Doctoral Dissertation

# Emotion Development and Interaction based on Multimodal Predictive Coding

Takato Horii

June 2018

Graduate School of Engineering
Osaka University

Thesis Supervisor: Minoru Asada
Title: Department of Adaptive Machine Systems,
Graduate School of Engineering, Osaka University

Thesis Committee: Minoru Asada, Chair
Hiroshi Ishiguro
Ko Hosoda
Yukie Nagai

# Preface

This work has been carried out by Takato Horii from 2013 to 2017 under the supervision of Professor Minoru Asada at the Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University, Japan.

# Abstract

The aims of the study described in this dissertation are

- to construct an artificial emotion model based on knowledge from psychology and neuroscience for understanding how human emotion develops during infancy and
- to reproduce four essential capabilities for emotion interaction by the artificial emotion model apply to communication robots.

We humans feel our physical and mental states as various emotions such as delight, anger, and sadness and express them in communication with others. Emotion influences human cognitive functions and interaction between humans from the findings of psychology, neuroscience, and cognitive science. However, the developmental mechanisms of human emotion have not been elucidated. On the contrary, the characteristics of emotion have attracted the attention of robotics researchers for actualization of communication between humans and robots. Affective developmental robotics is an interdisciplinary research field focusing on artificial empathy based on the findings of psychology, neuroscience, and cognitive science applied to robotics. The approach has the potential to reveal how humans acquire the capabilities of emotion through developmental changes by constructing an artificial emotion model for robots to acquire the above capabilities through interaction with humans.

This dissertation attempts to reproduce the developmental process of human emotions and essential capabilities for emotional interaction by proposing an artificial emotion model based on the idea of predictive coding from the viewpoint of affective developmental robotics. In revealing the development of human emotion, what capabilities are important for acquiring emotion in infancy should be considered. Furthermore, in realizing the emotional human{robot interaction, how capabilities of the

predictive coding model contribute to reproducing essential abilities for the realization of emotional interaction should be validated. Focusing on these points, the following two approaches are executed.

1. The first approach is a constitutive method focusing on the developmental process of human emotion based on neuroscience and cognitive science findings. We propose the emotion model based on the idea of predictive coding with two hypotheses that are a peculiarity of tactile sensation and perceptual improvement of sensation during infancy from findings of neuroscience and cognitive science. The proposed model is evaluated with and without our hypotheses in a virtual infant{caregiver interaction environment. Experimental results show that tactile dominance and perceptual improvement facilitate the development of emotion perception when both mechanisms are integrated into the proposed model.

2. The second approach attempts to reproduce abilities for emotional interaction by using the capabilities of a predictive coding model. The proposed model is constructed by a generative model, which can infer latent signals from observed signals and their inverse. Additionally, it can predict unobserved signals from observed ones through the above inference process. We evaluate the abilities of the model regarding the expression of emotional signals from the robot, estimation of others' emotional states, and active perception based on the predictive information in human{robot interactions. Experimental results demonstrate that the capabilities of the predictive coding model enable the robot to improve emotional interaction with humans.

It is expected that the proposed models incorporate the knowledge from psychology, neuroscience, and cognitive science, and provide suggestions for understanding how human emotions develops in infancy and how humans interact emotionally.

# Acknowledgments

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

First of all, I would like to express my gratitude to Professor Minoru Asada, who is a principal adviser of this work. He has given me many chances to work on high-level studies and make connections with many great researchers. The new research field of "Emotional Developmental Robotics," which I am interested in and tackling now, was advocated by him. His critical comments and great supports have enabled me to accomplish this work.

Dr. Yukie Nagai has given me detailed suggestions based on her work and ideas. Her valuable advice has helped me shape the way I think and enabled me to fulfill my Ph.D. work and this dissertation. She also has given me many chances to become a fully-fledged researcher. I would like to extend my gratitude to her.

Professor Hiroshi Ishiguro and Professor Koh Hosoda have spared their time to discuss the meanings of the research on cognitive developmental robotics. They have made me reacknowledge the difficulties and the significance to study cognitive science from the viewpoint of robotics. Their valuable comments give me a chance to reaffirm my position on this study. I am very grateful to them.

I also thank all the present and the past members in Emergent Robotics Laboratory at Osaka University.

I express my gratitude to Dr. Hisashi Ishihara and Dr. Yuki Sasamoto. Dr. Hisashi Ishihara has given me essential advice on how to discover interesting topics of study. Dr. Yuki Sasamoto has taught me the importance of understanding numerical expressions of computational models.

I would like to thank Dr. Yuji Kawai and Mr. Jihoon Park. They have discussed

research topics and ideas with me, and their supports have enabled me to think deeply about my study. I would like to lead the research area of cognitive developmental robotics with you.

I am grateful to Mr. Takumi Kawasetsu. He has spent his time for meaningful discussions with me about the study on flexible tactile sensors. I thank Mr. Ryo Iwaki, Mr. Jorge Luis Copete, and Mr. Jyh-Jong Hsieh. Other members who are not mentioned here also have supported me to accomplish my study.

Finally, I would like to express my appreciation to my family. My parents, Hiroyuki and Etsuko, have supported me to spend a (so) long student life. They have given me a chance to do whatever I like. None of this work and my life would be possible without their great love. Thank you so much.

June 1st, 2018
Takato Horii

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Humans feel their own physical and mental conditions as emotional states. Our feelings of emotion are sometimes influenced by a context, and sometimes we experience a change of emotions based on our own body reactions. We also express such emotional states through various modalities such as voluntary body gestures and involuntary changes in heart rate. For example, when our behavior is disturbed by someone, our heart pumps faster, we feel anger, and we grimace. However, if the cause comes from friends, we might forbear expressing an angry face. When we achieve goals, our heart might pump faster also; nevertheless, we feel a different emotion than in the previous situation and may smile.

The study of emotion has recently expanded in many research areas (e.g., psychology, cognitive science, neuroscience, and robotics). According to cognitive and neuroscience approaches, emotion influences the cognitive functions of individual humans (e.g., changing perception and action, modulating memory and learning performance, and modifying decision making [3]). On the other hand, emotion has been defined in many ways. For instance, James [4] simply defined that emotions are perceptions of bodily state. Damasio [5] considered body responses with a value of stimuli processed in the limbic system of the brain. Oatley and Johnson-Laird [6] reviewed cognitive theories of emotion in terms of evaluation, which is one of the cognitive functions in the human brain. The concept of evaluation is critical in everyday life because it allows us to learn how events change our physical and mental states. Then, we can predict causes of our own perceptions and the results of our actions in the world and

can select behaviors to avoid or bring about different events.

Related to the emotion mechanism of evaluation and prediction, some researchers in neuroscience and cognitive developmental robotics have focused on the idea of "predictive coding" for explaining and modeling the brain mechanisms of human cognition [7{ 12]. Predictive coding theory assumes that the fundamental mechanism of the human brain is to minimize a prediction error the difference between what we expect and what we experience. Computational models based on predictive coding have been proposed, and they allow us to understand the human brain mechanisms from a constructive approach viewpoint [7, 13]. Seth [14{ 16] attempts to explain the process of emotional cognition based on the idea of predictive coding. He considers multimodal sensory signals, which involve not only exteroceptive and proprioceptive perception but also interoceptive perception related to physiological homeostasis (i.e., embodiment signals).

Emotion plays important roles in not only the cognitive functions of individuals but also interactions between humans. Humans often attempt to communicate their several emotional states with others through their gestures, facial expressions, and vocalizations. They also recognize others' emotions from those multimodal signals. Furthermore, we can estimate others' emotions from partial observation of expressions by predicting them based on own experiences and choose an action to acquire any missing information from others for updating the estimation. We believe that these four mechanisms (i.e., perception, expression, estimation, and active inference) are essential capabilities for emotional interaction among humans.

Up to the present, we have described the roles of emotion in cognitive functions of individual humans and interactions among humans. However, two questions remain.

- how various types of human emotion develop during infancy, and
- how humans gain the capabilities of emotional interaction such as emotion perception, emotion expression, emotion estimation, and active perception.

This study addresses these two issues from the viewpoint of affective developmental robotics. We first propose an artificial emotion model for reproducing emotion development. Then, we apply the proposed model to robots, which interact with humans through processing of multimodal signals, to acquire the four capabilities

of emotional interaction and evaluate the model capabilities. The idea of predictive coding is employed as a key mechanism in our artificial emotion model to address the above issues. The model processes multimodal emotional signals in face-to-face interactions between humans and robots to reproduce development of emotions (emotion development for short) and to realize the four capabilities of emotional interaction.

This chapter first describes basic knowledge of emotion development and interaction in terms of cognitive and neuroscience research and modeling approaches. From the perspective of cognitive and neuroscience research, we first explain the developmental theory of human emotion, the connections between mirror neuron systems and emotion, and the essential capabilities for emotional interaction between humans. Then, the robotics approaches for the artificial emotion are described. In Section 1.3, we explain the idea of predictive coding, which is expected to explain human cognition, as a key mechanism for our models. Finally, the objective and the organization of this dissertation are provided.

## 1.1 Background

In this section, we explain theories on the development of emotion from psychology and cognitive science studies. The brain regions regarding emotion processing and the mirror neuron system, which represents a relevant mechanism to realize emotional interaction between humans, are introduced. Then, we describe the definitions of the essential capabilities for emotional interaction used in this dissertation. Finally, we introduce robotics approaches regarding artificial emotion and emotional interaction between humans and robots.

### 1.1.1 Human Emotion

#### Psychological Findings of Emotion

Humans recognize their own physical and mental states as various types of emotions such as delight, anger, sorrow, pleasure, and so on. We previously explained that ones' perception of emotion is influenced by different situations, and humans can regulate their own emotions to some extent. Humans also express their emotions

through various modalities such as facial expressions, gestures, vocalizations, etc. In social context, humans share their emotions in communication. In other words, people perceive others' emotions from others' expressions and react accordingly. However, it remains unclear how humans acquire these generalized emotional states as typified by happiness, anger, and so on.

Theories on the development of emotion have adhered to the classical dichotomy of nature versus nurture. Tomkins [17] and Izard [18] both hypothesized that emotional states are innate. Conversely, Bridges [1], Sroufe [19], and Lewis [20] all speculated that basic emotions are acquired through a developmental process from infancy to childhood, which they attempted to support through observing infants' behavior. In this dissertation, we suppose that emotions differentiate from primitive internal states such as pleasure and displeasure to detailed emotions involving six basic emotions. We discuss any other developmental studies regarding emotions in psychology and neuroscience, in Section 2.

### Neurological Findings of Emotion and Mirror Neuron System

Damasio [5] proposed the somatic marker hypothesis where emotion is associated with body responses to stimuli, and emotional processes guide one's behaviors. Namely, emotion works as an evaluation system for external (and internal) stimuli. Within the human brain, the evaluation mechanism is processed in the ventromedial prefrontal cortex (VMPFC) and the amygdala. The amygdala is well known as a human brain region regarding emotion processing and is a part of the limbic system [3]. It processes low-level sensory signals corresponding to appraisal of stimuli, especially negative emotion. The insular cortex is also a part of the limbic system and monitors sensory signals such as exteroceptive and interoceptive signals [21]. Adding to the above regions, it is known that the superior temporal sulcus (STS) integrates visual, audio, and tactile signals [22] and engages in multimodal information processing for emotion perception [23, 24]. The superior temporal gyrus (STG), which is near the STS, also responds to various non-verbal emotional stimuli [25], and the temporal area of infants' brains perceives and reacts to audiovisual emotional stimuli [26].

In social contexts, people perceive others' emotions from multimodal signals and

may share an emotion. Humans' physical and mental states, which are base of emotion, are very complex; however, we can share generalized emotional categories as typified by delight, anger, and so on with each other. A mirror neuron discovered by Rizzolatti et al. [27] is a single neuron in the brain of the macaque monkey that activates both when the macaque performs and when it observes the same action being performed by another. In other words, the neuron mirrors the others' behavior as though the observer were the one performing the behavior. Brain imaging experiments have shown that the human brain also has similar mechanisms in the inferior frontal cortex and superior parietal lobe. These brain regions perform the function of mirror neurons, and they have been defined as the mirror neuron system (MNS) [28{31] Interestingly, the regions of MNS and emotion processing overlap, and it is known that the mechanism of MNS enables people to imagine the emotional state of others based on their own experiences and of expressing the corresponding emotion in the context of emotional interaction [29, 32, 33]. The emotional states exchanged in the communication produce empathy in others based on the function of the MNS, and they can induce and maintain relationships such as happy cooperation or angry conflict.

### 1.1.2 Artificial Emotion

Sharing emotion with partners is an essential skill in social communication between humans. The skill is even more important for the social robots interacting with humans. In attempts to achieve emotional interaction, several empathic robots have been developed [34{46]. Kismet [41, 42] is one of the most popular social robots to have established emotional interaction with humans. Kismet has an artificial emotion system that is able to extract features corresponding to three affective values (specifically arousal, valence, and stance) from humans' speech and recognize one's emotional state. Then, Kismet generates a facial expression by interpolating preprogrammed expression prototypes based on the recognized emotional state of others. However, the empathic behaviors of many emotional robots are often preprogrammed by a designer and have been made to address specific contexts [47]. Therefore, it would be difficult for Kismet to understand the capabilities regarding emotional interaction

and to generalize them in different situations.

To address the above issues, Asada [48,49] advocate a new idea, "Affective Developmental Robotics", which extends the cognitive developmental robotics approach. This research approach connects neuroscience and biobehavioral studies to robotics studies more strongly than the cognitive developmental robotics approach. He proposed a conceptual model of artificial empathy based on an idea associated with embodiment, development of self{other cognition (including MNS), and social interaction. The model tries to explain the developmental process of artificial empathy from emotional contagion (a low-level phenomenon) to sympathy (a high-level cognition). An example of an artificial emotion model based on this approach was proposed by Watanabe et al. [50]. Their model acquired the relationships between its own emotional states and the facial expressions of interaction partners via face-to-face interaction. We suppose the approach of affective developmental robotics in this dissertation to propose our artificial emotion model for reproducing emotion development and humans' focused capabilities for emotional interaction.

## 1.2 Four Essentials for Emotional Interaction

Humans have many cognitive capabilities to participate in emotional interactions with each other. We focus on four capabilities that are essential for emotional interaction and describe them as follows.

Emotion Perception

: The capability of recognizing the emotions of others. This capability is necessary to recognize the emotional states of others from their expressions. This capability is basic for emotion estimation and sharing emotion in interactions. It is known that this capability is acquired via the developmental process from infancy to childhood [51{54].

Emotion Expression

: The capability of showing one's emotional state by generating signals. This capability is important to tell others what your emotional state is. We can share and transmit emotional states with each other through any signals based on

both emotion perception and expression capabilities. Some researchers hypothesize that the expression capability is innate and preprogrammed [17, 18, 55]. On the other hand, other behavioral studies show evidence that the variations of emotional expressions increase during infancy [1, 19, 20]. We support the developmental theories of emotion expression. In this dissertation, we focus on facial expressions, vocalizations, skin contact, and gestures as signals of multimodal emotional expressions.

