includes multiple objects. Each of the success rates of joint attention starts from a chance level because the gate module mainly selects the output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller in the early stage of learning. However, as learning advances, the gate module increases the selecting rate of the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module that has acquired a correlation in the sensorimotor coordination, and consequently improves the performance of joint attention. In the case that the number of objects is set to five, the success rate of joint attention improves from 20%, which is just a chance level, to 85%. Therefore, it can be concluded that the proposed bootstrap learning model enables the robot to acquire the ability of joint attention without any task evaluation from the caregiver.



(a) The gating function used in the experiments. The output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module is selected at the rate of the value of this sigmoid function, and the output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller is selected at the residual rate.



(b) The changes of the success rates of joint attention over learning. Each curve shows the result when the number of objects is set to one, three, five, or ten. The success rates of all cases are at chance levels at the beginning of learning; however, they increase to high levels at the end.

Figure 5.11: The gating function for selecting a motor output in the gate module and the changes of the success rates of joint attention over learning.

5.4.2 Staged Learning Process

It is investigated how the robot changes its behavior through the learning process. We focus on the result when the number of objects is set to five in Figure 5.11 (b) and examine the robot's behavior in three stages I, II, and III, of which learning periods are 2-5, 20-23, and 45-48 [$\times 10^4$], respectively. Figure 5.12 shows the change of the pan angle of the robot's camera head when it has realized visual attention, in which "○" and "×" indicate the success of joint attention and the failure, respectively. In other words, the former means that the robot has looked at the same object that the caregiver is looking at while the latter means that the robot has looked at a different one. Note that objects exist at the areas that do not include any mark, and the failure of visual attention is not indicated. The data are recorded every 100 steps. The pan angle of the robot's camera head becomes 0 [deg] when the robot is looking at the caregiver, and the view range of the robot is ±18 [deg]. In other words, the objects found within ±18 [deg] are observed in the field of the robot's view when the robot is looking at the caregiver.

From this result, we can see that the number of the success of joint attention increases as learning advances; at the same time, the range of the camera angle when the robot has realized joint attention gradually exceeds the range of ±18 [deg]. This result shows that the learning process of the robot's joint attention based on the bootstrap learning model is similar to the developmental process of infants' joint attention shown in Figure 2.2. It seems that the robot's behaviors in stages I, II, and III are equivalent to the infants' behaviors at 6-9, 12, and 18 months old, respectively. In stage I, the robot has a tendency to look at an interesting object in the field of its first view just like infants at 6 to 9 months old. The reason why the robot seldom or never achieves visual attention or joint attention over ±18 [deg] is that the gate module mainly selects the output from the visual feedback controller based on the gating function shown in Figure 5.11 (a). In stage II, the robot becomes to realize joint attention mostly in the field of the robot's first view, i.e. within ±18 [deg]. The robot follows the caregiver's gaze and looks at the object that the caregiver is looking at only when the object is observed in the robot's first view. This behavior is similar to infants' at 12 months old. Finally, in stage III, the robot as well as infants at

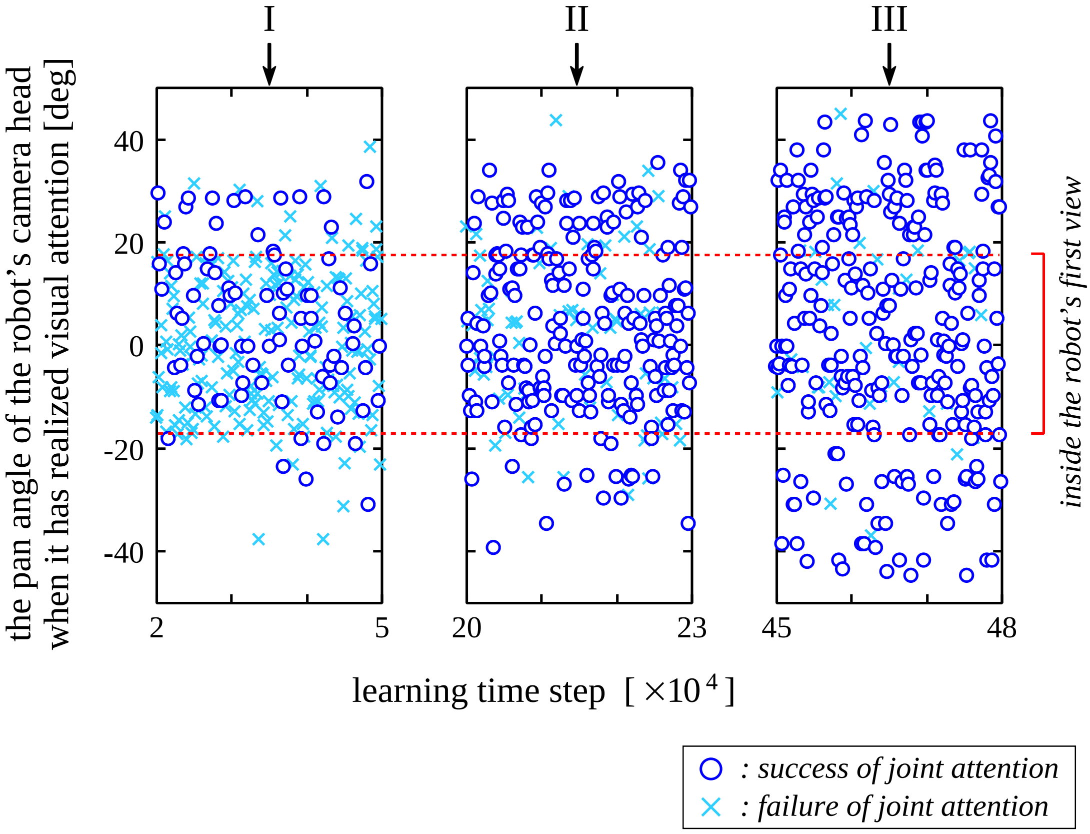


Figure 5.12: An experimental result of the staged learning process of joint attention. This graph plots the change of the pan angle of the robot's camera head when the robot has achieved visual attention, in which "○" and "×" denote the success of joint attention and the failure, respectively. In stage I, the robot has a tendency to look at an object inside the field of the robot's first view and realizes joint attention only at a chance level. In stage II, the robot becomes to realize joint attention within the robot's first view. Finally, in stage III, the robot realizes joint attention at every position. This learning process of the robot's joint attention is considered to be equivalent to the developmental process of infants' joint attention shown in Figure 2.2

18 months old can realize joint attention regardless of whether the object that the caregiver is looking at is within the field of the robot's or the infants' views or not. From the similarities described here, we may conclude that the proposed bootstrap learning model makes the robot reproduce the infants' development of joint attention.

5.4.3 Effect of Gate Module on Task Performance

The effectiveness of the gate module is verified. The experimental result shown in Figure 5.11 (b) when the gating function is defined as Figure 5.11 (a) is compared with the results when the gating functions are defined as constant values.

Figure 5.13 shows the changes of the success rates of joint attention when the selecting rate of ${}^{LM}\boldsymbol{\Delta\theta}$ is set to 0.7, 0.9, or 1.0 over learning. The result when the gate module uses the sigmoid function and the number of object is set to five is also presented in Figure 5.13. The horizontal axis and the vertical one respectively denote the learning time step and the success rate of joint attention. From the comparison of these results, it is confirmed that the gating function designed as a sigmoid one improves the task performance of joint attention. Only in this case, the success rate of joint attention overs 85 [%] at the end of learning. By contrast, when the selecting rate of ${}^{LM}\boldsymbol{\Delta\theta}$ is set to 1.0, the success rate of join attention cannot reach 30 [%]. The reason is considered that the learning module which has not acquired appropriate sensorimotor coordination for joint attention yet is utilized to perform visual attention in the early stage of learning, and consequently the learning data are biased to inadequate initial experiences. By the same reason, the task performances when the selecting rate is set to 0.7 and 0.9 cannot improve. From this experimental result, it can be concluded that the gating function should be designed so that the output ${}^{VF}\boldsymbol{\Delta\theta}$ from the visual feedback controller is selected at high proportion in the beginning of learning, and the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module gradually comes to be selected as learning advances as mentioned in Section 5.1.5.

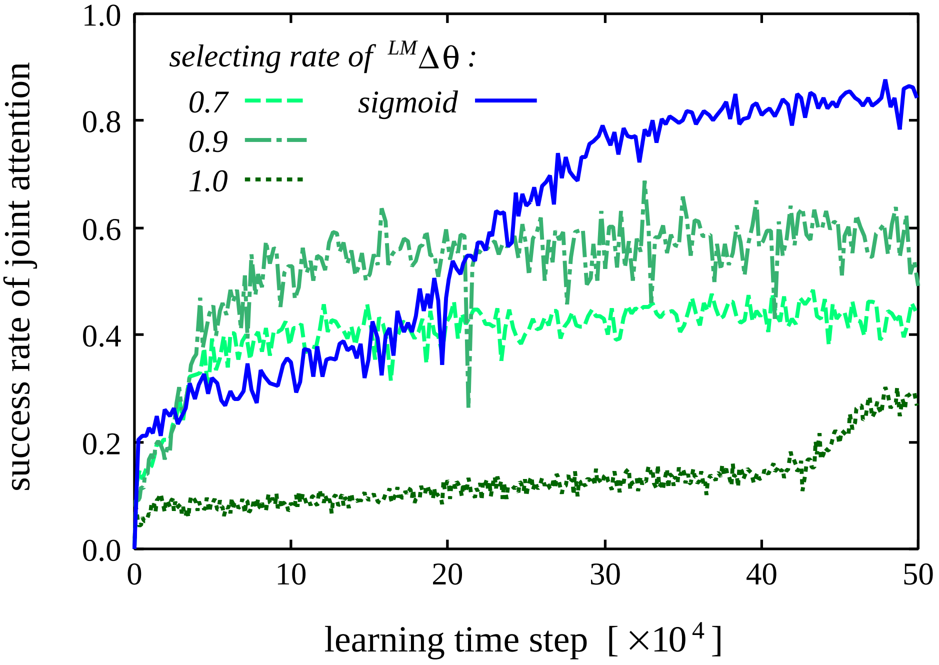


Figure 5.13: The comparison of the changes of the success rates of joint attention between when the selecting rate of $^{LM}\boldsymbol{\Delta\theta}$ is defined as the sigmoid function shown in Figure 5.11 (a) and when it is set to a constant value 0.7, 0.9, or 1.0. In the former case, the success rate of joint attention overs 85 [%] at the end of learning. In contrast, in the latter case, it cannot improve enough since learning data are biased to inadequate initial experiences based on the learning module that has not acquired appropriate sensorimotor coordination yet.

5.4.4 Acquisition of Correlation in Sensorimotor Coordination of Learning Module

It is investigated how the learning module acquires a correlation in the sensorimotor coordination for joint attention through bootstrap learning. In order to evaluate the process of structuring the sensorimotor coordination, the following procedure is taken.

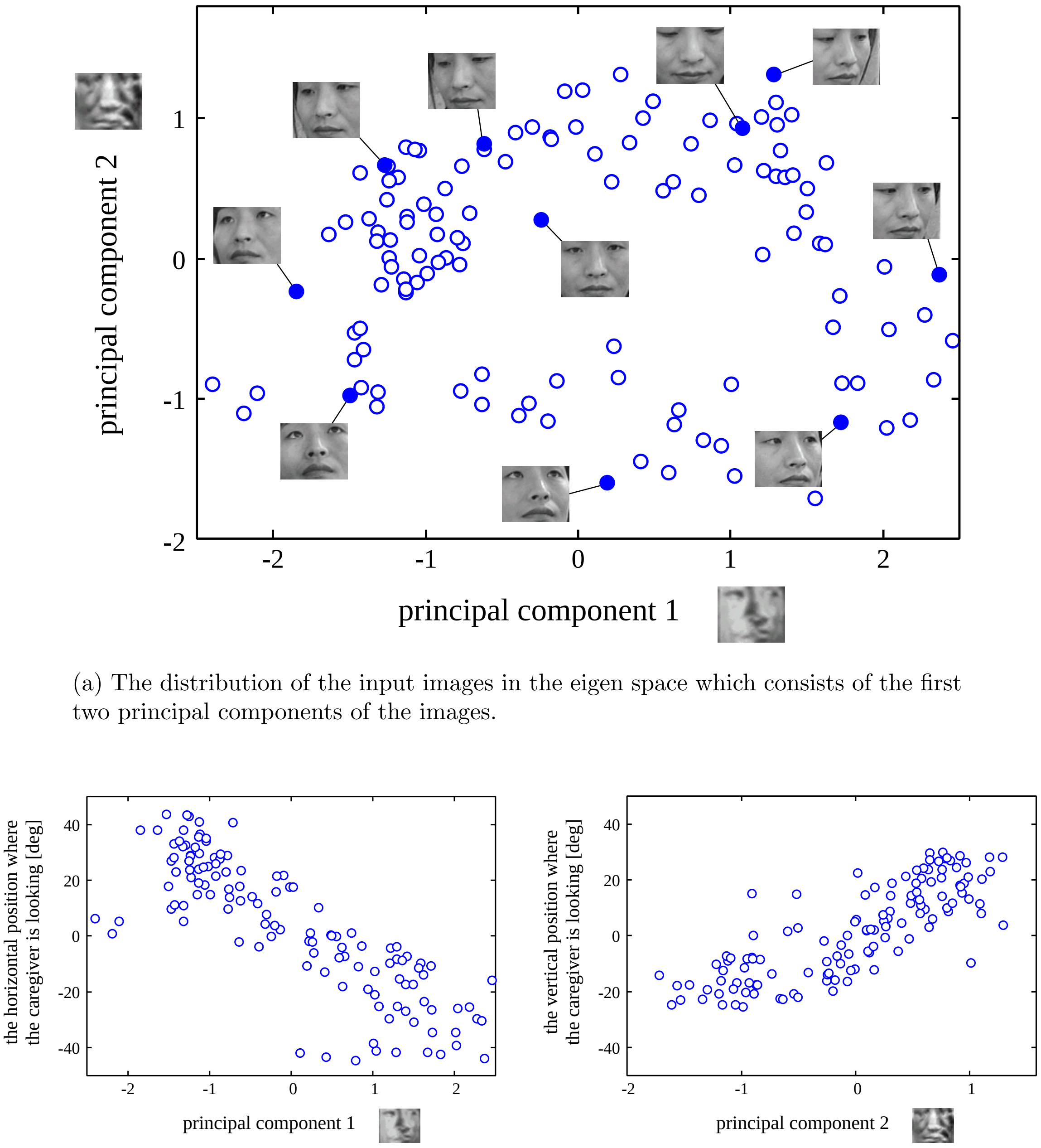
1. To visualize the sensorimotor coordination, principal component analysis is applied to the input images.