Emotion Estimation

The capability of understanding others' emotions from partial observation based on our own experience. In practical interactions, a few parts of others' emotional expressions are lacking; for instance, the partner is paying attention to others. However, humans are able to estimate others' emotion from not only complete observations of others' expressions but also partial observations by imagining uncertain signals from observed ones based on experience. This capability is called mental simulation and corresponds to MNS and mentalizing systems [29, 33].

Active Perception

The capability of selecting our own actions to obtain additional and effective information to update an estimation. This last capability, active perception, is when humans select their own behaviors (e.g., expression of emotion, asking questions, and so on.) to acquire effective information from others for updating and improving the estimation of the other's emotion. During interactions, humans try to estimate others' emotions by observing their expressions. However, it is not always true that all modality signals link to the actual state and complete at the same time. Some signals might be noisy or ambiguous. Then, we should collect missing information from others by selecting our own actions to bring out effective signals for the estimation.

## 1.3 Predictive Coding: Modeling Approach of Human Cognition

The original idea of predictive coding was proposed as the concept of unconscious inference by Helmholtz. Unconscious inference considers the interaction between sensory stimuli and conceptual knowledge, and these bottom-up and top-down processes influence each other. Recently, the idea of bottom-up and top-down processing was adopted into a computational modeling approach as a key idea to replicate the visual processing mechanisms in the human brain [7], modeling the dynamic interactions between the human brain and body [12], and so on. Nagai and Asada [11] applied predictive coding to reproduce the developmental process of cognitive abilities (e.g., such as self{other cognition, estimation of others' goals, and the emergence of helping behaviors). They also attempted to explain the characteristics of autism spectrum disorder based on the theory. From both bottom-up and top-down processing, we can calculate the prediction error, which is the difference between expected perception from the inner model by top-down process and actual perception from the bottom-up process. The computational models of the above modeling studies attempt to minimize this prediction error via updating the model parameters.

Friston, a neuroscientist, focused on the idea of predictive coding as a fundamental mechanism of the human brain. He extended it by including a Bayesian inference and proposed the "Free Energy Principle" [8, 9]. The free energy (named from an analogy with thermodynamics) is the sum of the prediction errors in the entire brain. From the viewpoint of Bayesian inference, the prediction error is represented by the divergence between distributions of prior knowledge and perceptions [56]. Friston assumes that the free energy principle is a unique mechanism in the human brain. Seth, a colleague of Friston's extended the free energy principle from modeling perception (in Seth's paper, it is called proprioception and exteroception) to interoception. He proposed that human subjective feeling known as emotion is generated by predictive systems that are built actively from causal interoceptive appraisals. This idea, which assumes that emotion is an evaluation process or mechanism of embodiment reaction associated with stimuli, has affinities with the cognitive approaches to study emotion. It is known that the interoceptive signals also relate to evaluation signals linked with

embodiment and self-awareness. Furthermore, the idea of the free energy principle (minimization) also has affinities with computer science, especially generative models. The generative model is a model for generating data that can be observed and includes latent variables that can only be computed from observed data. In other words, the generative model represents relationships among observations and latent variables. Many types of generative models such as the hidden Markov model, latent Dirichlet allocation [57], restricted Boltzmann machine [58], and generative adversarial networks [59] have been proposed in research fields of computer science. However, the objectives of those generative models are the same in terms of minimizing the divergence between a generative probability of data in the actual world and an approximated probability via training of the generative model using observed data (i.e., experience). This objective of generative models is also similar to the process of prediction error minimization in predictive coding. Accordingly, Friston pointed out the relationship between the human brain and the generative model [8, 9].

The restricted Boltzmann machine, which is one type of artificial neural network, attracted the attention of many researchers in the machine learning field because the model broke through an important barrier, the vanishing gradient problem, of neural networks by adopting layer-wise pre-training [60]. On the other hand, we view the different characteristics of the restricted Boltzmann machine as a generative model. In the training phase, the restricted Boltzmann machine minimizes the cross entropy between data probability and the probability of reconstructed data from hidden neuron activations. This training method corresponds to the minimization of the free energy of the network and also relates to the idea of predictive coding. In our study, we employ the restricted Boltzmann machine as a fundamental module in our model to reproduce emotion development and the capabilities in emotional interaction. In later sections, we describe the details of the restricted Boltzmann machine and the proposed models to address our objectives.

## 1.4 Overview

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

- to construct an artificial emotion model based on knowledge from psychology and neuroscience to understand how human emotion develops during infancy, and

- to reproduce four essential capabilities for emotional interaction with an artificial emotion model for communication robots.

To address these aims, we adopt the approach of affective developmental robotics in this dissertation. The key idea in our study is employing the predictive coding mechanism for constructing our artificial emotion model.

This dissertation consists of seven chapters including this one. The outlines of all chapters are as follows:

Chapter 1. Introduction

The grand challenge of this study is to understand human emotion as well as to realize empathic robots from a viewpoint of cognitive and affective developmental robotics. In this chapter, first we described current knowledge regarding emotion development and interaction in terms of psychology, cognitive science, neuroscience, and affective developmental robotics. Then, the basic idea of predictive coding and modeling studies has been explained. We introduced the reason why we employ the restricted Boltzmann machine as a fundamental model for our artificial emotion model.

Chapter 2. Related Work

The findings of emotion development and our focused capabilities for emotional interaction from psychology, cognitive science, neuroscience are described. We review knowledge regarding emotion development from the perspectives of emotion perception and emotion expression. Then, this chapter introduces previous studies that attempt to understand and mimic human emotion by constructing computational models and realizing emotional human{robot interactions with artificial emotion models. These studies are compared with our study.

*Chapter 3. Our Approach to Modeling Artificial Emotion*

The task definitions for reproducing emotion development and evaluating our

focused capabilities for emotional interaction are given. We focus on face-to-face emotional interaction between an infant and a caregiver or interaction between humans and robots. In both interactions, we assume that humans and robots perceive and express emotional states through multimodal signals. Subsequently, this chapter explains the restricted Boltzmann machine as the key mechanism of our artificial emotion model. The types of characteristics that correspond to predictive coding are described.

Chapter 4. Modeling Development of Multimodal Emotion Perception

Our first computational model based on predictive coding for reproducing the development of emotion perception is presented. We hypothesize that the capability of emotion perception is acquired owing to two important functions of infants: tactile dominance and perceptual improvement. The proposed model is evaluated in multimodal interaction situations assuming face-to-face infant{ caregiver interaction. Experimental results show how both functions facilitate the development of emotion perception from multimodal signals.

Chapter 5. Emotion Expression and Estimation by Mental Simulation

The second computational model is introduced. This model is also based on the idea of predictive coding. We utilize the restricted Boltzmann machine, which is the generative model for reproducing the capabilities of emotion expression and estimation. In particular, the estimation mechanism corresponds to the mental simulation of the MNS. Experimental results demonstrate that a robot using the proposed model can generate emotion expressions and can estimate human emotion from partial multimodal signals.

Chapter 6. Active Perception based on Energy Minimization

The second proposed model is used again to reproduce the capability of active perception. The model controls the robot's attention to humans' multimodal signals based on an energy minimization criterion that corresponds to the free energy principle. It is known that the energy values of the restricted Boltzmann machine reflect the amount of experiences. Therefore, the method enables the robot to select the modality that provides the lowest energy to exploit more

frequent experiences. The experimental results demonstrate the effectiveness of the proposed model in emotion estimation during human{robot interaction.

Chapter 7. Conclusion

Finally, we conclude this study and describe future work. The proposed emotion models based on the idea of predictive coding with our two hypothesis first reproduce the development of emotion perception through infant{caregiver interaction. This model enables us to verify knowledge from psychology and neuroscience studies. Furthermore, the second similar model handling multimodal signals in emotional interaction with humans demonstrates its capabilities in emotional interactions. However, several problems remain to be solved, e.g., generating continuous expressions (actions) and executing exploratory behavior during an actual interaction, and so on. These problems are discussed as future work.

# Chapter 2

# Related Work

Emotion is an interesting research topic in psychology, neuroscience, cognitive science, and so on. The researchers in these fields have investigated emotion from diversified viewpoints and have made many findings. Especially, the developmental process of emotion and its capabilities have been studied based on behavioral approaches. Furthermore, these findings have motivated robotics researchers to build empathic robots that interact with humans and to understand the mechanism of emotion by constructing artificial systems.

This chapter presents the knowledge regarding the developmental process and capabilities of emotion and reviews robotics and computational science approaches to develop and model emotion systems for realizing human{agent interactions. First, we introduce two points of view on the structure of emotions: discrete and continuous emotion. These two ideas relate to the theories of emotion development (generation). One argues that emotional states are innate and the other that emotions are acquired via development. We concur with the developmental theory of emotion and describe the developmental changes regarding emotional capabilities. Then, this chapter describes two robotics approaches, called "Cognitive Developmental Robotics" and "Affective Developmental Robotics," to build social and developmental robots based on the knowledge gained from the above research fields.

## 2.1 Findings on Emotion in Psychology and Cognitive Science

We humans, recognize our physical and mental states as various emotions such as delight, anger, sorrow, and pleasure and also express them through our own multimodal expressions (e.g., facial expressions, gestures, and vocalizations). For example, when we face a snake we might feel fear and exhibit avoidance behavior (e.g., run). When we recognize that the snake is a rope in reality, we experience a feeling of relief.

Categorical emotion theory assumes that there is an assembly of discrete emotions which constitutes the fundamental emotion. Ekman [55] defined discrete categories of emotion called six basic emotions: joy, surprise, anger, fear, sadness, and disgust. The definition of the basic emotions is based on analysis of facial expressions among different cultures. He claimed that the six basic emotions are evolutionarily common because people who belong to different cultures can recognize emotional facial expressions from other cultures. On the other hand, Russell [61–63] studied the relationships among various emotional states and proposed a two-dimensional space called the circumplex model, and developed a core affect theory. His model and theory assumed that underlying any emotion is a state with two dimensions: pleasure versus displeasure and level of arousal. Emotional categories such as anger and fear are not distinct and not evolutionarily common. They overlap in continuous space as prototypes of emotion.

### 2.1.1 Emotion Development

Adult humans are able to recognize and express several emotional states. However, it remains unclear how these emotional states are acquired. Theories on the development of emotion have adhered to the classical dichotomy of nature versus nurture. From an evolutionary perspective, Tomkins [17] and Izard [18] both hypothesized that emotional states are innate, which is the basis for their discrete emotion theory. Frijda et al. [64] also considered that the basic emotions are elementary response patterns or action tendencies. They claimed that infants have fixed emotional states corresponding to phenomena such as facial expressions and bodily reactions.

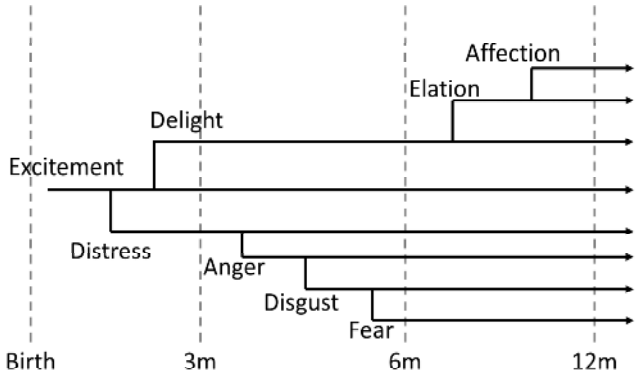



Figure 2.1: Bridges suggested a genetic theory of emotion by observing behavioral changes in infants [1]. These branches also correspond to a developmental process for an expression capability of emotions in infants.

Conversely, Bridges [1], Sroufe [19], and Lewis [20] all speculated that basic emotions are acquired through a developmental process from infancy to childhood, which they attempted to support through observing infants' behaviors. According to Bridges' study, infants' only internal state is general excitement. Namely, it corresponds to self-awareness for received stimuli. This primary state initially differentiates into delight and distress in response to positive and negative stimuli, respectively. After this first differentiation, the category of positive responses can be split into elation and affection, whereas the category of negative reactions can be subdivided into anger, disgust, and fear (see Figure 2.1). This figure also illustrates the developmental process of the capability for emotional expression during infancy.

To understand the developmental changes in emotion perception, several researchers displayed visual, audio, or audiovisual emotional expressions of others to infants and young children (e.g., [51{54]). Walker-Andrews [51] argued that human infants have only rudimentary capacities to detect, discriminate, and recognize (or perceive) others' emotional expressions at birth; however, the capabilities rapidly develop during the first year of life. She reviewed many articles pertaining to infants' capabilities of emotion perception and suggested that younger infants (around five months old) can detect only primary meanings in the emotional expressions of others. On

the other hand, older infants can discriminate and recognize various emotional categories of others' expressions (e.g., basic emotions) owing to perceptual development. Grossmann [54] reported that 12-month-old infants showed different reactions to an event-related potential in their brain when they faced angry and happy audiovisual stimuli (i.e., facial and vocal emotional expressions of others). We consider that there is more evidence in support of the developmental process of emotion perception in particular findings in that tactile interactions between infants and their caregivers appear to affect the ability of emotion perception [65–67]. However, these studies only observed changes in behavior when the subjects faced emotional stimuli. Thus, it remains unclear as to what causes the developmental changes in emotion perception because there are no methods to represent the actual emotional states experienced by infants.

## 2.2 Robotics Approaches to Modeling Emotion

### 2.2.1 Cognitive Developmental and Affective Developmental Robotics Approaches to Understanding Emotion

"Cognitive Developmental Robotics" is a research field aimed at understanding the mechanisms of the cognitive development process by synthetic approaches utilizing human-like robots and computer simulations [68]. Recently, Asada extends the idea of cognitive developmental robotics and advocates "Affective Developmental Robotics". Affective developmental robotics attempts to connect biobehavioral and neuroscience studies to robotics studies to extend cognitive developmental robotics to elucidate the developmental process of human empathy and to build artificial empathy in robots. A number of researchers in robotics have constructed emotional systems and reproduced the abilities and developmental process of emotion (e.g., [35, 38, 41, 44, 44, 50, 52, 69{ 73]), as discussed below.

### Robotics Approaches to Reproducing Emotion Capabilities

Blanchard and Canamero [70] proposed a general Perception-Action architecture, which involves imprinting experiences and reward-based learning methods. Their experimental results on simple human{robot interaction showed that the proposed model acquired affective behavior, driven by the parameters of the model (e.g., "distress" caused by a difference between the current and imprinted experiences, and "comfort" caused by reward stimuli). Hiolle et al. [72] and Lones et al. [73] examined influences of a novel environment and different experiences on the arousal regulation and behavior learning using agents. These studies evaluated how robots' behaviors and model parameters (e.g., arousal and comfort) differentiate through the interaction between robots and their environment.

### Preprogramed Systems for Realizing Emotion Interaction

Kismet [41, 42] is one of the most popular social robots to have established emotional communication with humans. The Kismet system extracts features corresponding to three affective values (specifically arousal, valence, and stance) from human speech and then generates a facial expression by interpolating preprogrammed expression prototypes. The prototypes of facial expression sit at the extremes of each component of affective information. These prototypes represent high to low arousal, positive and negative valence, and open and closed stance. In addition to these prototypes, three facial prototypes are used to distinguish the expressions for disgust, anger, and fear. The expression system calculates weights for each prototype based on affective information to generate a current facial expression as a linear combination of all prototypes. This interpolation approach was able to generate continuous expressions in different forms [38] and the recognition rate of Kismet's six basic emotional expression ranged from 47% to 83%.

Trovato et al. [38] and Kishi et al. [39] developed an emotional model for a humanoid robot named KOBIAN, based on psychological studies. Their model represented KOBIAN's internal state, which is modulated by external stimuli. It also included prototypes of facial expressions grounded on specific emotional states and expressed facial patterns as combinations of these prototypes [74].

Ogata et al. [69] developed an emotional robot that communicates with humans by modeling human hormone systems. Their robot, WAMOEBA-2, modulates its own emotional expressions and actions based on external stimuli and the homeostasis activation of virtual hormones. WAMOEBA-2 is adaptive to external stimuli to maintain a stable body. Therefore, the best behavior is sleeping in order to minimize energy reduction unless external signals arise. As a result, it seemed to actually avoid interaction with humans.

Watanabe et al. [50] built a mapping system for emotional states based on intuitive parenting observed during infant{caregiver interactions. They assumed that caregivers imitate the facial expressions of infants when infants are expressing a particular emotion. The internal state of the robot and facial expressions are defined using psychological studies and change dynamically in response to stimuli. Their model learned the relationships between its own emotional states and the facial expressions of interaction partners when its emotional states were changed through external stimuli. After training, the model was able to modulate its own emotional states according to not only the type of stimuli but also the facial expressions of the caregivers.

Lim and Okuno [44, 45] proposed multimodal emotional intelligence (MEI), which utilizes an integrated architecture to recognize the emotional states of others and generate its own emotional expressions. Their model was inspired by MNS, a mechanism underlying human cognition. MEI is composed of Gaussian mixture models (GMMs) to realize both recognition and generation in the same architecture. Recognition of the emotional states of others is represented by classification of input features by GMMs, whereas expressions of one's own emotional states are achieved by sampling features from selected Gaussian distributions corresponding to a specific state. An important characteristic of their model is that it computes four features (Speed, Intensity, irRegulation, and Extent (SIRE)) assumed to be very common among modalities. Therefore, following training from speech using the SIRE features, MEI was able to estimate categories of emotion from not only audio signals but also gait signals.

### 2.2.2 Other Approaches

Breazeal et al. [40] presented a creature robot called Leonardo that can imitate humans' facial expressions. Leonardo learns a direct mapping between a person's facial expression and its expression by using a neural network. Andra and Robinson [34] developed an android head robot that mimicked the facial expressions of humans with the aim of social-emotional intervention for autistic children. Their robot tracked facial feature points of subjects who expressed emotional states and directly converted them into corresponding control points to modify its own facial expression. Riek et al. [75, 76] developed a head gesture mimicry robot with the aim of eliciting empathic behavior from humans. They tracked a subject head movement and the robot executed an appropriate movement in response. However, the direct mapping of human expressions may lead to misalignment of emotional states. For example, humans may show a tearful face when crying with delight. Further, their expressions vary depending on context. Consequently mapping only facial expression (i.e., crying) can result in miscommunication of the emotional state (i.e., happiness). Therefore, it is better for robot systems to estimate the emotional states of communication partners and generate expressions based on the estimated states. Kanoh et al. [36, 37] proposed a continuous facial expression system called "iFbot". The system used neural networks to abstract prototypes of iFbot's facial expression patterns and to acquire their relationships. Matsui et al. [35] proposed a recurrent neural network that generates facial expressions from external stimuli through simulated emotional states. An anthropomorphic robot called BARTHOC is capable of recognizing human' emotion from speech and producing facial expressions corresponding to the six basic emotions [43].

For emotional recognition, many studies focused on facial, vocal, and audiovisual information to estimate humans' emotional states [77{80]. Wang et al. [78] used audiovisual signals to recognize emotional states. They compared the recognition precision to models such as the Gaussian Mixture Model, k-nearest Neighbors, Neural Network, and Fishier's Linear Discriminant Analysis with a feature selection method named the stepwise method. The stepwise method starts with only one feature from

audiovisual signals and progressively adds features. For each step in the selection process, one feature is added or removed from a subset of selected features to maximize the between-class Mahalanobis distance. The method increased the performance of emotional recognition from 70.0% which used the original feature, set to 75.7%. This result shows that there exists an efficient feature subset of emotional expressions to estimate emotional states. Mower et al. [79] also selected a set of emotional features from audiovisual signals by using information gain. This selection method was able to assign weights to feature sets for recognition of each emotional state. For instance, from their experiments, eyebrow movements were important for happy and neutral states while a sad emotion was conveyed more obviously by audio signals. They assumed that emotional expressions are composed of multi-emotional categories. This assumption of emotion profiles enabled them to consider the major and minor emotions of expression for classification. They estimated combinations of emotional states from human{human audiovisual interactions by using a support vector machine with margins. The recognition system performed well in estimating minor voted emotional states.

## 2.3 Summary

This chapter has described knowledge regarding the developmental process and capabilities of emotion. We also reviewed robotics approaches to understand the mechanisms of human emotion and build empathic robots. Studies in psychology, neuroscience, and cognitive science have produced many findings on emotion development and emotion capabilities. Theories on emotion development continue to be discussed in many research fields; however, in this dissertation, we support the developmental theory of emotion. Moreover, several robotics studies were inspired by these findings and the proposed emotional systems and/or robots that can develop emotional behaviors through interactions with their environments.

The next chapter introduces the assumed tasks and a key mechanism in our emotion model in this dissertation. The proposed models in later experiments are based on the findings described in this chapter.

# Chapter 3

# Our Approach for Modeling Artificial Emotion

This chapter first describes our assumed interaction between humans and robots and the task settings to reproduce emotion development and realization of four essential capabilities for emotional interaction (i.e., emotion perception, emotion expression, emotion estimation, and active perception). We assume that robots (infants) interact with humans (caregivers) by using multimodal signals with each other during face-to-face interaction. In this dissertation, the proposed emotion models with a robot are validated in the interaction situation, and the models enable the robot to acquire four capabilities for emotional interaction. In particular, the capability of emotion perception is evaluated in the infant–caregiver interaction to elucidate our hypothesis on emotion development based on psychology and neuroscience findings. Other capabilities that are reproduced by the proposed models are evaluated in human–robot interaction.

Subsequently, we explain the details of the restricted Boltzmann machine, which is a key component, to incorporate the idea of predictive cording with our approaches. The restricted Boltzmann machine is a generative model and is also a type of artificial neural network that consists of two layers. It can encode an input signal into latent information and decode the information into an original signal, and both functions are based on the learning algorithm applied for predictive coding. This chapter presents the basic idea for constructing artificial emotion models to address

our research objectives.

## 3.1 Assumptions and Task Definition

In this section, we explain the interaction that is assumed in this dissertation. Humans and robots (or infants and caregivers) can interact and collaborate with each other in many situations. However, we focus on face-to-face interaction in this dissertation. Figure 3.1 shows our intended interaction between the human and the robot. In this situation, we assume that the human and the robot use multimodal signals, such as facial expressions, speech, gestures, and touch to interact with each other. The main aim of this interaction is to exchange emotions between the human and the robot.

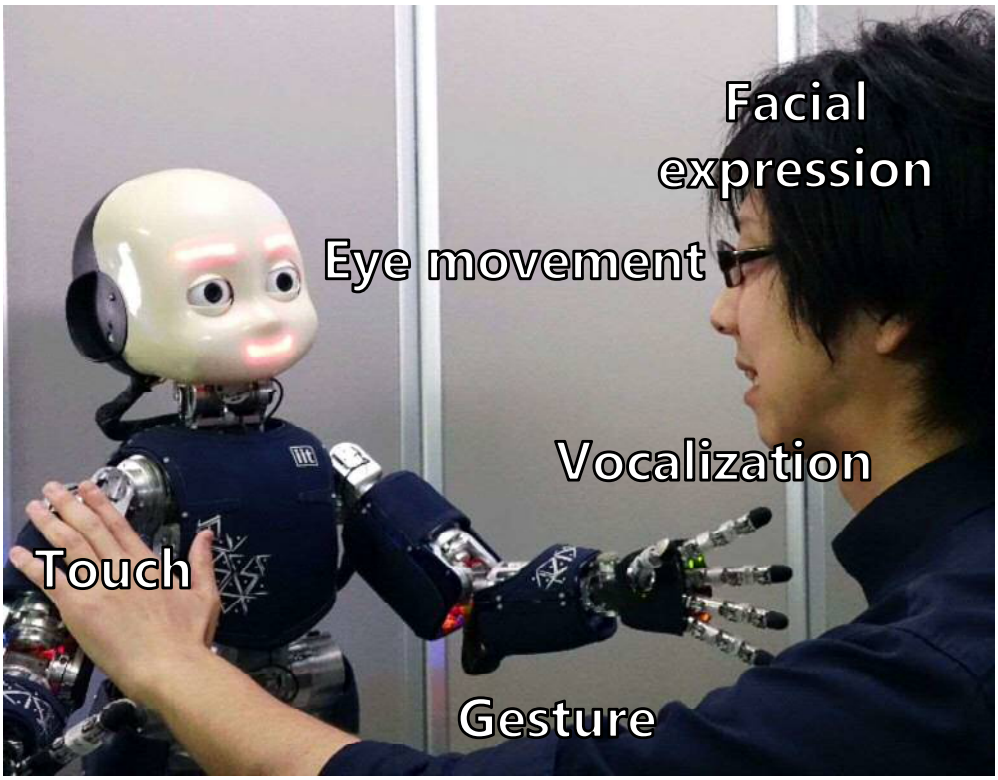



Figure 3.1: Overview of the assumed multimodal human{robot interaction

Figure 3.2 illustrates more detailed tasks that realize emotional interaction. We

describe the objectives of these interaction from the robot's perspective through correlations with the four emotion capabilities of interaction discussed in Chapter 1.

### First task: Emotion Perception

The first task in emotional interaction is to perceive the human's emotion. This task is defined as Emotion Perception in Chapter 1. The robot receives a human's complete multimodal expressions and recognizes his/her emotional state from them. To acquire this capability is the first step to conduct emotional interactions with humans.

We address this issue (to acquire the capability of emotion perception) by constructing an artificial neural network that considers psychological and neuroscience findings regarding emotion development, as described in Section 2.1.1. We focus on infant–caregiver interactions as an experimental interaction setting. The proposed model is evaluated from two perspectives. One is the possibility of the predictive coding model to acquire emotion capabilities. The other is determining which factor(s) is important to reproduce the development of emotion perception. This approach is discussed in Chapter 4.

### Second task: Emotion Expression

The second task is to express emotional signals to the partner. This task is defined as Emotion Expression The robot shows multimodal emotional expressions based on its emotional state. This capability is important for the human to understand what the robot's current emotional state is. It is necessary to continue an interaction because the human gets tired and stops the interaction if the robot does not react to him/her.

To acquire this capability, a similar model as discussed in Chapter 4 is employed for the robot. However, we focus on the adult human–robot interaction in later chapters. The robot learns how to express emotional states based on the expressions of the interaction partner. This approach is discussed in Chapter 5.

### Third task: Emotion Estimation

The third task is emotion estimation. In a practical interaction, sometimes, human expressions are no fully displayed and the robot receives only partial signals. It is difficult to recognize the correct emotional state of the human in this case; however, the robot should estimate the partner's emotion based on its own knowledge. For emotion estimation, the robot executes a mental simulation based on its own knowledge to compensate for missing signals. Actually, the capabilities of emotion perception and emotion estimation are similar, although we separate these functions here. In this dissertation, we define that emotion perception is the capability to recognize others' emotions based on complete information (i.e., full multimodal signals used for model training in this study). In contrast, emotion estimation is defined as the recognition process based on incomplete information (e.g, only the facial expression or reconstructed signals based on simulation) and where the estimated emotion includes uncertainty.

To realize the emotion estimation capability, we employ the same model as for emotion expression. One characteristic of the restricted Boltzmann machine is utilized for executing the mental simulation to update the estimation of a partner's emotion in the interaction. This approach is discussed in Chapter 5 with the result of emotion expression.

### Fourth task: Active Perception

The fourth task is active perception. When the robot receives partial multimodal signals from the human, the robot should select its own action to obtain additional information from the interaction partner to renew its estimation. The robot not only imagines lacked signals by mental simulation but also gains them using its own actions. This capability is important to conduct interactions based on accurate beliefs regarding partner's emotional state.

The active perception capability is achieved based on the idea of predictive coding on the same model as that proposed in Chapter 5. The robot selects its own actions to minimize the network energy in the proposed model. This approach is discussed in Chapter 6.

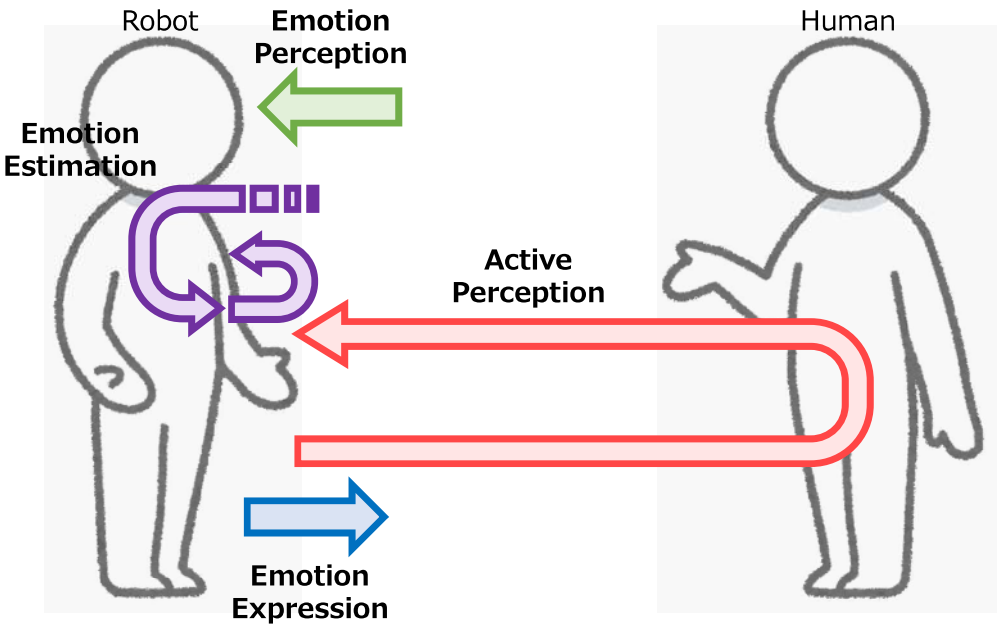



Figure 3.2: Tasks from the robot perspective in emotional interaction

## 3.2 Restricted Boltzmann Machine: A Key Mechanism of Artificial Emotion Model based on Predictive coding

To construct an artificial emotion model to address the described tasks in the previous section, we have many possible approaches available. This section explains the restricted Boltzmann machine, which is the key mechanism of our artificial emotion models based on predictive coding.