2. To define a basis of the evaluation, the correct sensorimotor coordination to realize joint attention is structured.

3. The process to acquire sensorimotor coordination is evaluated, and the acquired one is compared with the correct one.

First, in order to visualize the sensorimotor coordination in the learning module, principal component analysis [Murtagh and Heck, 1986] is applied to the input images $\boldsymbol{I}$ used in the learning experiments. The camera angle $\boldsymbol{\theta}$, which is the other input, is not taken into consideration in this evaluation because the relationship between the direction of the caregiver's gaze and the robot's motor output $\boldsymbol{\Delta\theta}$ to follow the direction has linearity when $\boldsymbol{\theta}$ is not large. The input images, each of which has 30×25 dimensions, are projected into a two dimensional space shown in Figure 5.14 (a), which is the eigen space consisting of the first two principal components of the images. The horizontal and the vertical axes denote the first and the second principal components, respectively. The eigen images are shown on the axes. 125 input images are distributed in the space, and several representative images are presented. Figures 5.14 (b) and (c) respectively show the relationship between the first principal component and the horizontal position where the caregiver is looking and the relationship between the second one and the vertical position. From these figures, it is confirmed that the first and the second principal components roughly correspond to the horizontal and the vertical components of the changes in the input images, respectively.

Next, in order to evaluate the sensorimotor coordination acquired through bootstrap learning, the correct sensorimotor coordination to realize joint attention is structured as a basis for the evaluation. Figure 5.15 (a) plots 125 correct learning data acquired when the robot has realized joint attention. The $x - y$ plane and the z axis indicate the input space and the output one, respectively. The input space is constructed by the eigen space shown in Figure 5.14 (a), and the output one shows the direction of the robot's gaze sift when the robot has realized joint attention, which is calculated by arctangent$({}^{corr}\Delta\theta_{tilt}, {}^{corr}\Delta\theta_{pan})$. This value has a linear relationship between the direction of the robot's gaze shift and ${}^{corr}\boldsymbol{\Delta\theta}$. To make the figure easier to understand, the data are approximated as a curved surface as shown in Figure 5.15 (b). The surface is formed by mean values at the grid points that are calculated based on 15 neighbor data. From this figure, it is confirmed that

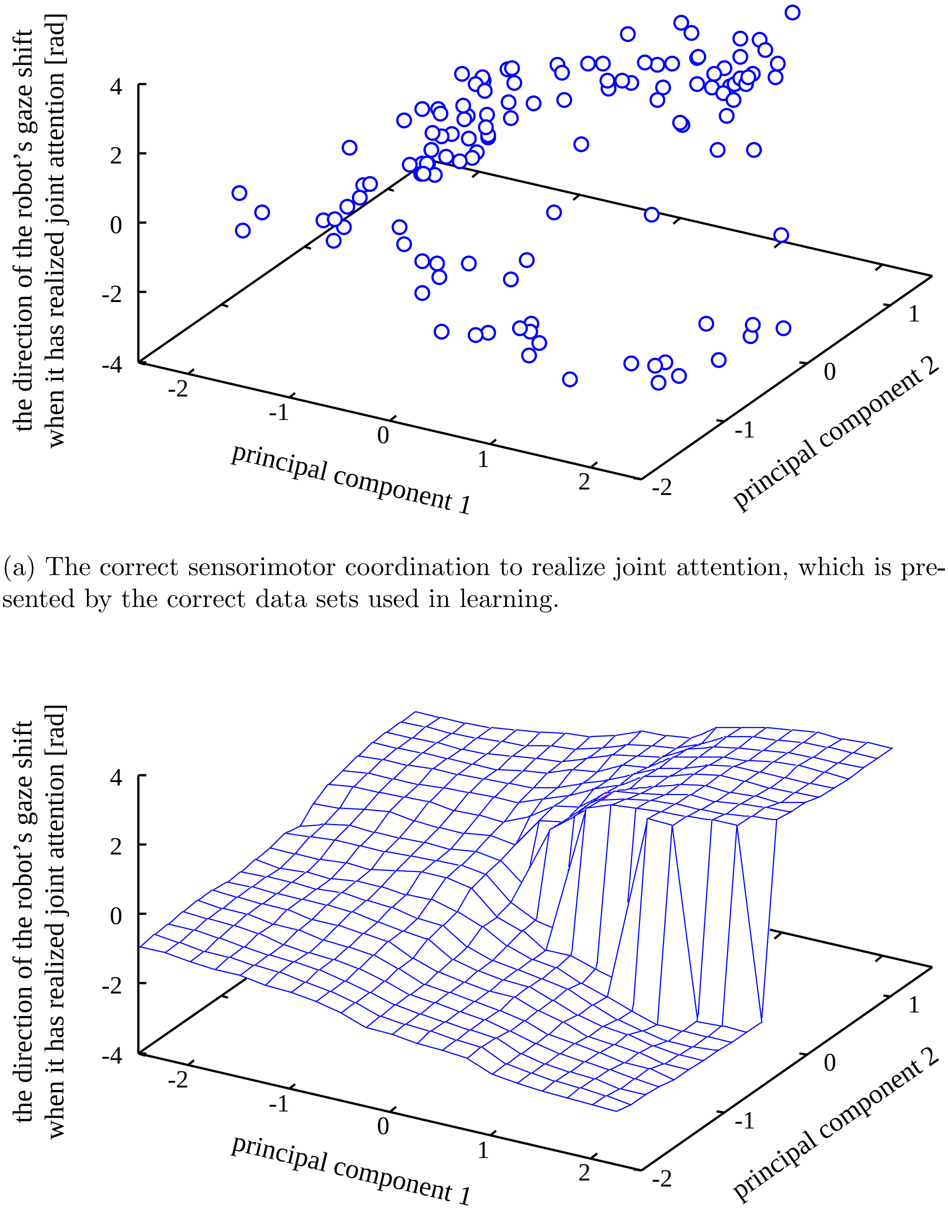


(a) The distribution of the input images in the eigen space which consists of the first two principal components of the images.

(b) The relationship between the first principal component of the input image and the horizontal position where the caregiver is looking.

(c) The relationship between the second principal component of the input image and the vertical position where the caregiver is looking.

Figure 5.14: The results of principal component analysis of the input images. The first and the second principal components seem to represent the horizontal and the vertical components of the changes in the input images, respectively.



(a) The correct sensorimotor coordination to realize joint attention, which is presented by the correct data sets used in learning.

(b) The correct sensorimotor coordination to realize joint attention, which is presented by an approximated surface of the above data.

Figure 5.15: The correct sensorimotor coordination to realize joint attention. The $x - y$ plane indicates the input space, and the z axis indicates the output space, in which the direction of the robot's gaze sift is represented as the value of arctangent$({}^{corr}\Delta\theta_{tilt}, {}^{corr}\Delta\theta_{pan})$. The discontinuity where $PC1 \geq 0$ and $PC2 \approx 0$ does not become a problem because it has continuity in the physical space of ${}^{corr}\boldsymbol{\Delta\theta}$.

the correct sensorimotor coordination forms a curved surface which has discontinuity where $PC1 \geq 0$ and $PC2 \approx 0$, in which $PC1$ and $PC2$ denote the values of the first and the second principal components. The discontinuity does not become a problem since it is caused by the computation of the function value of arctangent and has continuity in the physical space of ${}^{corr}\boldsymbol{\Delta\theta}$. The approximated surface seems to be similar to that of $z = \tan^{-1}(y/-x)$ shown in Figure 5.16. The discontinuity also appears where $PC1 \geq 0$ and $PC2 = 0$, and the surface shows the same tendency. The reason why the approximated surface of the correct learning data resembles the surface of $z = \tan^{-1}(y/-x)$ is that the principal components of the image and the motor output to realize joint attention have proportional relationships between the first component and ${}^{corr}\Delta\theta_{pan}$ and between the second and ${}^{corr}\Delta\theta_{tilt}$ as shown in Figures 5.14 (b) and (c). Such characteristics of the sensorimotor coordination for joint attention are ensured in other data sets if the input images are uniformly acquired in various situations. It means that the robot is required to find such sensorimotor coordination, i.e. a correlation, in the learning module through bootstrap learning.

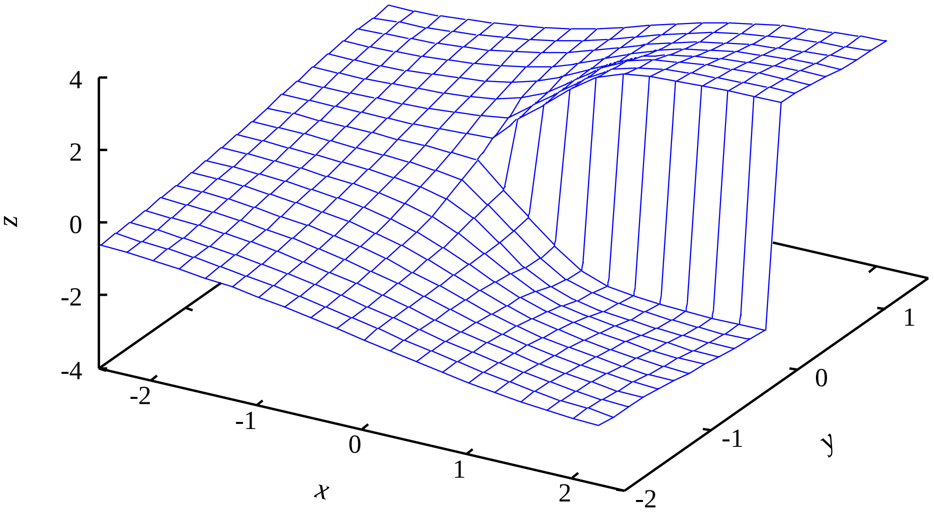


Figure 5.16: The surface of $z = \tan^{-1}(y/-x)$. The sensorimotor coordination to realize joint attention shown in Figure 5.15 (b) seems to have the same characteristics as this surface.

Based on Figure 5.15, it is evaluated how correct sensorimotor coordination the learning module acquires through bootstrap learning. To evaluate the convergence of the sensorimotor coordination, the variance of the motor output is first examined. Figure 5.18 shows the change of the variance in the motor output by the learning module, which is calculated by arctangent($^{LM}\Delta\theta_{tilt}, {}^{LM}\Delta\theta_{pan}$). (a), (b), and (c) respectively show the variance in stage I, II, and III, which are highlighted in Figure 5.11 (b). In each figure, the $x - y$ plane denotes the input space, and z axis shows the variance of the direction of the robot's gaze shift. From the result, it is confirmed that the sensorimotor coordination gradually converges as learning advances, and finally, a certain coordination is acquired. Only the area in which $PC1 \approx PC2 \approx 0$ has large variance since the value of arctangent($^{LM}\Delta\theta_{tilt}, {}^{LM}\Delta\theta_{pan}$) drastically changes as shown in Figure 5.15. Next, the acquired sensorimotor coordination through this learning process is examined. Figure 5.17 shows the sensorimotor coordination in the learning module acquired through bootstrap learning. It is presented as a surface which approximates the relationship between the input and the output of the learning module. From the comparison of Figure 5.15 (b) and Figure 5.17, we can see that the sensorimotor coordination acquired through bootstrap learning almost matches with the correct one. It means that the bootstrap learning model enables the robot to acquire the ability of joint attention by finding a correlation in the sensorimotor coordination.

Then, it is evaluated how the gate facilitates the learning module to acquire a correlation in the sensorimotor coordination through bootstrap learning. Figure 5.18 shows the result when the gate module is set to the sigmoid function shown in Figure 5.11 (a). The result is compared with that when the selecting rate of $^{LM}\boldsymbol{\Delta\theta}$ is set to zero. Figure 5.19 shows the change of the variance in the motor output by the learning module, in which (a), (b), (c), and (d) respectively show the variance in stage I, II, III, and the learning period 997-1000 [$\times 10^4$]. The representation of the variance is the same as that in Figure 5.18. From the result, it is confirmed that the sensorimotor coordination when the selecting rate of $^{LM}\boldsymbol{\Delta\theta}$ is set to zero begins to converge in stage II once; however, it gradually diverges as time goes on. The reasons are conjectured as follows.

- In the early stage of learning, initial conditions of the learning module and the bias of learning data enable the learning module to find some sort of correlation in a part of its sensorimotor coordination.

- However, as time goes on, a number of learning data including correct and incorrect ones at a certain proportion prevent the learning module to acquire a correlation in its sensorimotor coordination by making it explain the all data.