The restricted Boltzmann machine (RBM), which is one type of artificial neural network, was proposed by Hinton [58, 60]. It attracted the attention of many researchers in the machine learning field because the training mechanism broke through the vanishing gradient problem of neural networks by adopting a characteristic training method called layer-wise pre-training. On the other hand, we pay attention to a different characteristic of the RBM, which is its use as a generative model. The

generative model is a model that infers the generating process of data in the world. Namely, the generative model represents the relationships between observed data and their latent information. The RBM is also able to represent the generating process of data. The aim of RBM in the training phase is to minimize the cross entropy between data probability and the probability of reconstructed data from the latent information. In other words, the RBM tries to minimize the prediction error between real data and reconstructed data in an unsupervised manner. This training method mimics the idea of predictive coding and relates to the minimization in the free energy of the network. Therefore, in this dissertation, we employ the RBM as a fundamental module for our computational models to reproduce emotion development and the capabilities in emotional interaction. This section introduces two types of RBM: the conventional RBM, which deals with binary signals (i.e., the Bernoulli{Bernoulli RBM), and another type dealing with continuous signals (i.e. the Gaussian{Bernoulli RBM).

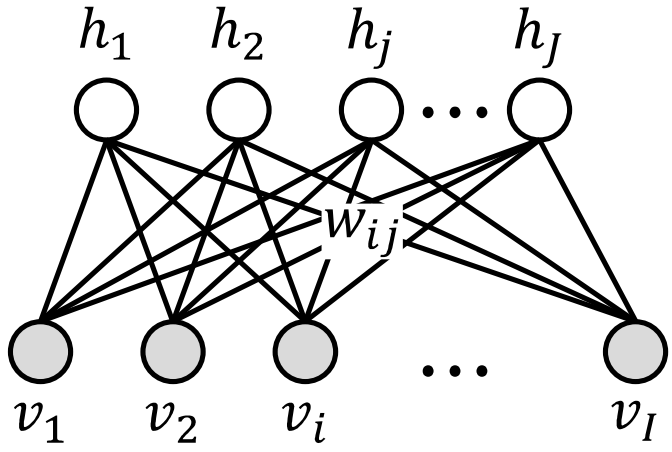



Figure 3.3: Structure of RBMs.$v_i$ and $h_j$ are the activations of the i-th visible layer and the j-th hidden layer, respectively, and $w_{ij}$ is the connection weight of their relative weight.

The RBM [58, 60] is a generative model that represents the generative process of data distribution and latent representation, and can generate data from latent signals [81–84]. The RBM is also a type of artificial neural network, which is consists

of two types of stochastic neurons: visible units $v_i$ and hidden units $h_j$, where i and j are index numbers for each neuron. Each layer has fully symmetrical connections with the other of layers (i.e., the connection weight $w_{ij} = w_{ji}$); however, it has no connections between units in the same layer (Figure 3.3). The visible units receive input signals and the hidden units generate output (latent) signals. The RBM is able to extract various features from input signals (e.g., camera images) by acquiring latent signals (e.g., the edges of the images), which can then be used to reconstruct the input signals in an unsupervised manner [81–84]. We used RBMs for our model because this characteristic allows the model to represent emotional states by integrating and abstracting multimodal sensory signals from interactions.

A Bernoulli–Bernoulli RBM handles only binary signals for both visible (i.e., $v_i \in \{0, 1\}$) and hidden units (i.e., $h_j \in \{0, 1\}$). The activation probabilities for these units are given by

$$p(h_j = 1|v; \theta) = g\left(b_j + \sum_i v_i w_{ij}\right); \tag{3.1}$$

$$p(v_i = 1|h; \theta) = g\left(a_i + \sum_j h_j w_{ij}\right); \tag{3.2}$$

where $g(x)$ is a logistic sigmoid function $1/(1 + \exp(-x))$, and $a_i$ and $b_j$ are biases corresponding to the i-th visible and j-th hidden units, respectively. $\theta = \{a, b, w\}$ is the model parameter. Thus, the RBM not only can extract features from input signals as hidden activations (per Eq. (3.1)) but also can reconstruct the input signals from the extracted features (per Eq. (3.2)).

The joint probability over the visible and hidden units represents the Boltzmann distribution that follows the energy function of the network. The probability distribution and the energy function of the Bernoulli–Bernoulli RBM, $E(v, h; \theta)$, are defined as

$$P(v, h; \theta) = \frac{1}{\mathcal{Z}(\theta)} \exp(-E(v, h; \theta)); \tag{3.3}$$

$$E(v, h; \theta) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i h_j w_{ij}; \tag{3.4}$$

where $\mathcal{Z}(\theta)$ is a normalizing constant to limit the probability values within the range 0 to 1.

The parameters $\theta$ of the RBM are trained through the minimization of the reconstruction error between actual input signals and reconstructed input signals calculated in Eqs. (3.1) and (3.2), respectively. This process is replaced by the minimization of cross entropy, denoted by L, between two probability distributions: $p(v)$ and $p(v|h,\theta)$.

$$L = -\sum_{x} p(v) \log p(v|h,\theta) \tag{3.5}$$

Cross entropy corresponds to the distance between two probability distributions, where $p(v)$ and $p(v|h,\theta)$ are the distributions of actual and reconstructed input signals, respectively. In fact, the minimization of cross entropy in the RBM corresponds to the minimization of the free-energy (i.e., this objective function relates to predictive coding [7{12]).

To derive update rules for the model parameters, we differentiate cross entropy using the traditional gradient-based method. The update rules for the parameters are given by

$$\Delta w_{ij} = \epsilon_w \left( \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recon} \right); \tag{3.6}$$

$$\Delta a_i = \epsilon_a \left( \langle v_i \rangle_{data} - \langle v_i \rangle_{recon} \right); \tag{3.7}$$

$$\Delta b_j = \epsilon_b \left( \langle h_j \rangle_{data} - \langle h_j \rangle_{recon} \right); \tag{3.8}$$

where the angle brackets $\langle \cdot \rangle_{data}$ and $\langle \cdot \rangle_{recon}$ denote the expectations under the distributions of actual and reconstructed input signals, respectively, and $\epsilon_w$, $\epsilon_a$, and $\epsilon_b$ are the learning rates for the corresponding model parameters. We update the parameters by adding their values to subsequent ones during training (i.e., $w_{ij}^{t+1} = w_{ij}^{t} + \Delta w_{ij}^{t}$; here, t is a learning step). For a detailed account of the learning process of RBMs, see [58].

To take into account the modeling of real values in the visible layer (the robot's sensory signals and their features are not binary values), we substitute Gaussian units $v_i \in R$ for binary units. [85]. The activation probabilities for the visible and the hidden units of this Gaussian{Bernoulli RBM are given as

$$p(h_j = 1|v) = g\left( b_j + \sum_{i} \frac{1}{\sigma_i^2} v_i w_{ij} \right); \tag{3.9}$$

$$p(v_i = v|h) = \mathcal{N}\left( v|a_i + \sum_{j} h_j w_{ij}, \sigma_i^2 \right); \tag{3.10}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes the probability of a normal distribution with a mean $\mu$ and a variance $\sigma^2$, and $\sigma_i$ is the standard deviation associated with the i-th Gaussian visible unit. The probability function of the hidden units is different from Eq. (3.1), because of the inclusion of the variance of the visible units. More specifically, visible activation with a small variance is more weighted toward the hidden activations than the large variance in Eq. (3.9). The energy function of Gaussian–Bernoulli RBMs is defined in [85] as

$$E(v; h; \theta) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_i \sum_j \frac{1}{\sigma_i^2} v_i h_j w_{ij}; \tag{3.11}$$

The update rules must be modified because the data probabilities, $p(v)$ and $p(v|h, \theta)$, of the Gaussian–Bernoulli RBM are different from those of the Bernoulli–Bernoulli RBM. The rules for the Gaussian–Bernoulli RBM are given by

$$\Delta w_{ij} = \epsilon_w \left( \left\langle \frac{1}{\sigma_i^2} v_i h_j \right\rangle_{\text{data}} - \left\langle \frac{1}{\sigma_i^2} v_i h_j \right\rangle_{\text{recon}} \right); \tag{3.12}$$

$$\Delta a_i = \epsilon_a \left( \left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{\text{data}} - \left\langle \frac{1}{\sigma_i^2} v_i \right\rangle_{\text{recon}} \right); \tag{3.13}$$

$$\Delta b_j = \epsilon_b \left( \langle h_j \rangle_{\text{data}} - \langle h_j \rangle_{\text{recon}} \right): \tag{3.14}$$

In addition to model parameters $w_{ij}$, $a_i$, and $b_j$, we can modulate another parameter $\sigma_i$ to minimize the cross entropy between $p(v)$ and $p(v|h, \theta)$ because these probabilistic functions are influenced by the variance parameters. We modulate each standard deviation $\sigma_i$ via a surrogate parameter $z_i$, which is defined as $z_i = \log \sigma_i^2$ because the variance $\sigma_i^2$ takes only positive values ($\sigma_i^2 > 0$). The update rule for $z_i$ is given by

$$\begin{aligned} \Delta z_i = {} & \epsilon_z e^{-z_i} \left\langle \frac{1}{2}(v_i - a_i)^2 - \sum_j v_i h_j w_{ij} \right\rangle_{\text{data}} \\ & - \epsilon_z e^{-z_i} \left\langle \frac{1}{2}(v_i - a_i)^2 - \sum_j v_i h_j w_{ij} \right\rangle_{\text{recon}}: \end{aligned} \tag{3.15}$$

Through this modulation, each variance is expected to be closer to the actual variance of the input signals. As the variance becomes closer to 0.0, the more strongly the input signals contribute to the probabilities of the hidden activations. Similarly, the noise in the reconstructed signals decreases with the variance.

## 3.3 Summary

This chapter has provided our intended interaction scenario and four challenges to realize emotional interactions between humans and robots. We assumed that humans (caregivers) and robots (infants) interact with each other in a face-to-face situation, and both of them use multimodal expressions such as facial expressions and vocalizations. In the above interaction, a robot is required to have several capabilities such as emotion perception, emotion expression, emotion estimation, and active perception. To acquire these capabilities in emotional interactions, we employ an RBM as a key mechanism for the proposed model based on the idea of predictive coding. Two types of RBM have been evaluated for our model: the Bernoulli–Bernoulli RBM and the Gaussian–Bernoulli RBM.

The following Chapters 4, 5, and 6 provide detailed descriptions of our challenges. The proposed models for each challenge based on the RBMs are explained and evaluated.

# Chapter 4

# Modeling Development of Multimodal Emotion Perception

Humans recognize others' emotional states such as delight, anger, sorrow, and pleasure through multimodal expressions during interactions. However, it is unclear how this capability of emotion perception is acquired during infancy. This chapter presents a neural network model that reproduces the developmental process of emotion perception through an infant–caregiver interaction. We employ an RBM, which is described in Chapter 3, as a key component of the proposed model. The model receives multimodal expressions from a caregiver (visual, audio, and tactile signals in our current experiment) and learns to recognize her/his emotional states. We hypothesize that emotional categories of multimodal stimuli are represented in a higher layer in the network owing to two important functions: tactile dominance and perceptual improvement. The former refers to tactile sensors that can detect emotional valence of stimuli such as positive, negative, and zero valence more directly than can other sensors owing to characteristics of the nerve system of the skin. The latter refers to developmental changes in perceptual acuity. Experimental results demonstrated that tactile dominance and perceptual improvement play the role of facilitating the differentiation of emotional states of multimodal expressions; however, the influences only appear when both functions are included in the model.

## 4.1 Background

Emotion perception refers to capabilities of recognizing emotions of others. We humans can estimate another person's emotional states such as joy, neutral, anger, etc. from multimodal expressions during interactions [18, 62]. Many researchers have studied how humans acquire the capability of emotion perception in infancy (e.g., [20, 86]) although it is not fully understood to date. To investigate the developmental changes in emotion perception, several researchers applied audiovisual emotional stimuli such as facial expressions and vocalizations to infants [51–54]. Walker-Andrews [51] claimed that human infants have immature capacities to detect, discriminate, and recognize others' emotions at birth; for instance, younger infants (around five months old) can detect only primary emotional expressions of others. However, these capabilities rapidly develop during the first year of life. Older infants can discriminate and recognize the emotional categories of others' expressions (e.g., basic emotions) owing to perceptual development. Grossmann [54] reported that 12-month-old infants showed different reactions to an event-related potential in their brain when they faced angry and happy audiovisual stimuli (i.e., facial and vocal emotional expressions of others).

We consider that there is more evidence in support of the developmental process of emotion perception, in particular the findings where the tactile interaction between infants and their caregivers and where developmental disorders appear to affect the ability of emotion perception [65–67, 87]. However, these studies observed only changes in behaviors and brain activations when the subjects faced emotional stimuli. Thus, it remains unclear what causes the developmental changes in emotion perception during infancy.

To address this issue lacking in prior studies, we proposed a computational neural network model based on the idea of predictive coding for reproducing developmental changes in emotion perception in infancy. This model relies on two key ideas based on psychological studies [51, 67]. The first idea is tactile dominance: that is, the fact that the sense of touch can detect the emotional valence of stimuli more directly than can other modalities. The second idea is perceptual development of infants' multimodal sensation (i.e., sense of vision, audio, and touch). We explain this and hypotheses

for representing the development of emotion perception in Section 4.2. The proposed model is composed of RBMs, which are stochastic neural networks. The first idea, tactile dominance, was modeled as a semi-supervised module of the proposed network, while the second idea, perceptual development, was modeled via a learning process of the variance parameters in the distribution of input modules (see Section 4.3 for more detail). We compared experimental results under four different conditions (i.e., with and without tactile dominance, and with and without perceptual development) by using virtual infant{caregiver interactions. Finally, we discuss the relationships between our model and brain regions related to emotion, the validity of including tactile dominance as a contributor to the development of emotion perception, which will be discussed in relation to congenital insensitivity to tactile sensation; and the future scope of this line of research in Section 4.5.

## 4.2 Our hypotheses

### 4.2.1 Tactile Dominance

Tactile sensation is extremely important for infants, and younger infants tend to use touch to interact with their external environment, owing to their undeveloped vision [67, 88]. Touch allows infants to perceive the emotional valences of their interactions as well as detect the shapes of objects. Interestingly, caregivers also employ the tactile modality more often than others (e.g., vision and audio) when interacting with infants [89].

Touch is also an important modality for emotional communication from a neuroscience perspective. When we suffer pain from tactile stimuli, for instance, two types of nerves in our skin are activated [90, 91]: A$\delta$-fiber, which is a myelinated fiber, and a C-fiber, which is an unmyelinated fiber. C-fibers are considered more primitive nerves than A$\delta$-fibers. Björnsdotter et al. [92] examined the different anatomical mechanisms of C-fibers that transmit positive emotional valence due to touch. Such C-fibers are called "C tactile (CT) afferents". CT afferents are distributed over hairy skin and respond to gentle contact (e.g., stroking the surface of the skin) at a velocity range of 1–10 cm/s. They concluded that CT afferents help humans experience positive

emotions and enhance the social aspects of human{human interaction through skin contact. Importantly, C-fibers, besides detecting the emotional valence of touch, deliver the information not only to the somatosensory area of the brain but also to the limbic system (e.g., the insular cortex and thalamus) that is known as an emotional brain region. It is conceivable that the emotional information ascribable to the tactile dominance has an important role in the abstraction and the integration of sensory information based on the emotional state. For example, it is difficult to recognize the state of the emotion to use only the intensity as a feature in auditory stimuli. However, it is possible that we separate the sensory information by slight differentiation of features to obtain emotional information from tactile stimulation. Craig [21] measured the activity of the insular cortex to tactile stimuli by using functional Magnetic Resonance Imaging (fMRI). Infants react more strongly in their cerebral cortex to tactile stimuli than to other sensory inputs. Brain imaging study has revealed that tactile stimulation activates a wider range of the neonatal cerebral cortex than visual or auditory stimuli [93].

The first evidence supposes that we should take the infant–caregiver tactile interaction into account for the modeling of emotion development. The second evidence suggests that dominant emotion signals affect the developmental differentiation of emotion in infancy. However, this communicative function (touch) has been neglected in modeling studies of emotion.

We hypothesize that tactile communication allows infants to perceive the emotional valence, a value of emotional stimuli (e.g., positive, negative, and non-valued (zero value) information) from others' multimodal expressions. For instance, touching the skin of infants softly might elicit a positive emotional valence, whereas more forceful contact or pinching might elicit a negative emotional valence during the interaction. Consequently, the emotional valence from the sense of touch might aid in perceiving emotional categories of other sensory signals during infancy.

### 4.2.2 Perceptual Development

Humans' sensory organs develop during the fetal period [94], whereas their perceptive faculties develop after birth over the course of one year. For instance, infants' visual

acuity increases from birth to 6 months of age [95]. Furthermore, for auditory perception, infants' ability to discriminate frequencies improves from 3-, 6-, to 12 months after birth [96].

Walker-Andrews [51] studied the development of emotion perception from the viewpoint of the influence of perceptual development during the first year of life in infancy. By reviewing many studies that considered the development of emotion perception and perceptual development in the sense of vision and audio, she claimed that younger infants (around five months old) can notice primal information of the emotional meaning of others' expressions (e.g., positive or negative). She also claimed that older infants (around one year old) are able to discriminate and then recognize the emotional states (e.g., basic emotions) of others from their expressions because perceptual development increases the differentiation of the information for affect from others' multimodal expressions.

Similarly, researchers have explored the effect of perceptual development on category generalization [97, 98]. They compared tendencies in classifying new objects among children of different ages. The results showed that younger children (around 5 years old) classified new objects based on holistic similarity, whereas older children (around 10 years old) used dimensional similarity (e.g., the size or color of objects) for such classification. They claimed that these differences in tendency for object classification depended on developmental changes in perceptual resolution.

More recently, several studies have considered perceptual development in modeling cognitive and motor skill learning in robotics. The authors in [99] proposed a learning model of joint attention employing a developmental mechanism of visual perception. Joint attention is the ability to gaze at an object that someone else is looking at, so visual input plays an important role in acquiring the ability. They introduced perceptual development in the visual sense of a robot, which changed the sharpness of the input image from coarse to fine through this development. Their experiment demonstrated that visual development enabled the robot to improve the accuracy of joint attention owing to better-structured visuomotor mapping.

In line with these past findings, we hypothesize that perceptual improvement in multiple modalities also affects the development of emotion perception. Specifically, changes in an infant's perceptual ability may induce gradual differentiation of emotion

perception from ambiguous emotion (e.g., emotional valences) to categorical emotions (e.g., basic emotions [18]).

## 4.3 Computational Model for Emotional Development

In this section, we first introduce our assumptions for modeling the development of emotion perception in infancy. Then, we describe the basic idea and its computational architecture using the proposed model and associated learning process. Finally, the dataset of multimodal emotional expressions that simulated infant–caregiver interactions is explained.

### 4.3.1 Assumptions about the Interaction

We focus on face-to-face multimodal interactions between an infant and a caregiver. Fig. 4.1 illustrates such an interaction, where the infant perceives stimuli from three modalities: vision (focusing on the caregiver's face), audition (hearing the caregiver's voice), and touch (experiencing the caregiver's touch). For instance, when the caregiver tries to make the infant happy, the infant will see the smiling face of the caregiver, hear her happy voice, and feel her gentle touch. Here, we assume that multimodal expressions of the caregiver are consistent among the three modalities. Furthermore, for the sake of simplicity, we suppose that the infant is receiving interaction signals only from the caregiver.

### 4.3.2 Proposed Model

Fig. 4.2 provides an overview of the proposed model. The model comprises two types of modules: sensory and emotion (see Fig. 4.2). There are three sensory modules related to three different sensory modalities (i.e., visual, auditory, and tactile), and these modules process low-level sensory signals obtained in infant–caregiver interactions. The emotion module acquires the representation of emotional states by

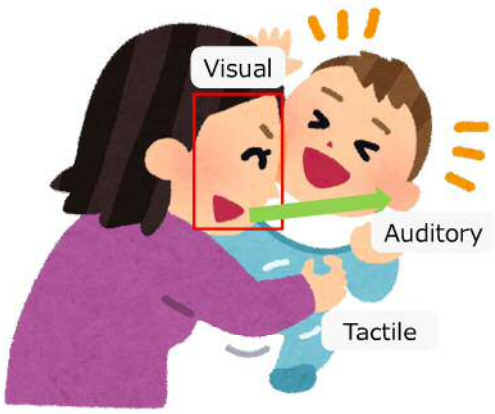



Figure 4.1: Example of face-to-face infant–caregiver interaction. The infant perceives emotional signals from the caregiver using three modalities: visual, auditory, and tactile. The caregiver's expressions are consistent among the three modalities and induce the same emotional state in the infant as in the caregiver.

integrating multimodal signals from the sensory modules. Our challenge is to propose a biologically- and neurologically-plausible mechanism for the development of emotion perception. To address this issue, we constructed each module by adopting the RBM [58, 60] that was described in Chapter 3. The reason why we employed the RBM is that its learning mechanism corresponds to a well-known theory of the human brain mechanism called the free energy principle [9]. The free energy principle proposed by Friston hypothesized that the fundamental mechanism of the human brain is the reduction of prediction error by free-energy minimization, and the theory is related to the idea of predictive coding. To propose the computational model based on the theory might aid us in comprehending the mechanism for the development of emotion perception.

### Sensory Module

Each sensory module comprises a Gaussian–Bernoulli RBM because the input signals from the sensors are continuous values. The visible layers receive sensory signals from the corresponding sensors (i.e., visual, auditory, and tactile). Each module processes

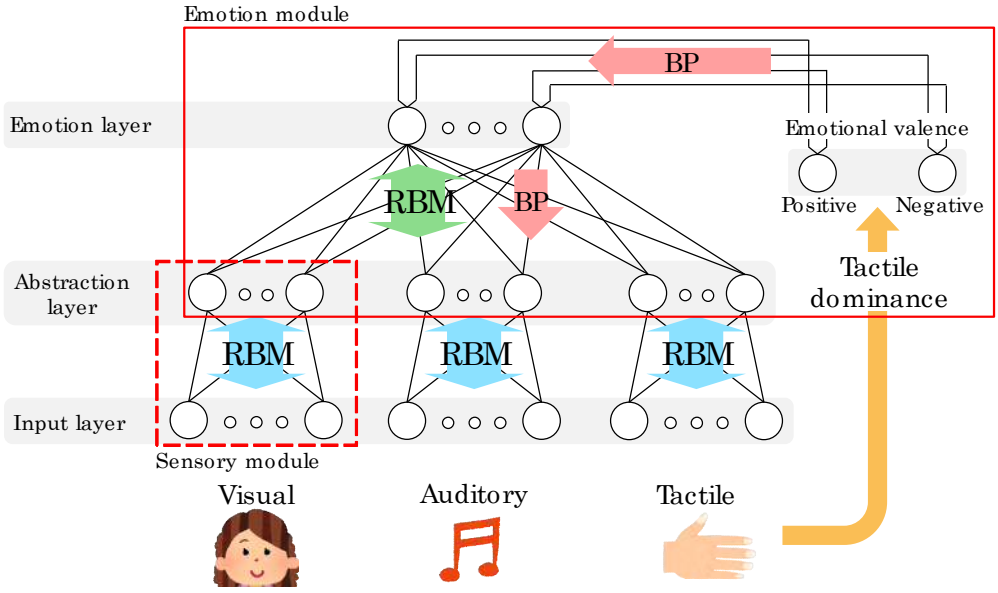



Figure 4.2: Computational model for the development of emotion perception based on tactile dominance and perceptual improvement in infancy. The three lower RBMs denoted by blue connections (i.e., the region enclosed in broken lines), constitute sensory modules, which process different modality signals independently. The region enclosed within the red solid line denotes the emotion module. The red arrows with "BP characters indicate that the connection weights were modulated by back propagation in the model training (see section 4.3.3).

these signals independently.

Perceptual development, which is one of the factors we proposed to drive the development of emotion perception, was modeled as modulations in the variance of the visible units. More specifically, the variance $\sigma_i^2$ is refined through the corresponding parameters $z_i$ (Eq. (3.15)) to reduce the error between the actual input signals and the reconstructed input signals from the hidden activations. Early in the model training, the variance of the visible units is large, which causes several Gaussian distributions to cover many input signals (Fig. 4.3(a)). Therefore, the hidden layer initially represents rough clusters of input data, which makes the reconstruction signals coarse. Later in the training, however, the variance is refined and the regions covered by the Gaussian distributions decrease. The reduced variance leads to more precise reconstructions

compared to when the variance is large (Fig. 4.3(c)). In parallel with this, the mean values of the Gaussian distribution should also be updated to approach the actual value of the input signals to improve accuracy. The refinement of both the variance and the mean value of Gaussian distributions reproduces perceptual development, just as in [99]. In Section 4.4.5, we illustrate the developmental changes in the variance and reconstructed signals in the visual sensory module as an example.

Following the training, the hidden activations of the sensory modules represent abstracted features of the corresponding sensory signals. These activations are then used as input signals for the RBM in the emotion module.

### Emotion Module

The Bernoulli{Bernoulli RBM used in the emotion module is called the multimodal RBM in the proposed model, given that it uses the combined hidden activations of the three sensory modules as input signals. The hidden neurons of the multimodal RBM are connected to the emotion valence layer. The emotion valence layer was introduced into the model to represent tactile dominance. In Section 4.2.1, we mentioned that human skin is equipped with specific nerve fibers (C-fibers) that can detect the emotional valence of touch. C-fibers are activated by specific tactile stimuli (e.g., a gentle stroke with slow velocity or a pinch) and transmit the emotional valence of the stimuli (i.e., positive or negative) to the brain regions that process emotion. To emulate this function, we implemented two units in the emotion valence layer that detect and transmit the positive and negative valence to the emotion layer. For example, when stroke stimuli were presented, the activation value of the positive unit was set to one, while that of the negative unit was set to zero. On the other hand, any unpleasant contact set the negative unit to one. When the contact had no emotional valence (e.g., weak pat and touch), both units were set to zero. How emotional valence is detected from various tactile stimuli was defined according to physiological evidence from past studies [90{92, 100].

In this module, the multimodal RBM was trained initially by using output signals from all sensory modules. It updates the parameters of the RBM to reduce the

reconstruction error in the abstraction layer. The module eventually learns the relationships between the hidden layers of the multimodal RBM and emotional valence in a supervised manner through a back propagation algorithm. In our model, the back propagation mechanism modulates the connection weights not only between the hidden layer and emotion valence nodes but also between the hidden and visible layers of the multimodal RBM. We believed that the emotion layer (i.e., the hidden layer of the multimodal RBM) acquires representations of emotional states from multimodal signals.

### 4.3.3 Learning Process of the Proposed Model

We trained the proposed model by performing ten sequences of the following three phases.

1. The parameters of the RBMs in the sensory modules were trained using Eqs.(3.12){(3.15). This phase is illustrated by blue arrows in Fig. 4.2.

2. The multimodal RBM in the emotion module was trained using Eqs. (3.6){(3.8). This phase is illustrated by the green arrow in Fig. 4.2.

3. The connection weights of the emotional valence units and the multimodal RBM as well as in the multimodal RBM were modulated by back propagation. This phase is illustrated by red arrows in Fig. 4.2.

We continued each training phase for 1000 steps. After the third phase, we began the first phase again. The overall sequence was repeated ten times.

The model structure and learning method are based on a deep belief net [60] and a multimodal deep belief net [83]. However, these previously used models only executed one sequence of the training phases; in contrast, we partitioned the training phases because we wanted to examine and illustrate the developmental changes in the proposed model. For further details of the training method of the RBM, see [58].

### 4.3.4 Multimodal Sensory Signals

We evaluated our model by using an interaction dataset that simulated infant–caregiver interactions. Each interaction datum contains sensory signals for the three modalities (i.e., visual, auditory, and tactile modalities) which represent the seven basic emotions (i.e., joy, surprise, anger, disgust, sadness, fear, and neutral). The data were collected by using a robotic system, which consisted of a USB camera, a microphone, and a soft tactile sensor. An experimenter faced the system and expressed emotional expressions, unlike typical face-to-face interactions but like infant–caregiver interactions, namely, exaggerated expressions. For instance, auditory signals, especially called infant-directed speech, have salient acoustic features [101, 102] (e.g., wide-range pitch and fundamental frequency), and tactile signals also have wide-range features [67]. Our dataset includes these characteristics of multimodal signals. We assumed that the caregiver's expressions of emotional states to the infant were consistent across all three modalities in each interaction, and that a given expression would evoke the same emotional state in the infant. For example, we assumed that when a caregiver showed the infant a smiling face, the infant would experience joy; furthermore, caregiver's auditory and tactile expressions in the same interaction would make the infant experience joy. It is important to note that the proposed model is not provided with the emotional labels of the input signals (i.e., the seven basic emotional states); instead, the model estimates the emotional states with a help of the emotional valences of the signals (i.e., positive, negative, or zero emotional valence) via tactile dominance.

Table 4.1 describes the infant–caregiver interaction dataset, while Fig. 4.4 shows a sample of the actual multimodal signals in the dataset. The first rows of the figure (i.e., Figs. 4.4(a), 4.4(d), and 4.4(g)) illustrate the visual, auditory, and tactile sensory signals for joyful emotion. The second and third rows show the signals for angry and neutral emotions. In the following subsections, we provide further details on the multimodal sensory signals and their features for the sensory modules.