The experimental results shown in Figures 5.17–5.19 suggest that the proposed model enables the robot to acquire a correlation in the sensorimotor coordination for joint attention through bootstrap learning. Moreover, it is indicated that the gate module plays a significant role in bootstrap learning and should be designed to shift the robot's behavior from the embedded one to the acquired one at an appropriate timing.

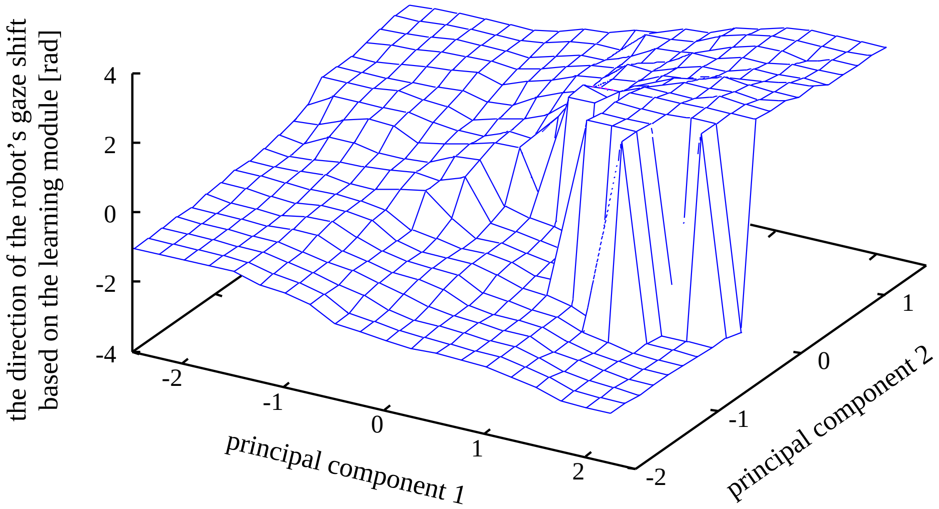
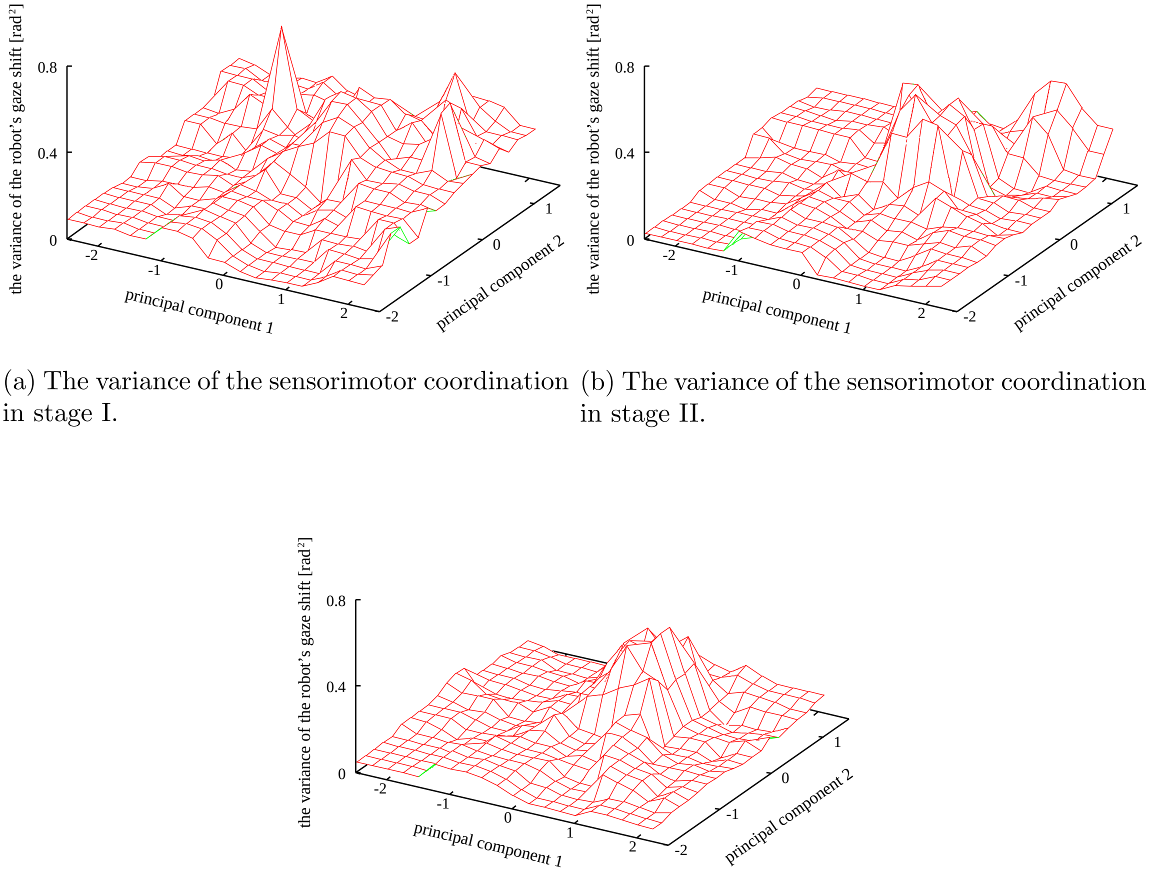


Figure 5.17: The approximated surface of the sensorimotor coordination acquired through bootstrap learning, in which the gate function was defined as the sigmoid one shown in Figure 5.11 (a). The z axis indicates the direction of the robot's gaze sift based on the learning module, which is calculated by arctangent$({}^{LM}\Delta\theta_{tilt}, {}^{LM}\Delta\theta_{pan})$. This result almost corresponds to the correct sensorimotor coordination to realize joint attention shown in Figure 5.15 (b). It means that the bootstrap learning model enables the robot to acquire the ability of joint attention by finding a correlation in its sensorimotor coordination.

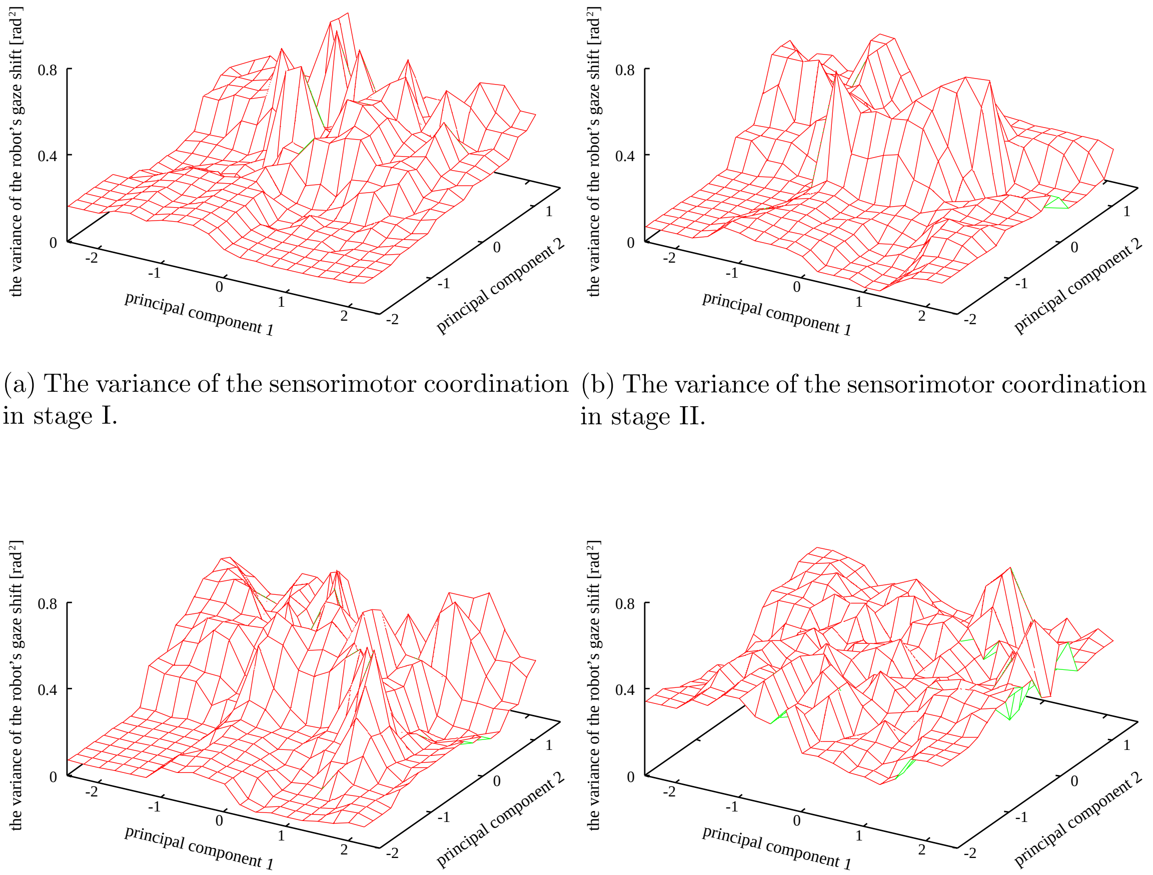


(a) The variance of the sensorimotor coordination in stage I.

(b) The variance of the sensorimotor coordination in stage II.

(c) The variance of the sensorimotor coordination in stage III.

Figure 5.18: The change of the variance of the sensorimotor coordination over learning. (a), (b), and (c) respectively show the variance in stages I, II, and III, which are highlighted in Figure 5.11 (b). The z axis shows the variance of the motor output by the learning module. In this case, the sensorimotor coordination converges to a certain one through bootstrap learning. Only the area in which $PC1 \approx PC2 \approx 0$ has large variance since the value of arctangent drastically changes in the area as shown in Figure 5.17.



(a) The variance of the sensorimotor coordination in stage I.

(b) The variance of the sensorimotor coordination in stage II.

(c) The variance of the sensorimotor coordination in stage III.

(d) The variance of the sensorimotor coordination in the learning period 997-1000 [$\times 10^4$].

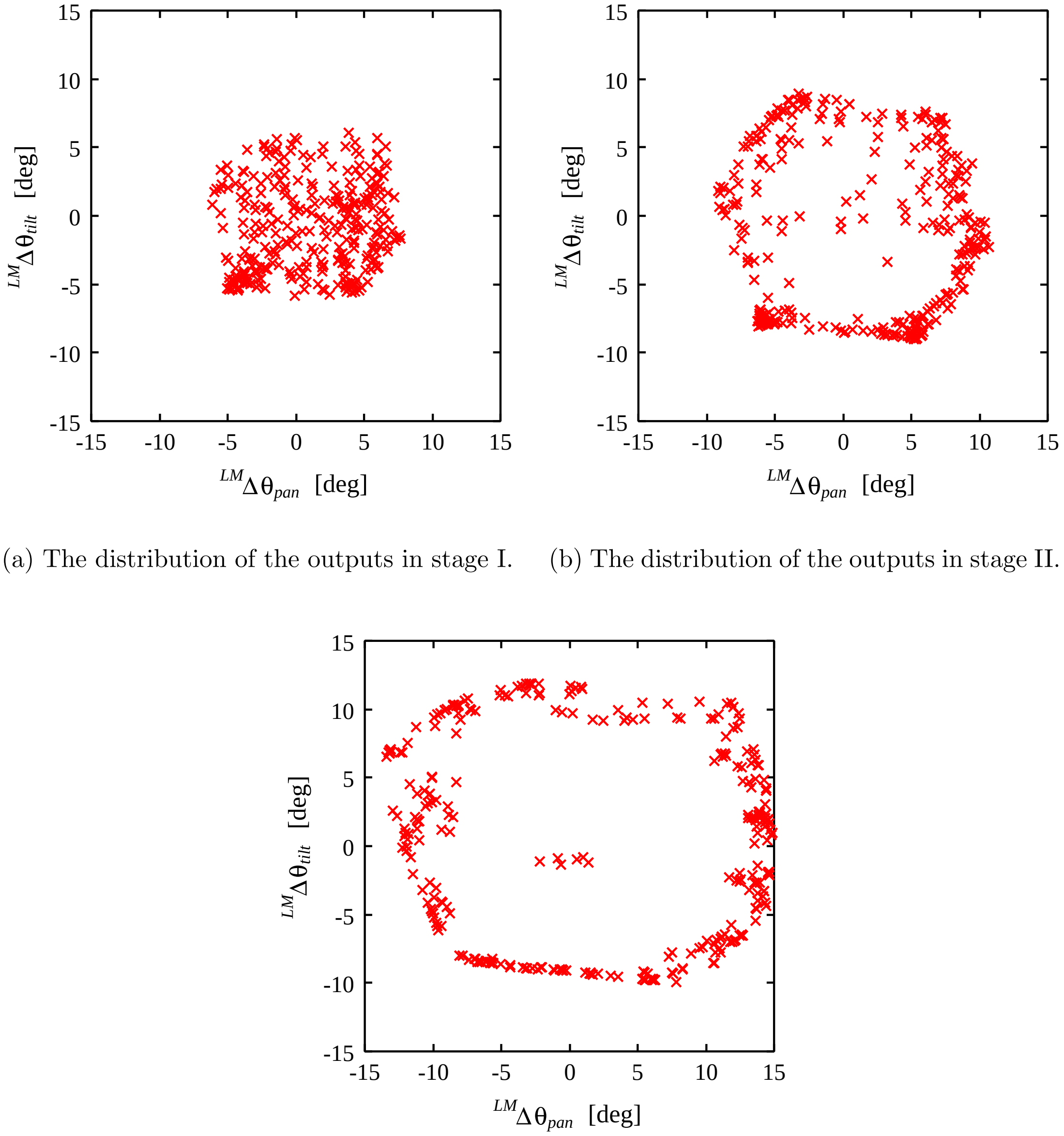
Figure 5.19: The change of the variance of the sensorimotor coordination over learning. (a), (b), (c), and (d) show the variance in stages I, II, III, and the learning period 997-1000 [$\times 10^4$], respectively. The gate module set the selecting rate of ${}^{LM}\boldsymbol{\Delta\theta}$ to zero. In other words, the robot learns its sensorimotor coordination based on the experiences of visual attention which includes joint attention only at a certain proportion. Because of this, the robot cannot find any correlation in its sensorimotor coordination.