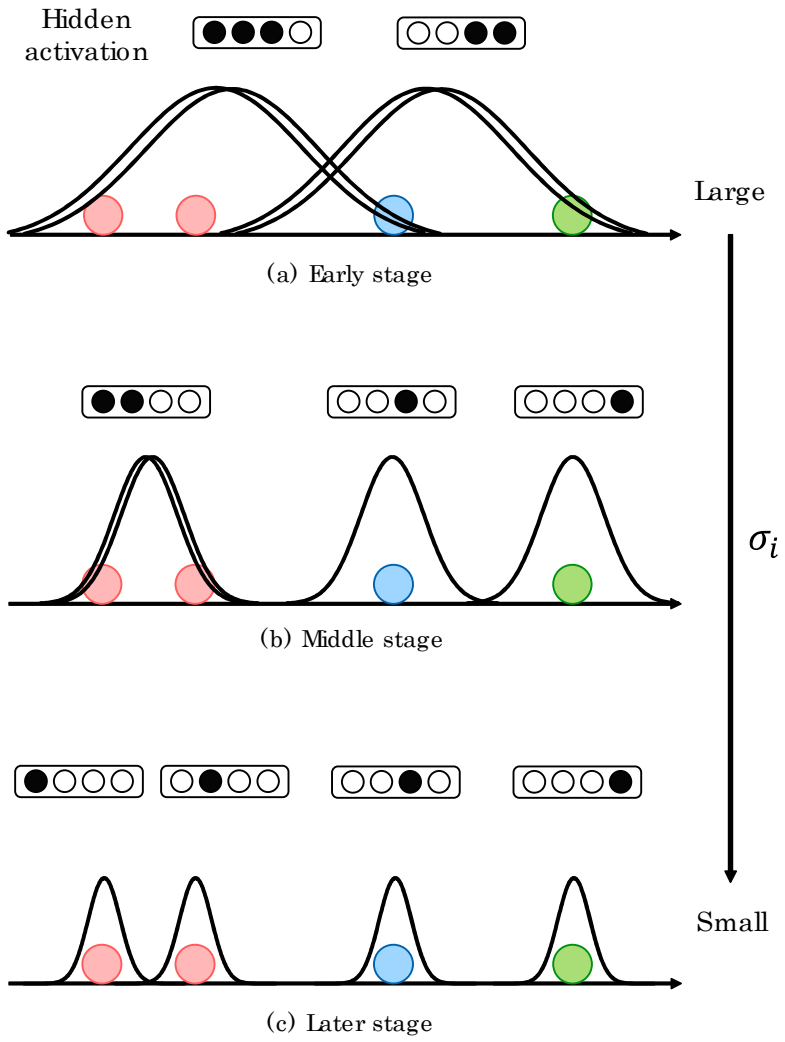



Figure 4.3: Examples of the sensory module's behavior through training. The vertical axis and color variation of the circles represent feature values of the sensory signals (e.g., intensity) and the different emotions, respectively. The boxes with black and white circles show the active and inactive hidden units, respectively, and the Gaussian curves represent the Gaussian distributions that correspond to the hidden activations. The variances are refined and the activation patterns of the hidden layer increase as wtraining progresses over the (a) early, (b)middle, and (c) after stages.

Table 4.1: Description of dataset simulating infant{caregiver interaction

| Emotional state | Visual stimuli | Auditory stimuli | Tactile stimuli (emotional valence) | Number of data |
|---|---|---|---|---|
| Joy | Smiling face | Pitch rise in voice | Stroke (positive) | 150 |
| Surprise | Surprised face | Loud voice | Touch (zero emotion) | 150 |
| Anger | Angry face | Loud voice | Pinch (negative) | 125 |
| Disgust | Worried face | Low tone voice | Pinch (negative) | 125 |
| Sadness | Tearful face | Quite voice | Weak pat (zero emotion) | 125 |
| Fear | Frighten face | High frequency voice | Pat (negative) | 125 |
| Neutral | Neutral face | Neutral voice | Touch (zero emotion) | 125 |

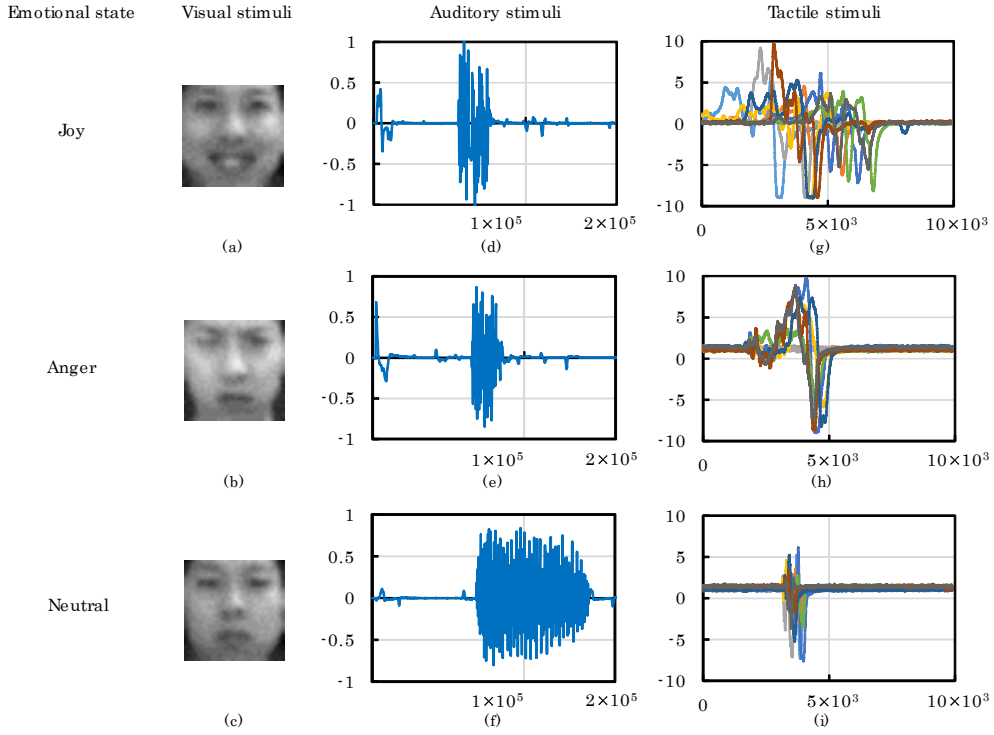



Figure 4.4: Samples of multimodal signals in our interaction dataset

### Visual Stimuli

The visual stimuli used were facial expressions produced by an experimenter playing a parent. We cropped the face region from images captured using a USB camera, and each image was converted to a gray scale image of size 30 × 30 pixels. The experimenter expressed facial expressions of the seven basic emotions, and each emotional face had ten variations. Fig. 4.4 (a){(c) shows examples of the converted facial images for joyful, angry, and neutral emotions, respectively, and Figs. 4.5 show other examples. The shapes of the eyebrow, eye, and mouth represented emotional characteristics of facial images. For instance, in many of the facial images of joy and surprise in our dataset, the mouth was open. By contrast, for the other emotional states, the mouth was closed. The first 20 principal components with contribution rates above 98 % were extracted from all of the converted images by the principal

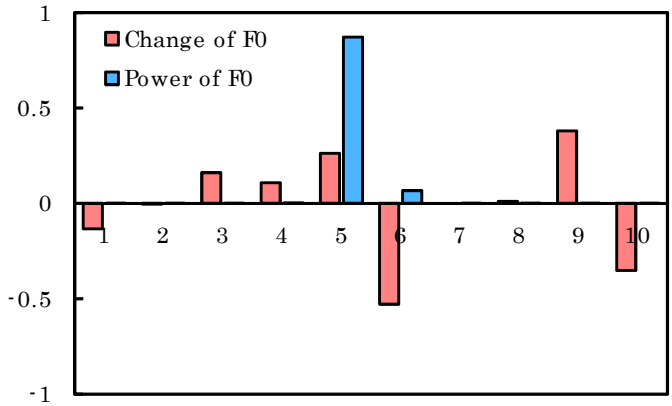
component analysis method to reduce the dimensions of the facial data. We utilized these 20 principal components as input signals for the visual sensory module.

(a) Joy (b) Surprise (c) Neutral (d) Anger (e) Disgust (f) Sadness (g) Fear

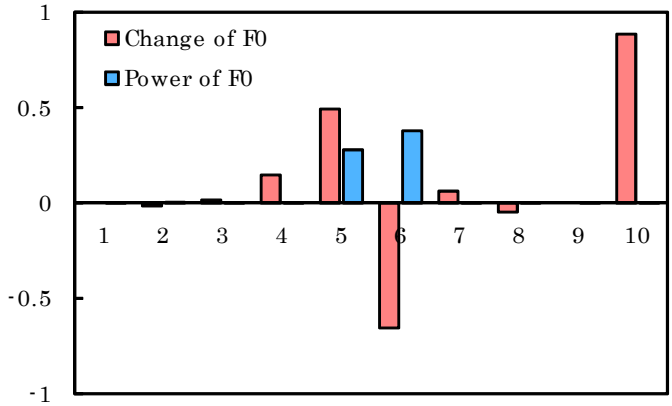
Figure 4.5: Examples of facial expressions of caregivers in communication with infants as visual stimuli.

### Auditory Stimuli

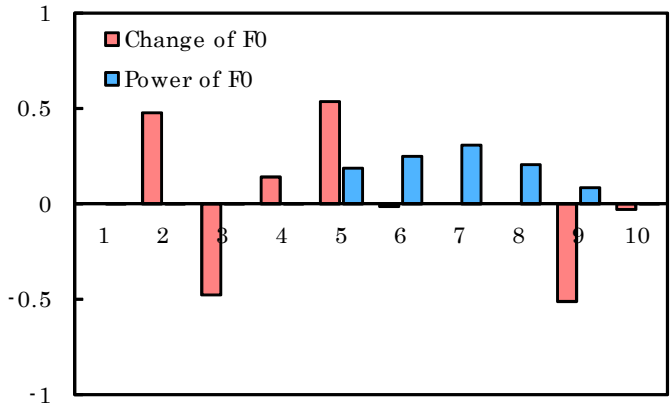
Auditory stimuli were single mora voices expressed as "Maa" corresponding to the seven basic emotions recorded from the same experimenter for the corresponding visual stimuli, and each emotional voice had ten variations. The reason that we used a single mora voice was that the acoustic characteristics were enhanced by simple utterances as well as for infant-directed speech [102], and we wanted to simplify the stimuli as much as possible for the experimental setting. The graphs in the center column of Fig. 4.4 show the raw signals (i.e., sound waves) of voices corresponding to joyful, angry, and neutral emotions. To extract features from these signals, we divided each signal into ten even sections and calculated acoustic features, namely the change in the fundamental frequency (F0) and the power of the F0 for the signals in each section (i.e., there are 20 features). All features were normalized in each section to fit the Gaussian distribution at a zero mean and a unit variance. We used these 20 features of audio signals as input values for the auditory sensory module. Fig. 4.6 shows an example of extracted acoustic features from the audio signals in Fig. 4.4. As evident in Fig. 4.6, joyful and angry voices were shorter than were neutral voices. The power of F0 (indicated by blue bars) was detected only during two consecutive sections. Joyful and angry voices, on the other hand, had similar characteristics such as a short duration and high-intensity sounds as seen in Figs. 4.4 and 4.6. This implies that only some emotional states can be discriminated by low-level acoustic characteristics, whereas ambiguities remain in certain types of emotional voices.




(a) Joy




(b) Anger




(c) Neutral

Figure 4.6: Extracted auditory features from Figs. 4.4 (d){(f) (i.e., joyful, angry, and neutral voices). The horizontal axis represents the divided sections, and the vertical axis shows the normalized value of each feature. The red and blue bars indicate the change of the F0 between the current section and the previous section and the power of the F0 in each section, respectively.

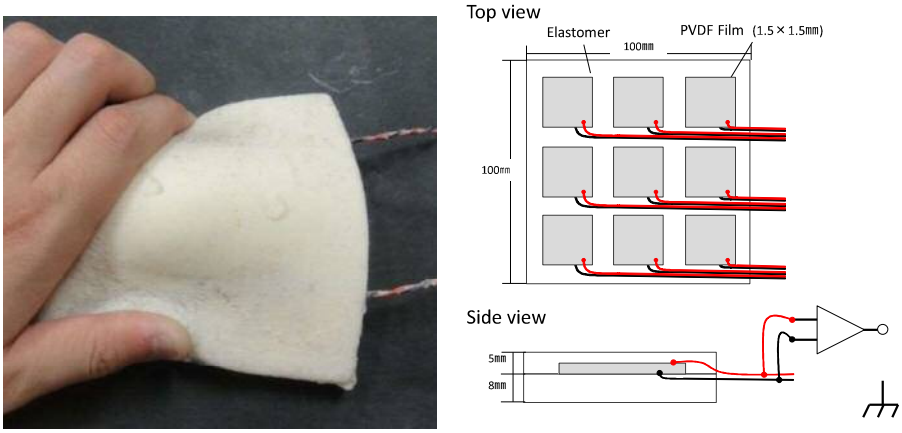
### Tactile Stimuli and Emotional Valence

Tactile stimuli and emotional valence were some of the most important signals in this experiment, as they were used to verify one of our hypotheses. We collected tactile stimuli simulating those used in infant{caregiver interaction via a human-skin-like tactile sensor. An overview and schematic diagram of our sensor, which was developed based on [103], are shown in Figs. 4.7(a) and 4.7(b). The tactile sensor is composed of two materials: polyvinylidene difluoride (PVDF) films and elastomer. More specifically, nine PVDF films (i.e., nine channels), which output voltage depending on the change rate of deformation (i.e., velocity of contact force), were sandwiched between two layers of human-skin-like elastomers (EXSEAL Co., Ltd). Figs. 4.4 (g){(i) provide examples of nine sensor output signals corresponding to joyful (stroke), angry (pinch), and neutral (touch) tactile stimuli, respectively. Each sensory signal was smoothed with a moving average filter based on the previous 100 samples to reduce noise. For instance, the stroke stimulus activated the tactile sensors for a longer duration than did other forms of contact, and the sensor output signals did not synchronize with each other (Fig. 4.4 (g)) because the contact point moved over large areas very slowly. On the other hand, the pinch and touch stimuli activated channels synchronously because their contact points did not move. Furthermore, the dynamic deformation of the sensor during the pinch stimulus was evident by the large values for the sensors' signals.