5.4.5 Organization of Motor Output from Learning Module

It is examined how the motor output from the learning module is organized through bootstrap learning. Figures 5.20 and 5.21 show the distribution of the outputs from the learning module, each of which correspond to the results shown in Figures 5.18 and 5.19. The horizontal and the vertical axes denote ${}^{LM}\Delta\theta_{pan}$ and ${}^{LM}\Delta\theta_{tilt}$, respectively.

From Figure 5.20, it is found that the outputs from the learning module converge at a certain scale as learning advances when the gate module uses the sigmoid function shown in Figure 5.11 (a). The outputs in stage I shown in Figure 5.20 (a) are dispersed in the small area near $\sqrt{{}^{LM}\Delta\theta_{pan}{}^2 + {}^{LM}\Delta\theta_{tilt}{}^2} = 0$ [deg] because the connecting weights in the learning module are set to small random values when the learning begins. However, as learning advances, the outputs enlarge concentrically. Finally, in stage III shown in (c), the outputs converge at $\sqrt{{}^{LM}\Delta\theta_{pan}{}^2 + {}^{LM}\Delta\theta_{tilt}{}^2} \approx 12$ [deg]. It is considered that the reason why the scale of the motor outputs from the learning module converge at 12 [deg] is that the robot has an error margin $\pm W_x/6 = \pm 36/6 = \pm 6$ [deg] for visual attention as mentioned in Section 5.3. In other words, the motor outputs of which scale are 12[deg] enable the robot to look everywhere by including the error margin in the most efficient manner. The motor outputs organized as shown in Figure 5.20 (c) are useful to incrementally follow the direction of the caregiver's gaze and to realize joint attention. It is very interesting that the outputs of the learning module are organized to converge at a certain scale even though any pressure is not given to it.

In contrast to the result shown in Figure 5.20, the motor outputs shown in Figure 5.21 are not organized. This shows the result when the selecting rate of ${}^{LM}\boldsymbol{\Delta\theta}$ is set to zero. At the end of learning shown in Figure 5.21 (d), most outputs collect around $\sqrt{{}^{LM}\Delta\theta_{pan}{}^2 + {}^{LM}\Delta\theta_{tilt}{}^2} = 0$ [deg]. The reason is conjectured that the learning data including incorrect ones at a certain proportion prevent the learning module to enhance the correct sensorimotor coordination as enough to be acquired. Therefore, the sensorimotor coordination cannot converge at the correct one, and the outputs from the learning module are not organized.

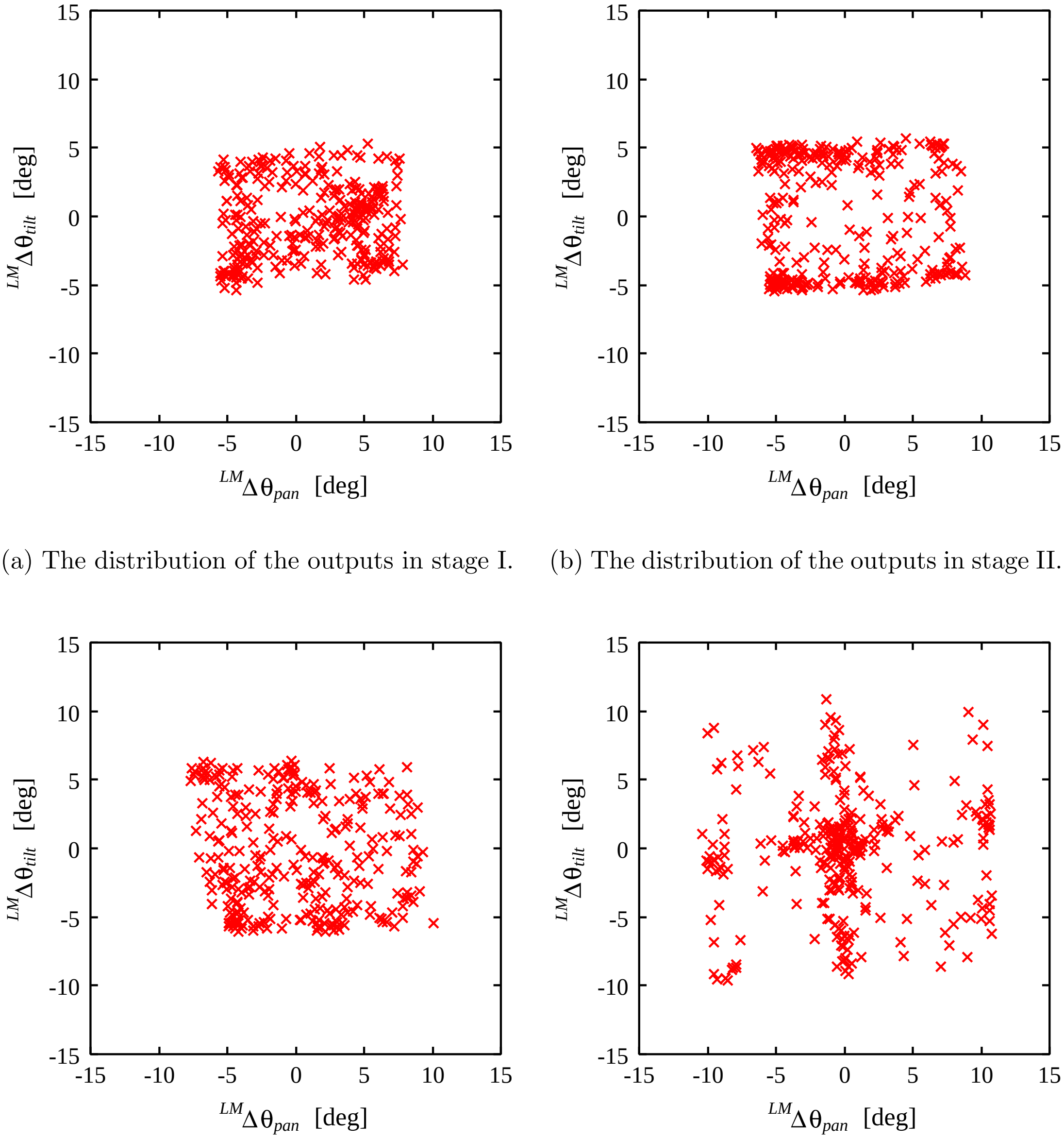


(a) The distribution of the outputs in stage I.

(b) The distribution of the outputs in stage II.

(c) The distribution of the outputs in stage III.

Figure 5.20: The process to organize the motor outputs of the learning module through bootstrap learning. (a), (b), and (c) show the distributions of the outputs in stages I, II, and III, respectively. This result corresponds to Figure 5.18, in which the gate module used the sigmoid function shown in Figure 5.11 (a). The motor outputs are organized to converge at a certain scale, i.e. $\sqrt{{}^{LM}\Delta\theta_{pan}{}^2 + {}^{LM}\Delta\theta_{tilt}{}^2} \approx 12$ [deg]. This organization enables the robot to incrementally follow the direction of the caregiver's gaze and to look everywhere by including the error margin in the most efficient manner.



(a) The distribution of the outputs in stage I.

(b) The distribution of the outputs in stage II.

(c) The distribution of the outputs in stage III.

(d) The distribution of the outputs in the learning period 997-1000 [$\times 10^4$].

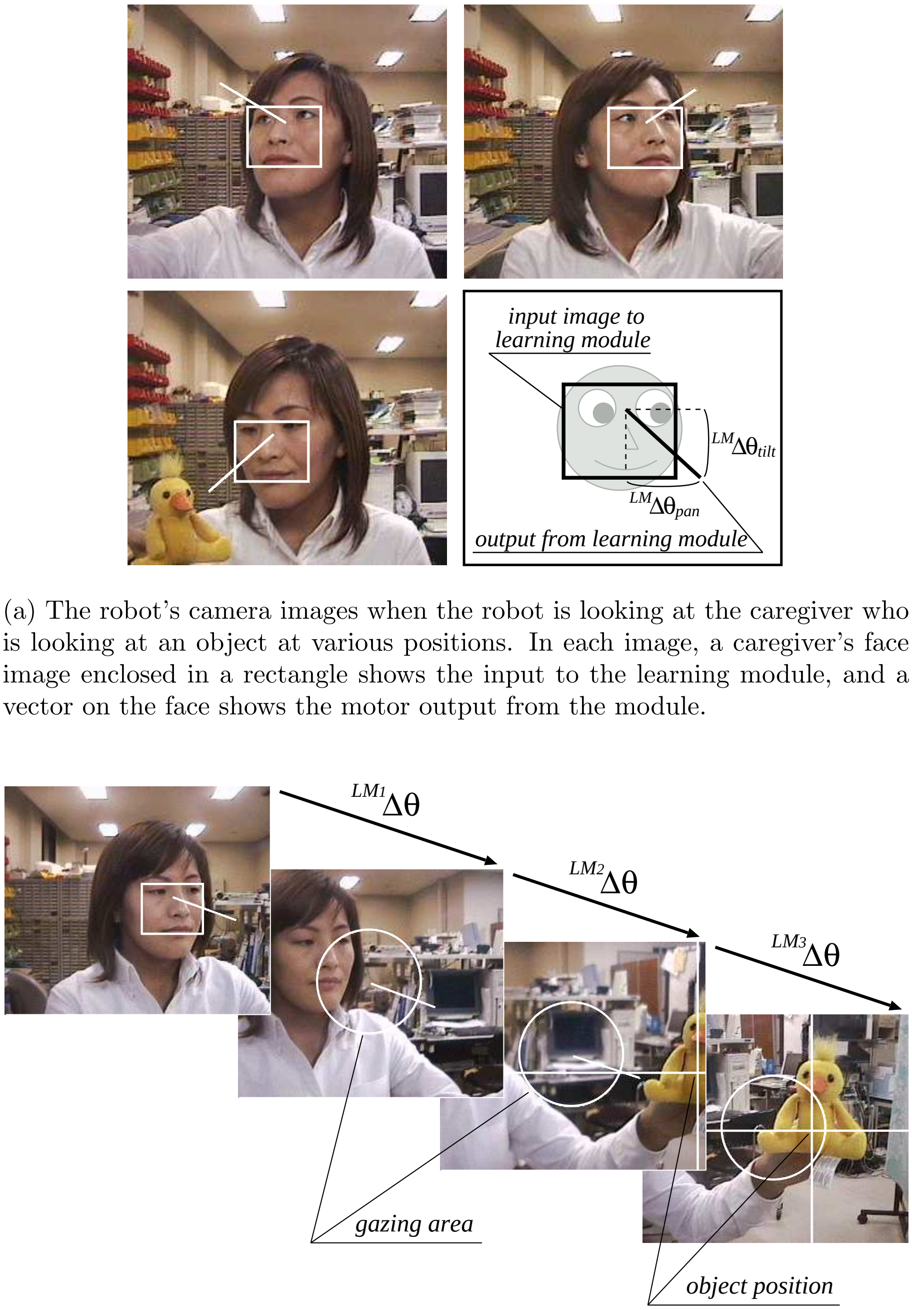
Figure 5.21: The process to organize the motor outputs of the learning module through bootstrap learning. (a), (b), (c), and (d) show the distributions of the outputs in stages I, II, III, and the learning period 997-1000 [$\times 10^4$], respectively. This result corresponds to Figure 5.19, in which the selecting rate of $^{LM}\boldsymbol{\Delta\theta}$ was set to zero. The motor outputs of the learning module are not organized compared to the result shown in Figure 5.20.

5.4.6 Experiments in Real Environment

Finally, the task performance of joint attention after learning is evaluated in a real environment. The learning module acquired through off-line learning when the number of objects was set to five is implemented in the actual robot shown in Figure 5.10. In the experiment in a real environment, the selecting rate of the output ${}^{LM}\boldsymbol{\Delta\theta}$ from the learning module is set to 1.0. The other conditions, e.g. the lighting condition and the objects' positions, are changed from those in learning while the caregiver is the same person. The experiment follows the procedure described in Section 5.1.

Figure 5.22 shows the experimental results in which the acquired sensorimotor coordination in the learning module is presented in the robot's camera images. Figure 5.22 (a) shows the camera images when the robot is gazing at the caregiver who is looking at an object at various positions. In each image, a caregiver's face image enclosed in a rectangle indicates the input to the learning module, and a vector on the face shows the output from the module. Note that a vector does not mean the gaze of the caregiver but means the motor command of the robot. In other words, the horizontal component and the vertical one of a vector show the pan and the tilt angles of the robot's motor command, respectively. The robot rotates its camera head based on the motor command and tries to find the object that the caregiver is looking at. From the results, it is confirmed that the learning module generates appropriate motor commands for joint attention since the vectors approximately correspond to the directions of the caregiver's gaze.

Figure 5.22 (b) shows the change of the robot's camera image when it shifts its gaze direction from the caregiver's face to the object based on the output from the learning module. A rectangle and a vector on the caregiver's face have the same meanings as described above. A circle and a cross line in each image respectively show the gazing area of the robot and the object's position detected by the salient feature detector. The robot generates motor commands by the learning module using the caregiver's face image in the top-left image and the angle of the camera head until finding any object in the gazing area. In this trial, the robot incrementally outputs the motor commands ${}^{LM_1}\boldsymbol{\Delta\theta}$, ${}^{LM_2}\boldsymbol{\Delta\theta}$, and ${}^{LM_3}\boldsymbol{\Delta\theta}$ at each step, and consequently finds the object that the caregiver is looking at. The success rate of joint attention



(a) The robot's camera images when the robot is looking at the caregiver who is looking at an object at various positions. In each image, a caregiver's face image enclosed in a rectangle shows the input to the learning module, and a vector on the face shows the motor output from the module.

(b) The change of the robot's camera image when it shifts its gaze direction from the caregiver's face to the object that the caregiver is looking at based on outputs from the learning module.

Figure 5.22: Experimental results of joint attention in a real environment based on the bootstrap learning model. Each result shows the robot's camera image, in which the sensorimotor coordination acquired in the learning module is indicated. The robot realized joint attention at high performance based on the acquired mechanism.

was 85 [%] (=17/20 [trials]) under various conditions in a real environment. From this result, it is concluded that the learning module has acquired the adequate ability of joint attention through bootstrap learning.

5.5 Discussion and Future Work

This chapter has presented the bootstrap learning model by which a robot learns the sensorimotor coordination for joint attention based on its embedded mechanisms in an unstructured environment without any evaluation from a caregiver. It has been suggested in cognitive developmental science that human infants are not always provided evaluation for the learning of joint attention from their caregivers. Moreover, infants seem to inherently have various capabilities that allow them to acquire the ability of joint attention. On the basis of the insight, the proposed model consists of the robot's embedded mechanisms: visual attention and learning with self-evaluation. The former is to detect and gaze at a salient object in the robot's view, and the latter is to evaluate the success of visual attention and then to learn the sensorimotor coordination. In an unstructured environment including multiple objects, the success of visual attention does not always correspond to the success of joint attention. However, the proposed model enables the robot to acquire the appropriate sensorimotor coordination for joint attention by finding a correlation only when joint attention has succeeded. From the experimental results, the followings were drawn.

- The proposed bootstrap learning model enables a robot to acquire the ability of joint attention without any evaluation from a caregiver even if the environment is not structured.

- The proposed model makes the robot reproduce the staged learning process of joint attention that is similar to the staged developmental process of infants' joint attention.

The proposed model was constructed based on the knowledge that the abilities like visual attention and learning with self-evaluation are also inherent in infants. This means that the robot has similarities with infants in both the mechanisms and the developmental phenomena of joint attention. It suggests that the proposed model could be one of the models to explain the development of infants' joint attention.

As future work, the following issues should be addressed.

- *How should a robot shift its attention mechanism from the embedded one to the acquired one?*
  The proposed bootstrap learning model utilized a gating function that was designed as a sigmoid function to shift its attention mechanism. The sigmoid function was determined by a designer in advance. However, it is conjectured that human infants shift their behaviors not based on a pre-defined schedule but based on the performance of their own behaviors. In the early stage, infants utilize their innate capabilities to behave and explore environments since the innate ones enable the infants to get some sort of pleasure through the experiences. As infants develop and acquire new capabilities through experiences, they become to utilize the acquired capabilities since the experiences based on the new capabilities are expected to bring better pleasure to the infants. Hence, the gate module in the bootstrap learning model should be designed to be adaptive according to the performance of the robot's own behaviors. For example, the performance of visual attention, not joint attention, by the learning module could be a reference to shift the robot's attention mechanism. The solution of this issue will make the proposed model more valuable in explaining the mechanism how infants shift their behaviors.

- *Do human infants really develop their ability of joint attention without any external evaluation?*
  The proposed bootstrap learning model showed that the ability of joint attention could emerge based on the robot's embedded mechanisms without any external evaluation. This is a very interesting finding not only in robotics but also in cognitive developmental science. It suggests that human infants have potentials to acquire new capabilities through their various experiences based on the innate capabilities of themselves without any external evaluation. However, in real environments where we are, infants receive not a little evaluation and help from their caregivers as discussed in Chapter 4. Caregivers are attentive to infants

and supports them in various ways. Therefore, the scheme of bootstrap learning should be integrated with the scheme of learning with external evaluation.

Through addressing the above issues, the proposed bootstrap learning model is expected to become more significant in understanding the development of infants' joint attention.

Chapter 6

Conclusions

The work presented in this dissertation has aimed at understanding the developmental mechanisms of human infants' joint attention. For the purpose, this dissertation has proposed two kinds of constructivist models for the development of joint attention and verified the validity of the models through some experiments.

The developmental processes of human infants are extremely complicated; therefore, it has attracted a number of researchers who have studied human intelligence for a long time. Observational studies and analytical ones in cognitive science, developmental psychology, and neuroscience have made many findings about developmental phenomena of infants, e.g. age-related changes of infants' behaviors, developmental disorders by psychological disabilities, brain activities for cognitive functions, and so on. However, the developmental mechanisms are still not clear. On the other hand, cognitive developmental robotics has potentials to reveal the mechanisms by constructing artificial models for robots to develop and learn like infants and demonstrating the interactions between an infant and a caregiver in the form of the interactions between a robot and a human caregiver.

This dissertation has focused on the ability of joint attention, which is a cornerstone for the further cognitive developments of infants, and discussed the mechanisms by which infants acquire the ability through interactions with their caregivers from a viewpoint of cognitive developmental robotics. Through experimental verifications of the proposed constructivist models, we have aimed at finding new knowledge that has not been known in cognitive science, developmental psychology, and neuroscience.

This chapter summarizes the two proposed models for the development of joint attention and describes the knowledge acquired through the experiments. Then, research issues that should be solved in the future are given toward deeper understanding the development of joint attention.

6.1 Summary of Two Proposed Models

The two proposed models have been constructed with the focus on external evaluation from a caregiver. Caregivers' evaluation makes a significant difference in infants' learning. Therefore, our study has discussed the two models that have been constructed based on the scheme of learning with/without caregivers' evaluation. The followings summarize the structures of the proposed models and the knowledge found through the experiments.

Developmental Learning Model with Caregiver's Evaluation

In Chapter 4, we have discussed the learning model by which a robot learns joint attention based on task evaluation from a caregiver. In this case, it is expected that the caregiver facilitates the robot's learning by adjusting how to evaluate the robot's behavior according to its performance. At the same time, the robot is able to mature its internal mechanism so that it makes learning easier. These changes are called the caregiver's development and the robot's development. In Chapter 4, it was examined how the caregiver's development and the robot's helped the learning of joint attention.

From the experimental results, the followings were drawn.

- The caregiver's development accelerates the learning of joint attention.
- Besides, the caregiver's development triggered by the progress of the robot's performance increases the effectiveness of the acceleration.
- The robot's development, which is a visual development, improves the final task performance of the robot.

- The robot's development enables the robot itself to acquire downsized and well-organized internal representation for the task of joint attention. This is considered as a reason for the improvement in the task performance.

The first and the third results demonstrated the knowledge which had been indicated in cognitive developmental science. The second and the fourth findings give new knowledge for understanding the developmental mechanisms of infants' joint attention. These findings are expected to help the researchers in cognitive developmental science to understand human developments.

Bootstrap Learning Model based on Robot's Embedded Mechanisms

In Chapter 5, we have discussed the learning model by which a robot learns joint attention based on its embedded mechanisms without any evaluation from a caregiver. Such learning is called bootstrap learning. In this case, the followings have to be considered: what mechanisms should be embedded in the robot and how the environment should be structured. The proposed model embedded the two mechanisms of visual attention and learning with self-evaluation into the robot, and it was assumed that the environment is not structured for joint attention, that is, the environment randomly changes every trial. In Chapter 5, it was verified whether the robot could acquire the ability of joint attention without any evaluation from the caregiver based on the proposed model.

From the experimental results, the followings were confirmed.

- The robot can acquire the ability of joint attention based on the embedded mechanisms of visual attention and learning with self-evaluation without any task evaluation from the caregiver.

- The robot finds a correlation in its sensorimotor coordination when joint attention has succeeded through the experiences of visual attention. Thus, it enables the robot to acquire the ability of joint attention.

- The proposed model makes the robot reproduce the staged learning process of joint attention that is similar to the developmental process of infants'.

The first two results provide a new suggestion that the ability of joint attention could emerge through bootstrap learning. It has been believed in cognitive developmental science that caregivers' supports enable infants to acquire the ability of joint attention. It is sure that caregivers' evaluation facilitates the infants' learning. However, our experimental results show that infants have a potential to acquire the ability by themselves. This is a very interesting suggestion not only in cognitive developmental science but also in robotics. The last finding is also valuable since it demonstrates that the proposed model could be one of the models to explain the developmental mechanisms of infants' joint attention. It is expected that the bootstrap learning model help the researchers in cognitive developmental science to understand the development of infants' joint attention.

6.2 Toward Deeper Understanding

The specific future work to each proposed model were described in each chapter. This section points out the common issues to the two models for deeper understanding of the development of joint attention.

Online Learning
A mechanism that enables a robot to learn not *off-line* but *online* should be invented. In the experiments based on the proposed models, a robot learns joint attention off-line using data which were acquired in a real environment in advance. The reason is that the models utilize the input data which have a high dimension and the learning method of which speed is not fast such as back-propagation. However, human infants learn joint attention online through real-time interactions with their caregivers. Furthermore, the real-time interactions seem to include essential problems of the acquisition of communication abilities. For example, an infant does not always perform joint attention in exact timing with his/her caregiver since each of them has different intention. It is interesting to discuss whether a robot can acquire the ability of joint attention through such asynchronous interactions. As another example, it is expected that a caregiver changes its behaviors according to the reaction of a robot, which is moving in the presence of the caregiver. It is also interesting to examine how the

caregiver changes its behaviors and what effects the changes have through learning. To address these issues, a mechanism of online learning should be invented.

Two-way Joint Attention

Not only *one-way* joint attention but also *two-way* joint attention should be realized. Both of the proposed models assume one-way joint attention in which a robot follows the direction of a caregiver's gaze and finds the object that the caregiver is looking at. It is adequate in the definition of joint attention; however, two-way joint attention is required to realize advanced communications. In other words, a robot is required to shift its gaze direction so that a caregiver follows the direction. In fact, human infants show various behaviors to get their caregivers' attention. Such behaviors are significant for the cognitive developments of infants. To realize such behaviors in a robot, we should construct a mechanism by which a robot acquires some kind of "intention" and an ability to make the caregiver follow the direction of the robot's gaze. For this purpose, the mechanism of real-time learning is also required.

Joint Attention with Various Caregivers

The experiments should be conducted in a situation where a robot learns and performs joint attention through interactions with *various caregivers* who are not designers. All of the experiments described in this dissertation were conducted with only one caregiver, the designer. Therefore, the learning of the robot has possibilities to be biased by the knowledge of the caregiver. Moreover, the acquired ability is not absolutely effective to realize joint attention with any other caregivers. To resolve these problems, the experiments should be conducted with several caregivers who are not the designer. This will enable the robot to acquire a better understood ability of joint attention. At the same time, it will allow us to discuss the characteristics of the acquired ability of joint attention.

Change of the Key to Realize Joint Attention from Face to Eye

It should be discussed why an infant *shifts the key* to realize joint attention from a face to eyes of his/her caregiver. Besides the knowledge described in Section 2.1, it is known that an infant changes the target to be attended to from his/her caregiver's face to the eyes [Corkum and Moore, 1995]. Infants at 12-16 months old mainly pay their attention to the face direction of their caregivers. Then, infants at 18 months old become to pay their attention not only to the face direction but also the gaze direction of their caregivers and perform joint attention only when these two directions coincide. It is considered that this change is related to the other development of infants' capabilities, e.g. visual development. Infants at 12-16 months old are conjectured not to have enough visual resolution to recognize the gaze of their caregivers. Therefore, they pay their attention to only the direction of the caregivers' face. However, when they acquire enough visual resolution, they pay their attention not only to the caregiver's face but also to the caregiver's eyes because the target which the caregiver is looking at is on the direction of the caregiver's gaze. Such a change of the key to realize joint attention is very interesting in robotics as well as in cognitive developmental science when we discuss how robots and infants extract appropriate inputs for joint attention.

A part of these issues are already addressed in the author's research group. Resolving these issues will enable us to understand deeper the development of human infants' joint attention.

Bibliography

[Adams *et al.*, 2000] Bryan Adams, Cynthia Breazeal, Rodney Brooks, and Brian Scassellati. Humanoid Robots: A New Kind of Tool. *IEEE Intelligent Systems*, 15(4):25–31, 2000.

[Asada *et al.*, 2001] Minoru Asada, Karl F. MacDorman, Hiroshi Ishiguro, and Yasuo Kuniyoshi. Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robotics and Autonomous Systems*, 37:185–193, 2001.

[ATR-IRC, 2003] ATR-IRC. Robovie Project. `http://www.irc.atr.co.jp/~m-shiomi/Robovie/index.html`, 2003.

[Baldwin, 1995] Dare A. Baldwin. *Joint Attention: Its Origins and Role in Development*, chapter 7, pages 119–143. Lawrence Erlbaum Associates, 1995.

[Banks and Dannemiller, 1987] Martin S. Banks and James L. Dannemiller. Infant visual psychophysics. In P. Salapatek and L. Cohen, editors, *Handbook of Infant Perception*, volume 1, pages 115–184. New York: Academic Press, 1987.

[Banks and Ginsburg, 1985] Martin S. Banks and Arthur P. Ginsburg. Infant visual preferences: a review and new theoretical treatment. In H. W. Reese, editor, *Advances in Child Development and Behavior*, volume 19, pages 207–246. New York: Academic Press, 1985.