We calculated nine features from the sensory signals to extract the characteristics of tactile stimuli. Fig. 4.8 shows the relationships between a single sensory signal and the calculated features from a stroke stimulus. First, we extracted five features: (i) the maximum absolute velocity of contacts, (ii) the number of code changes in the signals, and the intensity of (iii) low, (iv) middle, and (v) high frequency bands (low: 1{60 Hz; middle: 61{100 Hz; and high: 101{200 Hz) from the raw signal. Next, we calculated the integral values of the signal in terms of time to estimate the contact force and extracted the remaining four features: (vi) the duration of contact, (vii) duration of a contact with strong force, (viii) the maximum force of the contact, and (ix) the number of channels that detected contact. The maximum features (i.e., (i) and (viii)) were calculated from the whole channel (i.e., nine channels) values. The

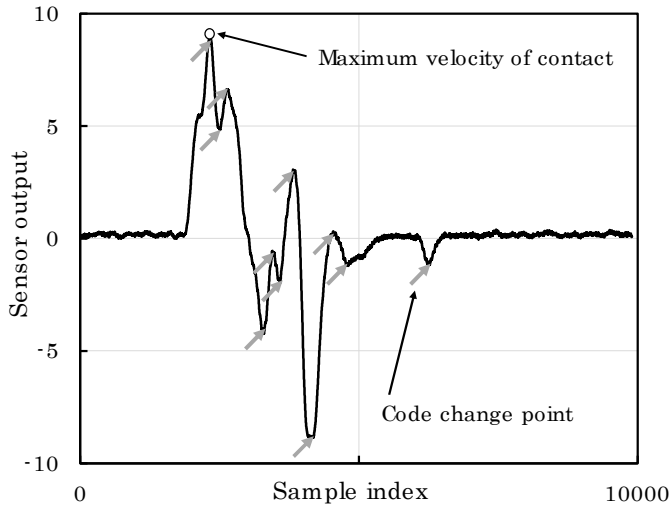
number of channels that detected contact (i.e., (ix)) was determined by counting the channels where the integral value exceeded a threshold. The duration features (i.e., (vi) and (vii)) were logical disjunctions between all channels. The thresholds for contact detection and strong contact detection were 0.8 and 1.5, respectively. The other features (i.e., (ii), (iii), (iv), and (v)) were averaged across the whole channel, individually. These nine features were determined based on our knowledge of tactile receptors and the most important properties of touch [67, 104], and they were used as input signals for the tactile sensory module (i.e., the tactile sensory module has nine input nodes).

In this experiment, we used three types of emotional valence: positive, negative, and zero. Emotional valences were predetermined for each tactile stimulus by the designer based on the inherent properties of C-fibers in the human skin [90, 92, 100]. Specifically, the stroke stimuli induced a positive emotional valence, while the touch and weak pat stimuli corresponded to zero emotional valence. The pinch and pat stimuli were considered to generate a negative emotional valence because the high pressure is known to activate C-fibers and evoke pain in humans. Table 4.1 indicates all relationships between tactile stimuli and emotional valences. These emotional valence signals were represented as neuron activations of the emotion valence layer in the emotion module. As described in Section 4.3.2, positive emotional valence set the positive unit of the emotion valence layer to one, while the negative unit was set to zero. On the other hand, negative valence set the negative unit to one and the positive unit to zero. When the emotional signal had zero valence, both units were set to zero.

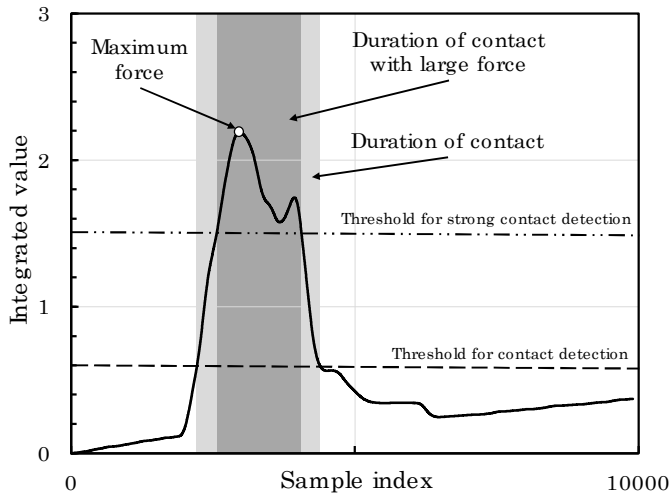



(a) Tactile sensor (b) Structure of the tactile sensor

Figure 4.7: Overview and structure of our tactile sensor composed of urethane foam and PVDF films.




(a) Sensor signals




(b) Integrated sensor signals

Figure 4.8: Relationships between sensor signals and extracted features. Graph (a) represents a smoothed sensor signal (using the moving average filter) with features (i) and (ii). Graph (b) shows the integrated values of sensor signal (a) with features (vi), (vii), and (viii).

## 4.4 Experiments and Results

Section 4.4.1 outlines the experimental conditions used to verify our hypotheses regarding tactile dominance and perceptual development. Four conditions were designed to investigate the roles of these two functions. Then, in Sections 4.4.2{4.4.5, we demonstrate the influences of these functions on emotional development by comparing emotion differentiation under the four conditions.

### 4.4.1 Experimental Conditions and Settings

We conducted experiments under the following four conditions to investigate how emotion develops differently with and without the two functions of interest:

A. A model with tactile dominance and perceptual development (i.e., wTD{wPD condition)

B. A model with only perceptual development (i.e., w/oTD{wPD condition)

C. A model with only tactile dominance (i.e., wTD{w/oPD condition)

D. A model without either function (i.e., w/oTD{w/oPD condition)

Condition A (wTD{wPD) included both functions. Tactile dominance was modeled using emotional valence units (which represented positive, negative, and zero emotional valence signals based on tactile stimuli), whereas perceptual development was achieved by refinement of the variance of the input nodes in the sensory modules. This was the main condition for verifying our hypotheses.

Condition B (w/oTD{wPD) omitted tactile dominance from the wTD{wPD condition. In this case, we assumed that the infant was not able to perceive positive or negative valences from tactile stimuli; this disorder has been observed in infants that are born without tactile nerves [105]. To replicate this situation, we removed the emotional valence units and their connections from the emotion module, and therefore skipped the third phase in the training process (see section 4.3.3). We used this condition to assess the role of tactile dominance in emotional differentiation by comparing the results with the first condition.

Condition C (wTD{w/oPD) excluded perceptual development instead of tactile dominance. In this case, we assumed that the infant's perception had matured at the start of the developmental processes. To represent this we fixed the variance of sensory modules' input nodes to 0.1 and excluded the modulation of variance in Eq. (3.15). By contrast, the initial values of the variance in the previous conditions (wTD{wPD and w/oTD{wPD) were set to 1.0 and modulated by Eq. (3.15). This modulation represented the development of perceptual abilities from the immature sensation to the mature one. We used this condition to verify the influence of perceptual development on the differentiation of specific emotional states.

In Condition D (w/oTD{w/oPD), we excluded both functions from the wTD{wPD condition. In other words, the emotional valence units and refinement process of the variance of the input nodes were removed from the proposed model as in the Conditions B and C, respectively.

We utilized the dataset of the simulated infant{caregiver interaction described in section. 4.3.4. The parameters for the proposed model are listed in Table 4.2.

Table 4.2: Parameters of the sensory modules and emotion module

| Parameter | Explanation | Value |
|---|---|---|
| $v_{t\_num}$ | Number of visible nodes of tactile sensory module | 9 |
| $h_{t\_num}$ | Number of hidden nodes of tactile sensory module | 10 |
| $v_{a\_num}$ | Number of visible nodes of auditory sensory module | 20 |
| $h_{a\_num}$ | Number of hidden nodes of auditory sensory module | 10 |
| $v_{v\_num}$ | Number of visible nodes of visual sensory module | 20 |
| $h_{v\_num}$ | Number of hidden nodes of visual sensory module | 10 |
| $\epsilon^{s}_{w}$ | Learning rate for weights of sensory module | 0.001 |
| $\epsilon^{s}_{a}$ | Learning rate for biases of visible units of sensory module | 0.001 |
| $\epsilon^{s}_{b}$ | Learning rate for biases of hidden units of sensory module | 0.001 |
| $\epsilon_{z}$ | Learning rate for log-variance | 0.001 |
| $v_{e\_num}$ | Number of visible nodes of emotion module | 30 |
| $h_{e\_num}$ | Number of hidden nodes of emotion module | 20 |
| $\epsilon^{e}_{w}$ | Learning rate for weights of emotion module | 0.01 |
| $\epsilon^{e}_{a}$ | Learning rate for biases of visible units of emotion module | 0.01 |
| $\epsilon^{e}_{b}$ | Learning rate for biases of hidden units of emotion module | 0.01 |
| $\eta$ | Learning rate for weights on back propagation | 0.0001 |

### 4.4.2 Results

We first present the experimental results for all four conditions. Then, in Section 4.4.3, we compare the results with a focus on the tactile dominance to test our first hypothesis, while in Section 4.4.4, we compare the results with focus on perceptual improvement to test the second hypothesis.

To visualize and evaluate the acquired representations of emotion, we carried out principal component analysis (PCA) on the activations of the emotion layer. Fig. 4.9 shows one example of the PCA results under the four conditions. We selected the first three principal components (PCs) and illustrated the first and second PCs in Figs. 4.9(a), 4.9(c), 4.9(e), and 4.9(g) and the first and third PCs in Figs. 4.9(b), 4.9(d), 4.9(f), and 4.9(h). All plotted data are labeled using the emotional states of input signals. Note that these labels were not used in the model training, they are shown only as a visual aid.

To quantitatively evaluate the representations of emotion in the PC space, we calculated the separation metric, given by

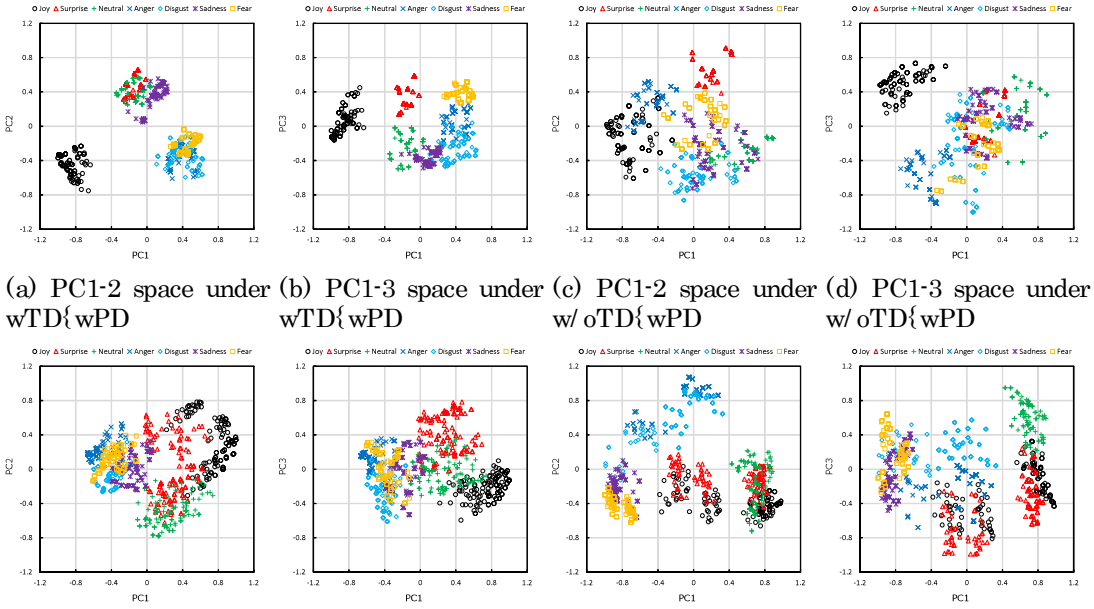
$$J_\sigma = \frac{s_b^2}{s_w^2}; \tag{4.1}$$

$$s_b^2 = \frac{1}{n}\sum_{c=1}^{C} n_c(m_c - m)^t(m_c - m); \tag{4.2}$$

$$s_w^2 = \frac{1}{n}\sum_{c=1}^{C}\sum_{x 2 X_c} n_c(x - m_c)^t(x - m_c); \tag{4.3}$$

where $s_b^2$ and $s_w^2$ are the between-class and within-class variance, respectively; C and c are the number and index of classes; n and $n_c$ are the number of all data used for training and the number of data belonging to class c; and $x \in X_c$, $m_c$, and m are data belonging to class c, the mean of values of $X_c$, and the mean of all data, respectively. The larger the separation metric is, the greater the separation of the clusters in the PC space. Figs. 4.10(a) and 4.10(b) summarize the separation metrics for the categories of emotional valences (i.e., positive, negative, and zero emotional valences) and the seven basic emotions under the four experimental conditions. We calculated the averages and standard deviations of the separation metrics over ten times of each experiment with different initial parameters for the network.

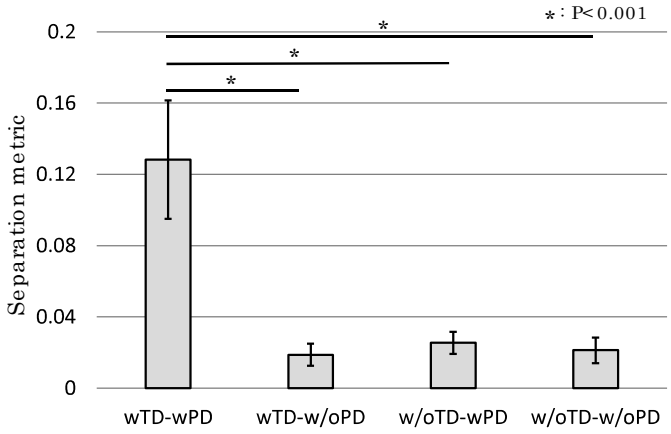
The change of distribution of interaction data in PC spaces at steps 1, 5,000, and 10,000 are depicted in Figure 4.11. Through the learning sequence from 1 to 5,000 steps, the confused distributions of interaction data are separated into positive and others in the first and second PC space. After 5,000 steps, they split into positive, negative, and neutral states. At learning step 1, interaction data were also scattered in the first and third PC space. However, the distribution of data narrowed and formed clusters corresponding to emotional states as learning progressed. Table 4.3 describes the change of separation metrics from the start to end of learning.



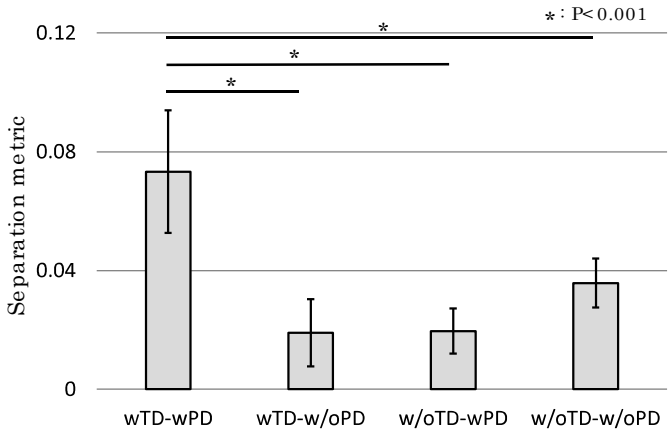

(a) PC1-2 space under wTD{wPD (b) PC1-3 space under wTD{wPD (c) PC1-2 space under w/ oTD{wPD (d) PC1-3 space under w/ oTD{wPD

(e) PC1-2 space under wTD{w/ oPD (f) PC1-3 space under wTD{w/ oPD (g) PC1-2 space under w/ oTD{w/ oPD (h) PC1-3 space under w/ oTD{w/ oPD

Figure 4.9: Acquired low-dimensional representations of emotional stimuli by PCA for the emotion layer activations in the proposed model under the four conditions. (a), (c), (e), and (g) refer to PC1-2 spaces, and (b), (d), (f), and (h) refer to PC1-3 spaces for each condition.