[Banks, 1980] Martin S. Banks. The Development of Visual Accommodation during Early Infancy. *Child Development*, 51:646–666, 1980.

[Baron-Cohen, 1995] Simon Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1995.

[Bartleby.com, 2003] Bartleby.com. The american heritage dictionary of the english language. `http://www.bartleby.com/61/`, 2003.

[Breazeal and Aryananda, 2002] Cynthia Breazeal and Lijin Aryananda. Recognition of Affective Communicative Intent in Robot-Directed Speech. *Autonomous Robots*, 12(1):83–104, 2002.

[Breazeal and Scassellati, 1999] Cynthia Breazeal and Brian Scassellati. A Context-Dependent Attention System for a Social Robot. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1146–1151, 1999.

[Breazeal and Scassellati, 2000] Cynthia Breazeal and Brian Scassellati. Infant-like Social Interactions between a Robot and a Human Caregiver. *Adaptive Behavior*, 8(1):49–74, 2000.

[Breazeal and Scassellati, 2001] Cynthia Breazeal and Brian Scassellati. Challenges in Building Robots that Imitate People. In Kerstin Dautenhahn and Chrystopher Nehaniv, editors, *Imitation in Animals and Artifacts.* MIT Press, 2001.

[Breazeal and Scassellati, 2002] Cynthia Breazeal and Brian Scassellati. Robots that imitate humans. *Trends in Cognitive Science*, 6:481–487, 2002.

[Breazeal *et al.*, 2001] Cynthia Breazeal, Aaron Edsinger an Paul Fitzpatrick, and Brian Scassellati. Active Vision for Sociable Robot. *IEEE Transactions on System, Man, and Cybernetics*, 31(5):443–453, 2001.

[Breazeal, 2000] Cynthia Breazeal. *Sociable Machines: Expressive Social Exchange Between Humans and Robots.* PhD thesis, Massachusetts Institute of Technology, 2000.

[Breazeal, 2002] Cynthia Breazeal. Regulation and Entrainment in Human-Robot Interaction. *The International Journal of Robotics Research*, 21(10):883–902, 2002.

[Breazeal, 2003a] Cynthia Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155, 2003.

[Breazeal, 2003b] Cynthia Breazeal. Emotive qualities in lip-synchronized robot speech. *Advanced Robotics*, 17(2):97–113, 2003.

[Breazeal, 2003c] Cynthia Breazeal. Towards Sociable Robots. *Robotics and Autonomous Systems*, 42(3-4):167–175, 2003.

[Bremner, 1994] J. Gavin Bremner. *Infancy.* Blackwell, 1994.

[Brooks *et al.*, 1998] Rodney A. Brooks, Cynthia Breazeal, Robert Irie, Charles C. Kemp, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. Alternative Essences of Intelligence. In *Proceedings of the American Association of Artificial Intelligence*, pages 961–968, 1998.

[Brooks *et al.*, 1999] Rodney Brooks, Cynthia Breazeal, Matthew Marjanović, Brian Scassellati, and Matthew Williamson. The Cog Project: Building a Humanoid Robot. In Chrystopher L. Nehaniv, editor, *Computation for Metaphors, Analogy, and Agents*, pages 52–87. Springer Verlag, 1999.

[Brooks, 1991] R. A. Brooks. New Approaches to Robotics. *Science*, 253:1227–1232, 1991.

[Butterworth and Jarrett, 1991] G. E. Butterworth and N. L. M. Jarrett. What minds have in common is space: Spatial mechanisms serving joint visual attention in infancy. *British Journal of Developmental Psychology*, 9:55–72, 1991.

[Butterworth, 1991] George Butterworth. The ontogeny and phylogeny of joint visual attention. In Andrew Whiten, editor, *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, pages 223–232. Oxford: Basil Blackwell, 1991.

[Butterworth, 1995] George Butterworth. *Joint Attention: Its Origins and Role in Development*, chapter 2, pages 29–39. Lawrence Erlbaum Associates, 1995.

[Butterworth, 2000] George Butterworth. Joint Attention is Based on the Facts of Embodiment and Not on a Theory of Mind. `http://www.warwick.ac.uk/fac/soc/Philosophy/consciousness/abstracts/Butterworth.html`, 2000.

[Capurro *et al.*, 1997] C. Capurro, F. Panerai, and G. Sandini. Dynamic Vergence using Log-polar Images. *International Journal of Computer Vision*, 24(1):79–49, 1997.

[Cassia *et al.*, 2001] Viola Macchi Cassia, Francesca Simion, and Carlo Umilta. Face preference at birth: the role of an orienting mechanism. *Developmental Science*, 4:101–108, 2001.

[Charman *et al.*, 2000] Tony Charman, Simon Baron-Cohen, John Swettenham, Gillian Baird, Antony Cox, and Auriol Drew. Testing joint attention, imitation, and play as infancy precursors to language and theory of mind. *Cognitive Development*, 15:481–498, 2000.

[Corkum and Moore, 1995] Valerie Corkum and Chris Moore. *Joint Attention: Its Origins and Role in Development*, chapter 4, pages 57–76. Lawrence Erlbaum Associates, 1995.

[Currie and Manny, 1997] Debra C. Currie and Ruth E. Manny. The Development of Accommodation. *Vision Research*, 37(11):1525–1533, 1997.

[de Haan *et al.*, 2002] Michelle de Haan, Kate Humphreys, and Mark H. Johnson. Developing a Brain Specialized for Face Perception: A Converging Methods Approach. *Developmental Psychobiology*, 40:200–212, 2002.

[Deák *et al.*, 2000] Gedeon O. Deák, Ross A. Flom, and Anne D. Pick. Effects of gesture and target on 12- and 18-month-olds' joint visual attention to objects in front of or behind them. *Developmental Psychology*, 36(4):511–523, 2000.

[Deák *et al.*, 2001] Gedeon O. Deák, Ian Fasel, and Javier Movellan. The Emergence of Shared Attention: Using Robots to Test Developmental Theories. In *Proceedings of the First International Workshop on Epigenetic Robotics*, pages 95–104, 2001.

[Dickinson, 2001] Anthony Dickinson. Causal Learning: Association Versus Computation. *Current Directions in Psychological Science*, 10(4):127–132, 2001.

[Dominguez and Jacobs, 2003] Melissa Dominguez and Robert A. Jacobs. Developmental Constaints Aid the Acquisition of Binocular Disparity Sensitivities. *Neural Computation*, 15(1):161–182, 2003.

[Dominguez, 2003] Melissa Dominguez. *Developing Vision.* PhD thesis, University of Rochester, 2003.

[Dunham and Dunham, 1995] Philip J. Dunham and Frances Dunham. *Joint Attention: Its Origins and Role in Development*, chapter 8, pages 145–178. Lawrence Erlbaum Associates, 1995.

[Elman *et al.*, 1996] Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A connectionist perspective on development.* MIT Press, 1996.

[Elman, 1993] Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48:71–99, 1993.

[Emery, 2000] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604, 2000.

[Fantz, 1961] R. L. Fantz. The origin of form perception. *Scientific American*, 204(5):66–72, 1961.

[Fantz, 1963] Robert L. Fantz. Pattern Vision in Newborn Infants. *Science*, 140:296–297, 1963.

[Fasel *et al.*, 2002] Ian Fasel, Gedeon O. Deák, Jochen Triesch, and Javier Movellan. Combining Embodied Models and Empirical Research for Understanding the Development of Shared Attention. In *Proceedings of the 2nd International Conference on Development and Learning*, pages 21–27, 2002.

[Fong *et al.*, 2003] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.

[Hains and Muir, 1996] Sylvia M. J. Hains and Darwin W. Muir. Effects of stimulus contingency in infant-adult interactions. *Infant Behavior and Development*, 19:49–61, 1996.

[Imai *et al.*, 2001] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. Physical Relation and Expression: Joint Attention for Human-Robot Interaction. In *Proceedings of 10th IEEE International Workshop on Robot and Human Communication*, 2001.

[Ishiguro *et al.*, 2001] Hiroshi Ishiguro, Tetuo Ono, Michita Imai, Takeshi Maeda, Takayuki Kanda, and Ryohei Nakatsu. Robovie: an Interactive Humanoid Robot. *International Journal of Industrial Robot*, 28(6):498–503, 2001.

[Ishimura, 1992] 石村貞夫 Sadao Ishimura. 分散分析のはなし. 東京図書, 1992.

[Jacobs and Dominguez, 2003] Robert A. Jacobs and Melissa Dominguez. Visual Development and the Acquisition of Motion Velocity Sensitivities. *Neural Computation*, 15(4):761–781, 2003.

[Johnson, 1997] Mark H. Johnson. *Developmental Cognitive Neuroscience*. Blackwell, 1997.

[Kanda *et al.*, 2002a] Takayuki Kanda, Hiroshi Ishiguro, Michita Imai, Tetsuo Ono, and Kenji Mase. A constructive approach for developing interactive humanoid robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1265–1270, 2002.

[Kanda *et al.*, 2002b] Takayuki Kanda, Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, and Ryohei Nakatsu. Development and Evaluation of an Interactive Humanoid Robot "Robovie". In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1848–1855, 2002.

[Kanda *et al.*, 2003] Takayuki Kanda, Hiroshi Ishiguro, Michita Imai, and Tetsuo Ono. Body Movement Analysis of Human-Robot Interaction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 177–182, 2003.

[Kozima and Yano, 2001] Hideki Kozima and Hiroyuki Yano. A Robot that Learns to Communicate with Human Caregivers. In *Proceedings of the First International Workshop on Epigenetic Robotics*, 2001.

[Kozima and Zlatev, 2000] Hideki Kozima and Jordan Zlatev. An Epigenetic Approach to Human-Robot Communication. In *Proceedings of the IEEE International Workshop on Robot and Human Communication*, 2000.

[Kozima, 1998] Hideki Kozima. Attention-Sharing and Behavior-Sharing in Human-Robot Communication. In *Proceedings of the IEEE International Workshop on Robot and Human Communication*, pages 9–14, 1998.

[Kozima, 2000] Hideki Kozima. Infanoid: An Experimental Tool for Developmental Psycho-Robotics. In *Proceedings of the International Workshop on Developmental Study*, 2000.

[Kozima, 2002] Hideki Kozima. Infanoid: A Babybot that Explores the Social Environment. In Kerstin Dautenhahn, Alan H. Bond, Lola Canamero, and Bruce Edmonds, editors, *Socially Intelligent Agents: Creating Relationships with Computers and Robots*, pages 157–164. Kluwer Academic Publishers, 2002.

[Kozima, 2003] Hideki Kozima. Research on Pre-Verbal Communication. `http://www2.crl.go.jp/jt/a134/xkozima/research/index.html`, 2003.

[Kumashiro *et al.*, 2003] Mari Kumashiro, Hidetoshi Ishibashi, Yukari Uchiyama, Shoji Itakura, Akira Murata, and Atsushi Iriki. Natural imitation induced by joint attention in japanese monkeys. *International Journal of Psychophysiology*, 50:81–99, 2003.

[Leslie and Keeble, 1987] Alan M. Leslie and Stephanie Keeble. Do six-month-old infants perceive causality? *Cognition*, 25:265–288, 1987.

[Leslie, 1994] Alan M. Leslie. ToMM, ToBy and Agency: Core architecture and domain specificity. In Lawrence A. Hirschfeld and Susan A. Gelman, editors, *Mapping the mind: Domain specificity in cognition and culture*, pages 119–148. Cambridge University Press, 1994.

[LIRA-Lab, 2003] LIRA-Lab. The Babybot project. `http://www.lira.dist.unige.it/babybotmain.htm`, 2003.

[Lungarella and Metta, 2003] Max Lungarella and Giorgio Metta. Beyond gazing, pointing, and reaching: A Survey of Developmental Robotics. In *Proceedings of the 3rd International Workshop on Epigenetic Robotics*, pages 81–89, 2003.

[Metta *et al.*, 1999] G. Metta, G. Sandini, and J. Konczak. A developmental approach to visually-guided reaching in artificial systems. *Neural Newtorks*, 12:1413–1427, 1999.

[Metta *et al.*, 2000] Giorgio Metta, Francesco Panerai, Riccardo Manzotti, and Giulio Sandini. Babybot: an artificial developing robotic agent. In *The Sixth International Conference on the Simulation of Adaptive Behavior*, pages 11–16, 2000.

[Metta *et al.*, 2001] Giorgio Metta, Giulio Sandini, Lorenzo Natale, and Francesco Panerai. Development and Robotics. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*, pages 33–42, 2001.

[Metta, 2000] Giorgio Metta. *Babyrobot: A Study on Sensori-motor Development*. PhD thesis, University of Genova, 2000.

[MIT-AI-Lab, 2003a] MIT-AI-Lab. Cog. `http://www.ai.mit.edu/projects/humanoid-robotics-group/cog/cog.html`, 2003.

[MIT-AI-Lab, 2003b] MIT-AI-Lab. Kismet. `http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/kismet.html`, 2003.

[Miyashita and Ishiguro, 2003] Takahiro Miyashita and Hiroshi Ishiguro. Natural Behavior Generation for Humanoid Robots. In *Proceedings of the 3rd IEEE International Conference on Humanoid Robots*, 2003.

[Mondloch *et al.*, 1999] Catherine J. Mondloch, Terri L. Lewis, D. Robert Budreau, Daphne Maurer, James L. Dannemiller, Benjamin R. Stephens, and Kathleen A. Kleiner-Gathercoal. Face Perception During Early Infancy. *Psychological Science*, 10:419–422, 1999.

[Moore and Corkum, 1994] Chris Moore and Valerie Corkum. Social Understanding at the End of the First Year of Life. *Developmental Review*, 14(4):349–450, 1994.

[Moore and Dunham, 1995] Chris Moore and Philip J. Dunham, editors. *Joint Attention: Its Origins and Role in Development*. Lawrence Erlbaum Associates, 1995.

[Morales *et al.*, 1998] Michael Morales, Peter Mundy, and Jennifer Rojas. Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development*, 21(2):373–377, 1998.

[Morales, 2000] Michael Morales. Responding to Joint Attention Across the 6-Through 24-Month Age Period and Early Language Acquisition. *Journal of Applied Developmental Psychology*, 21(3):283–298, 2000.

[Morton and Johnson, 1991] J. Morton and M. H. Johnson. Conspec and conlearn: a two process theory of infant face recognition. *Psychological Review*, 98:164–181, 1991.

[Movellan and Watson, 2002] Javier R. Movellan and John S. Watson. The Development of Gaze Following as a Bayesian Systems Identification Problem. In *Proceedings of the 2nd International Conference on Development and Learning*, pages 34–40, 2002.

[Mundy and Gomes, 1998] Peter Mundy and Antoinette Gomes. Individual differences in joint attention skill development in the second year. *Infant Behavior and Development*, 21(3):469–482, 1998.

[Murtagh and Heck, 1986] Fionn Murtagh and Andre Heck. *Multivariate Data Analysis*. Kluwer Academic, 1986.

[Nadel *et al.*, 1999] Jacqueline Nadel, Isabelle Carchon, Claude Kervella, Daniel Marcelli, and Denis Réserbat-Plantey. Expectancies for social contingency in 2-month-olds. *Developmental Science*, 2(2):164–173, 1999.

[Natale *et al.*, 2002] Lorenzo Natale, Giorgio Metta, and Giulio Sandini. Development of auditory-evoked reflexes: Visuo-acoustic cues integration in a binocular head. *Robotics and Autonomous Systems*, 39:87–106, 2002.

[Newman and Newman, 2003] Barbara M. Newman and Philip R. Newman. *Development Through Life: A Psychosocial Approach*, chapter 6, pages 132–175. http://newtexts.com/newtexts/book.cfm?book_id=837, 2003.

[Newport, 1990] Elissa L. Newport. Maturational constraints on language learning. *Cognitive Science*, 14:11–28, 1990.

[Ono *et al.*, 2000] Tetsuo Ono, Michita Imai, and Ryohei Nakatsu. Reading a Robot's Mind: A Model of Utterance Understanding based on the Theory of Mind Mechanism. *Advanced Robotics*, 13(4):311–326, 2000.

[Pfeifer and Scheier, 1999] Rolf Pfeifer and Christian Scheier. *Understanding Intelligence.* The MIT Press, 1999.

[Reddy, 2003] Vasudevi Reddy. On being the object of attention: implications for self-other consciousness. *Trends in Cognitive Sciences*, 7(9):397–402, 2003.

[Sandini *et al.*, 1997] G. Sandini, G. Metta, and J. Konczak. Human Sensori-motor Development and Artificial Systems. In *Proceedings of AIR&IHAS*, 1997.

[Scaife and Bruner, 1975] M. Scaife and J. S. Bruner. The capacity for joint visual attention in the infant. *Nature*, 253:265–266, 1975.

[Scassellati, 1996] Brian Scassellati. Mechanisms of Shared Attention for a Humanoid Robot. In *Proceedings of the 1996 AAAI Fall Symposium on Embodied Cognition and Action*, 1996.

[Scassellati, 1998] Brian Scassellati. Eye Finding via Face Detection for a Foveated, Active Vision System. In *Proceedings of the American Association of Artificial Intelligence*, 1998.

[Scassellati, 1999] Brian Scassellati. Imitation and Mechanisms of Joint Attention: A Developmental Structure for Building Social Skills on a Humanoid Robot. In Chrystopher L. Nehaniv, editor, *Computational for Metaphors, Analogy, and Agents*, pages 176–195. Springer Verlag, 1999.

[Scassellati, 2000] Brian Scassellati. Theory of mind for a humanoid robot. In *Proceedings of the First IEEE-RAS International Conference on Humanoid Robots*, 2000.

[Scassellati, 2001a] Brian Scassellati. Discriminating animate from inanimate visual stimuli. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2001.

[Scassellati, 2001b] Brian Scassellati. *Foundations for a Theory of Mind for a Humanoid Robot.* PhD thesis, Massachusetts Institute of Technology, 2001.

[Scassellati, 2001c] Brian Scassellati. Investingating Models of Social Development Using a Humanoid Robot. In Barbara Webb and Thomas Consi, editors, *Biorobotics.* MIT Press, 2001.

[Scassellati, 2002] Brian Scassellati. Theory of Mind for a Humanoid Robot. *Autonomous Robots*, 12:13–24, 2002.

[Shankle, 2003] Rodman Shankle. The data of behavior development of human infants. This data was contributed to us with Dr. Shankle's good intent., 2003.

[Simion *et al.*, 2001] Francesca Simion, Viola Macchi Cassia, Chiara Turati, and Eloisa Valenza. The Origins of Face Perception: Specific Versus Non-specific Mechanisms. *Infant and Child Development*, 10:59–65, 2001.

[Sutton and Barto, 1998] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* MIT Press, 1998.

[Taga, 2002] 多賀厳太郎 Gentaro Taga. 脳と身体の動的デザインー運動・知覚の非線形力学と発達. 金子書房, 2002.

[Triesch *et al.*, 2003] Jochen Triesch, Eric Carlson, Gedeon Deák, and Javier Movellan. Investigating the Emergence of Shared Attention through an Embodied Computational Modeling Approach: A Progress Report. In *Proceedings of the Joint International Conference on Neural Networks*, 2003.

[Turing, 1950] A. M. Turing. Computing Machinery and Intelligence. *Mind*, 49:433–460, 1950.

[Uchibe *et al.*, 1998] Eiji Uchibe, Minoru Asada, and Koh Hosoda. Environmental Complexity Control for Vision-Based Learning Mobile Robot. In *Proceedings of IEEE International Conference on Robotics and Automation*, pages 1865–1870, 1998.

[UCSD, 2003] UCSD. MESA Project. `http://mesa.ucsd.edu/`, 2003.

Appendix A

Statistical Test of Final Task Performance by Tukey's Method

This appendix describes a statistical test of the experimental result shown in Figure 4.10, which indicates the robot's task performance of joint attention acquired by the four learning models: *RC-dev. model*, *R-dev. model*, *C-dev. model*, and *Matured model*. The significance of the experimental result is examined by Tukey's method, which is one of the methods for multiple comparison [Ishimura, 1992].

Table A.1 lists the average of the normalized output error $\bar{x}_i$ $(= E_k)$ and its standard deviation of the each learning model. The size of the samples of each model n is 45, which is the number of the unknown data applied to the learning models.

Table A.1: The average of the normalized output error $\bar{x}_i$ and its standard deviation. The data correspond to the experimental result shown in Figure 4.10.

learning model	the average of the normalized output error $\bar{x}_i$	standard deviation
RC-dev. model	0.12773	0.08086
R-dev. model	0.12528	0.04548
C-dev. model	0.17055	0.08701
Matured model	0.18871	0.06692

The fluctuation within the each model S_E is given as

$$S_E = \sum_{i=1}^{4} \sum_{m=1}^{45} (x_{im} - \bar{x}_i)^2 = 0.92943, \tag{A.1}$$

and its mean square V_E is

$$V_E = \frac{S_E}{a(n-1)} = \frac{0.92943}{4 \times (45-1)} = 0.00528, \tag{A.2}$$

where $a = 4$ denotes the number of the learning models. The studentized range $q(a, a(n-1); \alpha)$ at the level of the significance 5 [%] ($\alpha = 0.05$) is defined as

$$3.6332 < q(a, a(n-1); \alpha) = q(4, 176; 0.05) < 3.6846, \tag{A.3}$$

from the table [Ishimura, 1992]. This derives the following.

$$0.03936 < q(a, a(n-1); \alpha)\sqrt{\frac{V_E}{n}} = q(4, 176; 0.05)\sqrt{\frac{0.00528}{45}} < 0.03991 \tag{A.4}$$

In this equation, if the condition of

$$|\bar{x}_i - \bar{x}_j| \geq q(a, a(n-1); \alpha)\sqrt{\frac{V_E}{n}} \tag{A.5}$$

is satisfied between the two learning models, i and j, they have a significant difference between themselves. The differences of the output error $|\bar{x}_i - \bar{x}_j|$ are shown in Table A.2, in which "$*$" means $|\bar{x}_i - \bar{x}_j| \geq 0.03991$.

Table A.2: The differences of the output error $|\bar{x}_i - \bar{x}_j|$ between two learning models, i and j. "$*$" indicates $|\bar{x}_i - \bar{x}_j| \geq 0.03991$.

	R-dev. model	*C-dev. model*	*Matured model*
RC-dev. model	0.00245	0.04282*	0.06098*
R-dev. model	—	0.04527*	0.06343*
C-dev. model	—	—	0.01816

* There is a significant difference at the level of 5 [%].

From Table A.2, it can be confirmed that the experimental result of the robot's task performance shown in Figure 4.10 has significant differences between

- *RC-dev. model* and *C-dev. model*,
- *RC-dev. model* and *Matured model*,
- *R-dev. model* and *C-dev. model*, and
- *R-dev. model* and *Matured model.*

at the level of the significance 5 [%]. In other words, all the models that include the robot's development have significant differences compared to all the models that do not include the development.

Appendix B

Mechanism of Bootstrap Learning of Joint Attention

This appendix explains the mechanism of bootstrap learning of joint attention. First, a mathematical proof of bootstrap learning is described. The proof shows how experimental conditions enable a robot to acquire the ability of joint attention through bootstrap learning. Then, an example in a simple environmental setup is given. The process to acquire the sensorimotor coordination in which the coordination when joint attention has succeeded is relatively enhanced is illustrated.

B.1 Mathematical Proof of Bootstrap Learning

This section describes the mathematical proof of bootstrap learning. We assume an environmental setup shown in Figure B.1. A robot is seated face-to-face with a caregiver. An environment is quantized in n positions, in which the i-th position is denoted as $\boldsymbol{x}_i$. The positions where the robot and the caregiver are looking are defined as $\boldsymbol{r}$ and $\boldsymbol{c}$, respectively. In this setup, the following assumptions are set:

(A-1) m objects are placed at random positions in the environment in every trial, and

(A-2) the robot and the caregiver look at one object that was randomly selected by each other.

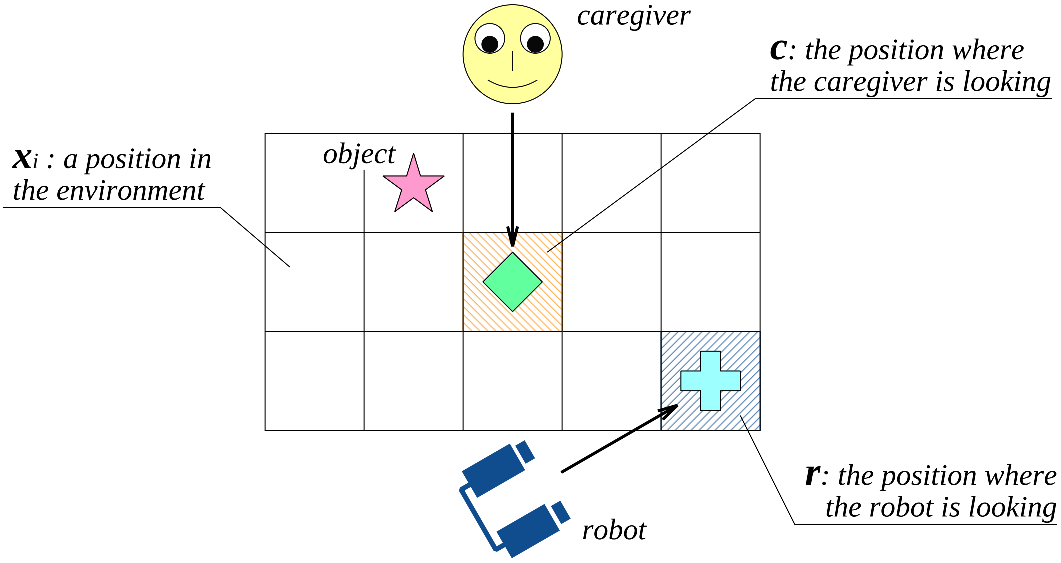

caregiver
c: the position where the caregiver is looking
x_i : a position in the environment
object
r: the position where the robot is looking
robot

Figure B.1: An environmental setup for the mathematical proof of bootstrap learning, in which the number of the quantized environment and that of objects are denoted as n and m, respectively. From the mathematical proof, it is derived that the sensorimotor coordination acquired when joint attention has succeeded is relatively enhanced at $(n-1)/(m-1)$ times compared to the coordination when joint attention has failed. It means that bootstrap learning enables a robot to acquire the ability of joint attention if m is adequately smaller than n.

The assumption (A-2) means that the probability that an object is gazed at by the robot or the caregiver becomes $1/m$.

Consider a situation in which the caregiver is looking at the position $\boldsymbol{x}_j$, i.e. $\boldsymbol{c} = \boldsymbol{x}_j$. Based on the assumption (A-2), the probability that an object is placed at the position $\boldsymbol{x}_j$ equals to 1. At the same time, based on the assumption (A-1), the probability that an object is placed at the other position $\boldsymbol{x}_i$ $(i \neq j)$ equals to $(m-1)/(n-1)$. In this situation, the probability that the robot looks at the position $\boldsymbol{x}_j$, where the caregiver is looking, becomes as below.

$$\Pr(\boldsymbol{r} = \boldsymbol{x}_j | \boldsymbol{c} = \boldsymbol{x}_j) = 1 \times \frac{1}{m} = \frac{1}{m} \tag{B.1}$$

On the other hand, the probability that the robot looks at the position $\boldsymbol{x}_i$ $(i \neq j)$ becomes as below.

$$\Pr(\boldsymbol{r} = \boldsymbol{x}_i | \boldsymbol{c} = \boldsymbol{x}_j, i \neq j) = \frac{m-1}{n-1} \times \frac{1}{m} = \frac{m-1}{m(n-1)} \tag{B.2}$$

The learning of the sensorimotor coordination of the robot depends on the number of the experiences of visual attention. From Eqs. (B.1) and (B.2), the ratio of the number that the sensorimotor coordination when joint attention has succeeded is learned to the number that the sensorimotor coordination when joint attention has failed is learned is driven as the following.

$$\frac{\#\text{ of J.A.}}{\#\text{ of non-J.A.}} = \frac{\Pr(\boldsymbol{r} = \boldsymbol{x}_j | \boldsymbol{c} = \boldsymbol{x}_j)}{\Pr(\boldsymbol{r} = \boldsymbol{x}_i | \boldsymbol{c} = \boldsymbol{x}_j, i \neq j)} = \frac{\frac{1}{m}}{\frac{m-1}{m(n-1)}} = \frac{n-1}{m-1} \tag{B.3}$$

This means that the sensorimotor coordination when joint attention has succeeded is relatively enhanced at $(n-1)/(m-1)$ times compared to the coordination when joint attention has failed. In other words, bootstrap learning enables the robot to acquire the ability of joint attention if the number of objects m is adequately smaller than that of the quantized environment n under the assumptions (A-1) and (A-2).

B.2 Example of Bootstrap Learning of Joint Attention

This section illustrates a simple example in which the sensorimotor coordination when joint attention has succeeded is relatively enhanced through bootstrap learning. Figure B.2 shows an environmental setup and the sensorimotor coordination of the robot. It is supposed that an environment is quantized in three positions, and two objects are placed by rotation, i.e. $n = 3$ and $m = 2$. Under this condition, the robot has three possibilities in each of the visual inputs and the motor outputs and learns the connections between them based on the mechanisms of visual attention and learning with self-evaluation. It is expected that the sensorimotor coordination when joint attention has succeeded is relatively enhanced compared to the coordination when joint attention has failed, and the ratio of the enhancement becomes $(n-1)/(m-1) = (3-1)/(2-1) = 2$ times.

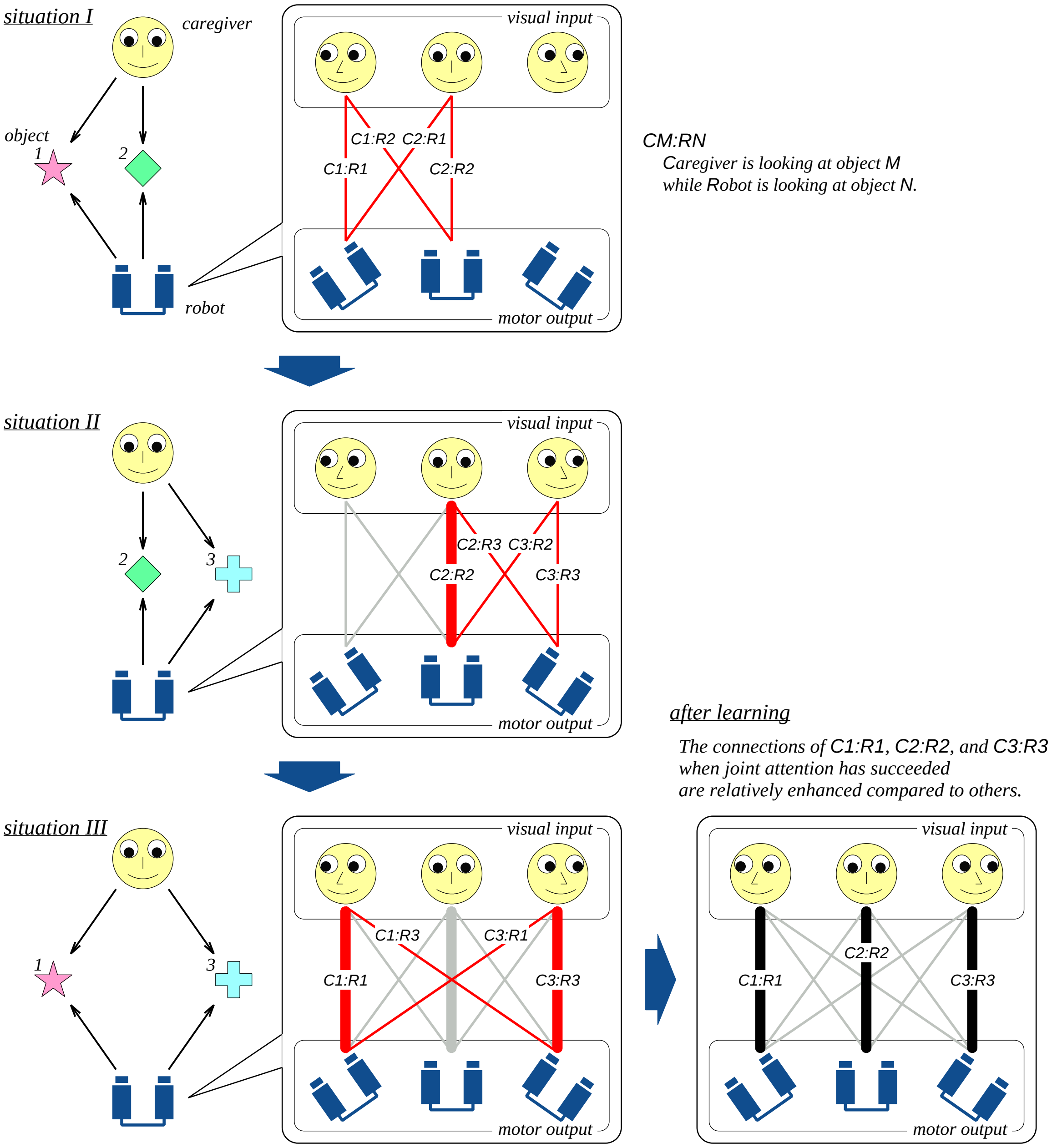


Figure B.2: The process of finding a correlation for joint attention in the sensorimotor coordination through bootstrap learning. In each situation, the robot learns its sensorimotor coordination when visual attention has succeeded. As a result, the robot acquires the sensorimotor coordination shown in the lower-right corner, in which the connections of *C1:R1*, *C2:R2*, and *C3:R3* when joint attention has succeeded are relatively enhanced twice compared to others. This result means that the proposed model enables the robot to acquire the ability of joint attention by finding the correlation in the sensorimotor coordination. In addition, this result is consistent with the mathematical proof described in the previous section.

The robot learns its sensorimotor coordination through the following process.

situation I: First, it is supposed that the environment includes the objects 1 and 2, and the robot as well as the caregiver have two choices to look at an object based on the mechanism of visual attention. In this situation, the robot learns its sensorimotor coordination in all possible cases. The robot acquires the connections of *C1:R1* (which means that the *C*aregiver is looking at the object *1*, and the *R*obot is looking at the object *1*), *C1:R2*, *C2:R1*, and *C2:R2*. Among the acquired connections, only *C1:R1* and *C2:R2* correspond to the success of joint attention. However, the strength of all four connections are equivalent at this time since these connections have been learned at a same rate.

situation II: Next, the environment changes into the situation including the objects 2 and 3. The sensorimotor coordination acquired in situation I is maintained as it is. In this situation, the robot learns its sensorimotor coordination in the same manner as in situation I. The robot newly acquires the connections of *C2:R2*, *C2:R3*, *C3:R2*, and *C3:R3* at a same rate. As a result, it can be found that only the connection of *C2:R2* has a double strength against the others since it has been learned at two times compared to the others.

situation III: Finally, the environment changes into the situation including the objects 1 and 3. The sensorimotor coordination acquired in situations I and II is maintained as it is. In this situation, the robot learns its sensorimotor coordination in the same manner as in situations I and II. The robot acquires the connections of *C1:R1*, *C1:R3*, *C3:R1*, and *C3:R3* at a same rate. As a result, the connections of *C1:R1* and *C3:R3* have a double strength just like *C2:R2* since they have also learned at two times compared to the others.

In the consequence of the above process, the robot acquires the sensorimotor coordination as shown in the lower-right corner of Figure B.2, in which the connections of *C1:R1*, *C2:R2*, and *C3:R3* are enhanced twice compared to the others. It can be found that these enhanced connections were acquired when the robot looked at the same object that the caregiver was looking at, that is, when joint attention succeeded. Furthermore, the connections show a correlation in the sensorimotor coordination. From the result, it can be concluded that the proposed bootstrap learning model enables the robot to acquire the ability of joint attention by finding the correlation in the sensorimotor coordination. In addition, the result that the sensorimotor coordination when joint attention has succeeded is relatively enhanced twice compared to the others is consistent with the mathematical proof described in the previous section.

Published Papers by the Author

This appendix lists the papers published by the author from 2001 to 2004. Parts of the publications and the presentation files for conferences are available on the author's website (`http://www.er.ams.eng.osaka-u.ac.jp/user/yukie/`).

Articles in Journals

- Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. "A constructive model for the development of joint attention." *Connection Science*, to appear, 2004.

- Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. "Emergence of Joint Attention through Bootstrap Learning based on the Mechanisms of Visual Attention and Learning with Self-evaluation." *Transactions of the Japanese Society for Artificial Intelligence (in Japanese)*, Vol. 19, No. 1, pp. 10-19, January 2004.

- Yukie Nagai, Minoru Asada, and Koh Hosoda. "Acquisition of Joint Attention by a Developmental Learning Model based on Interactions between a Robot and a Caregiver." *Transactions of the Japanese Society for Artificial Intelligence (in Japanese)*, Vol. 18, No. 2, pp. 122-130, March 2003.

Papers in Proceedings of International Conferences

- Yukie Nagai, Koh Hosoda, and Minoru Asada. "Joint Attention Emerges through Bootstrap Learning." In *Proceedings of the IEEE/RSJ International*

Conference on Intelligent Robots and Systems (IROS'03), pp. 168-173, October 2003.

- Yukie Nagai, Koh Hosoda, and Minoru Asada. "How does an infant acquire the ability of joint attention?: A Constructive Approach." In *Proceedings of the Third International Workshop on Epigenetic Robotics (EpiRob'03)*, pp. 91-98, August 2003.

- Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. "Emergence of Joint Attention based on Visual Attention and Self Learning." In *Proceedings of the 2nd International Symposium on Adaptive Motion of Animals and Machines (AMAM'03)*, SaA-II-3, March 2003.

- Minoru Asada, Yuichiro Yoshikawa, Yukie Nagai, and Koh Hosoda. "A Constructive Approach towards Emergence of Communication: Body Scheme Acquisition and Primary Joint Attention." In *Proceedings of the First International Symposium on Emergent Mechanisms of Communication (IEMC'03)*, pp. 41-48, March 2003.

- Yukie Nagai, Minoru Asada, and Koh Hosoda. "Developmental Learning Model for Joint Attention." In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'02)*, pp. 932-937, October 2002.

- Yukie Nagai, Minoru Asada, and Koh Hosoda. "A Developmental Approach Accelerates Learning of Joint Attention." In *Proceedings of the 2nd International Conference on Development and Learning (ICDL'02)*, pp. 277-282, June 2002.

Papers in Proceedings of Japanese Conferences

- Koh Hosoda, Yukie Nagai, and Minoru Asada. "Bootstrap for Emergence of Joint Attention." *Technical Report of IEICE (PRMU&NC)*, Vol. 103, No. 390, pp. 25-30, October 2003.

- Akio Morita, Yuichiro Yoshikawa, Koh Hosoda, Yukie Nagai, and Minoru Asada. "Acquisition of Joint Attention based on Self-organizing Map." In *Proceedings* *of the 21th Annual Conference of the Robotics Society of Japan*, 3B34, September 2003.

- Akio Morita, Yukie Nagai, Koh Hosoda, and Minoru Asada. "Incremental Learning of Joint Attention by Visual Feedback and Self Evaluation." In *Proceedings of JSME Conference on Robotics and Mechatronics*, 2P1-3F-C1, May 2003.

- Yukie Nagai, Minoru Asada, and Koh Hosoda. "Joint Attention Acquisition based on Synchronized Development and Learning." In *Proceedings of the 20th Annual Conference of the Robotics Society of Japan*, 3H13, October 2002.

- Yukie Nagai and Minoru Asada. "Human-Robot Communication based on Theory of Mind: Developmental Model for Shared Attention." In *Proceedings of the 19th Annual Conference of the Robotics Society of Japan*, pp. 117-118, September 2001.

Miscellaneous

- Yukie Nagai, Koh Hosoda, and Minoru Asada. "How do infants acquire the ability of joint attention?: An Approach from Cognitive Developmental Robotics." *The 3rd Annual Meeting of the Japanese Society of Baby Science*, poster presentation, May 2003.

- Yukie Nagai. "A constructive approach to understanding the development of joint attention." *The Workshop on Interactive DEsign for Adaptation (IDEA)*, invited talk, April 2003.

- Yukie Nagai, Koh Hosoda, Akio Morita, and Minoru Asada. "Bootstrap Learning into Joint Attention." *The First International Symposium on Emergent Mechanisms of Communication (IEMC'03)*, poster presentation, March 2003.