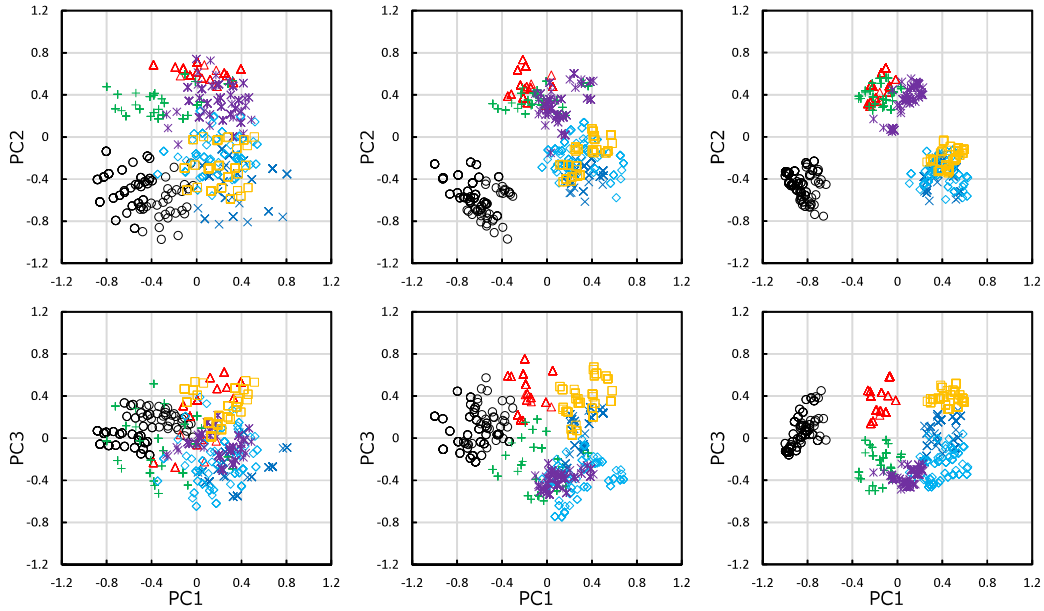



(a) Separation metric for the categories of emotional valences




(b) Separation metric for the categories of seven basic emotions

Figure 4.10: Separation metrics for the categories of emotional valences (i.e., positive, negative, and zero) and the categories of the seven basic emotions under the four conditions




(a) Step 1 (b) Step 5,000 (c) Step 10,000

Figure 4.11: Transitions of representation in the first to second and first to third PC spaces under the wTD{wPD condition.

Table 4.3: Separation metrics for each step under the wTD{wPD condition

| Learning steps | 1 | 5,000 | 10,000 |
|---|---|---|---|
| Separation metrics $J_\sigma$ | 0.021 | 0.039 | 0.082 |

### 4.4.3 Influence of Tactile Dominance on Differentiation of Emotion

We compared the results of the four conditions to demonstrate the influence of tactile dominance on the development of emotion perception. In the first and second PC space under the wTD{wPD condition (Fig. 4.9(a)), the positive (i.e., joy) and the negative (i.e., anger, disgust, and fear) emotional valence clusters are separated by the first PC axis, while the second PC represents the differentiation between the zero

emotional valence cluster and the others. Hence, the interaction data are differentiated into fundamental emotional clusters (i.e., positive, negative, and zero emotional valence) in this space. In the first and third PC space (Fig. 4.9(b)), the vertical axis (i.e., PC3) subdivides the clusters of negative and zero emotional valence into seven emotional states except joy. More specifically, the cluster of negative emotional valence is differentiated into sub-categories of fear, anger, and disgust from the top of the graph. The space composed of the first and third PCs shows that the clusters of emotional valence were subdivided into the seven basic emotions. The separation metrics for both emotional categories under the wTD–wPD condition showed significant differences between wTD–wPD and the other three conditions (in Figs. 4.10(a) and 4.10(b)). In contrast to the above, the results for the other conditions show unclear differentiation of emotional categories in the PC spaces. Only the PCA results of wTD–w/oPD condition demonstrated that the interaction data weakly differentiated into clusters of emotional valences in both spaces; however, the separation metric for categories of emotional valences showed non-significant differences between other conditions.

The overall comparison of these conditions demonstrates that tactile dominance leads to better separation of the emotional categories; however, it is also necessary to include perceptual development in the proposed model as seen in Figs. 4.9(a) and 4.9(b). When tactile dominance was excluded from the model (i.e., w/ oTD{ wPD and w/ oTD{ w/ oPD conditions), the representation of emotional valence clusters could not be obtained using the same interaction data. A potential reason is that visual and auditory signals contained ambiguous features in terms of positive and negative valence. For instance, joy and anger had similar auditory characteristics due to the intensity of those stimuli (Section 4.3.4). By contrast, when tactile dominance was included, the emotional valence inherent in the tactile stimuli disambiguated such situations.

### 4.4.4 Influence of Perceptual Development on Differentiation of Emotion

We assessed the influence of perceptual improvement on the development of emotion perception. From the comparison of results between the wTD{ wPD and wTD{ w/ oPD conditions, we found that the distribution of interaction data, which was weakly clustered based on the emotional valences in Fig. 4.9(f), showed clear differentiation of the seven basic emotions in Fig. 4.9(b) owing to perceptual development. It was clear that perceptual development also facilitated the differentiation of the clusters relevant to the emotional valences as shown in Figs. 4.10(a) and 4.9(a), although this effect was not clearly observed in the comparison between w/oTD–wPD and w/ oTD{ w/ oPD conditions.

Taken together, these results indicate that perceptual development does enhance differentiation of emotional categories in terms of both emotional valences and the seven basic emotions; however, the effect appears only when tactile dominance is included in the model. The result suggests that the two functions facilitate the developmental differentiation of emotion perception but their effects become significant only when they are both present in the model.

### 4.4.5 Perceptual Development Produced by Modulation of RBM Parameter $\sigma_i^2$

We then closely analyzed how perceptual development was reproduced by the modulation of the variance parameters, $\sigma_i^2$, in the sensory RBMs. Fig. 4.12 shows the transition of $\sigma_i^2$ (i = 0; $\cdots$ ; 20) of the vision module over the learning process as an example. All variances of the visible nodes with perceptual development were initialized at 1.0 and updated using Eq. (3.15). The results showed that the variance parameter $\sigma_i^2$ was properly adjusted through training.

We also illustrate the changes in reconstructed images across the learning steps. Fig. 4.13(d) shows four randomly selected input images depicting facial images of joy (left top), neutral (right top), anger (lower left), and sadness (lower right). Figs. 4.13(a) to 4.13(c) show the reconstructed images in Fig. 4.13(d). For some images,

in the early stages of learning (Fig. 4.13(a) { 4.13(b)), the reconstructed images were unclear, making it difficult to determine the emotional state. Furthermore, some reconstructed images appeared to represent different emotional states from the input images. For example, the top left in Fig. 4.13(a) looks similar to an angry face although the input was a happy face. We described the reason for this result in Section 4.3.2. The Gaussian distribution with a large variance covered many input signals and thus generated highly-ambiguous reconstructions. However, in the later stages of learning (Fig. 4.13(c)), the reconstructed images became more similar to the input images. These results indicate that the sensory modules were able to simulate perceptual development by updating the variance of their visible nodes.

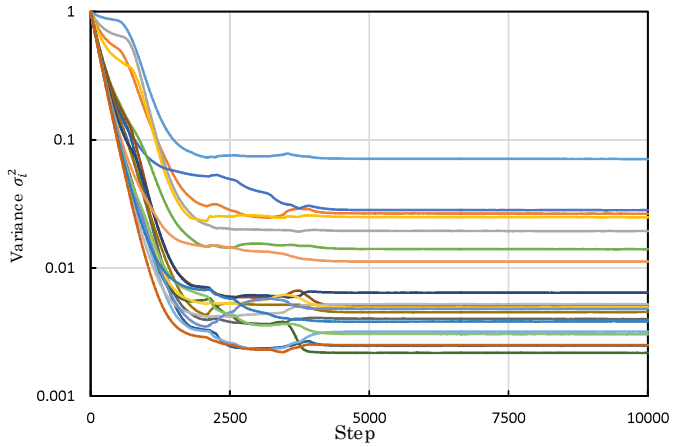
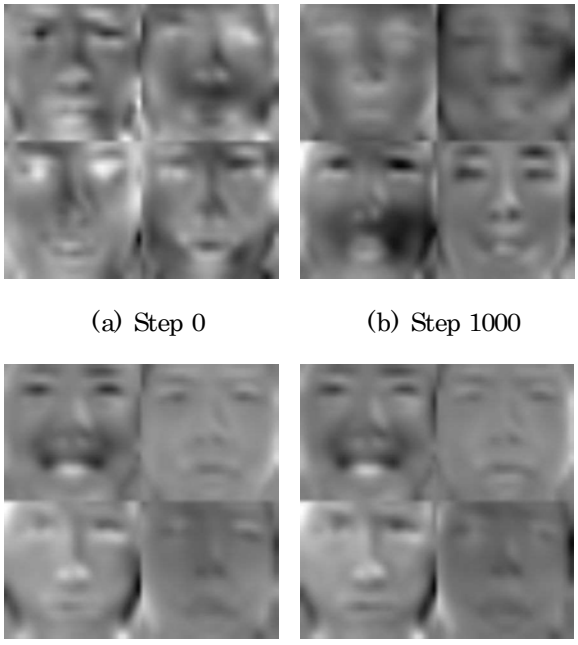



Figure 4.12: Transition of visual nodes' variance for the vision sensory module



(a) Step 0 (b) Step 1000

(c) Step 10000 (d) Input data

Figure 4.13: Examples of input stimuli and reconstructed images during the learning process in the visual sensory module

## 4.5 Discussion and Future Work

We proposed a computational neural network model comprising two modules (i.e., the sensory module using three RBMs and the emotion module) to verify our hypotheses regarding the development of emotion perception in infancy. The sensory module components processed multimodal sensory signals individually, similar to the sensory area of the cerebral cortex (i.e., the visual, auditory, and somatosensory cortices). The emotion module, which was on a higher-level than the sensory module, was used to integrate the abstracted signals obtained from the sensory module RBMs, and the emotion layer in the module further integrated the information of emotional valences based on tactile dominance. It is known that the STS integrates visual, audio, and tactile signals [22] and engages in multimodal information processing for emotion perception [23, 24]. The STG, which is near the STS, also responds to various non-verbal emotional stimuli [25], and the temporal area of an infant's brain perceives and reacts to audiovisual emotional stimuli [26]. There are also known neural connections between the STS and the amygdala; Björnsdotter et al. [92] reported that tactile C-fibers deliver the signals of positive and negative touches to areas of the limbic system, such as the amygdala and insular cortex. All of this prior knowledge suggests that the structure of the proposed model successfully reproduced that of the human brain, at least in relation to emotional processing. In other words, the sensory module corresponds to the sensory areas of the brain, while the emotional valence layer and its function reproduce the amygdala and tactile C-fibers. Furthermore, the emotion layer in the emotion module emulates brain functions, such as perception of categorical emotions in the temporal region (i.e., the STS and the STG). Furthermore, we consider that the proposed model would replicate not only the development of emotion perception but also multimodal sensory processing in general in the human brain. In fact, our challenge was to design a biologically- and neurologically-plausible mechanism for the development of emotion perception based on the latest knowledge on multimodal sensory association (e.g., the free-energy principle [9]). We believe that the generality and the scalability of the proposed model will increase the scientific impact of our work.

In Section 4.4.3, we compared the experimental results between the four conditions

to investigate how emotion develops differently with and without tactile dominance to test our first hypothesis. The results demonstrated that tactile dominance facilitated the differentiation of emotional categories when perceptual development was also included in the model (i.e., the wTD–wPD condition). There are two types of C-fiber in the human skin that perceive emotional valence produced by tactile stimuli. The first type of C-fibers, CT afferents, specializes in the detection of pleasurable touch. This type of fiber tends to be distributed over hairy skin and is activated by light strokes with a velocity of 1–10 cm/s [92]. The other type of C-fiber responds to chemical substances, thermal stimuli, and otherwise negative stimuli such as tactile stimuli pain. In addition to this second type of C-fiber, there are A$\delta$-fibers, which transmit pain signals from the skin to the brain, especially to the somatosensory area. There is a condition called congenital insensitivity to pain with anhidrosis (CIPA), whereby individuals are born without the second type of C-fiber and the A$\delta$-fibers; in lacking these fibers, individuals with CIPA are unable to feel pain. Past studies have also shown that patients with CIPA have impairments in the perception, recognition, and modulation of emotion [105]. Danziger et al. [87], in an experiment designed to estimate others' emotional states, demonstrated that patients with a similar condition called congenital insensitivity to pain (CIP) were inhibited in their ratings of others' painful situations or propensity to infer pain from others' facial expressions using control subjects. Note that CIP patients lose only the A$\delta$-fibers and do not lack the C-fibers in their skin; therefore, they are able to roughly detect pain (i.e., negative sensations) through the C-fibers, unlike CIPA patients. These past findings suggest that tactile C-fibers play an important role in emotion perception and understanding others' emotions. In our study, the experimental conditions where tactile dominance was excluded appear to simulate the characteristics of patients with CIPA, which suggests that our results both demonstrate the effects of CIPA and the importance of C-fibers for the development of emotion perception.

In Section 4.4.4, we assessed the influence of sensory improvement on the development of emotion perception. The comparison results of the four conditions showed that perceptual development facilitated the differentiation of emotional categories when tactile dominance was also included in the proposed model (i.e., the wTD{ wPD condition). This result supports the claim in Walker-Andrew's study [51] that

perceptual development increases the differentiation of information for affect. The idea of perceptual development was modeled by refining the variance of the input nodes in the sensory modules during the training process. The experimental results in Section 4.4.5 further demonstrated that the sensory modules simulated perceptual development by modulating the variance parameters. Other cognitive developmental robotics studies have proposed models whose perceptual capabilities develop according to their performance on their corresponding tasks. For instance, Nagai et al. [99] modified visual filters in their model according to the error rate of the joint attention task. However, perceptual development in our model was independent from the performance of the development of emotion perception. Thus, our model is more generalizable than those of other studies because it demonstrates that perceptual development does not depend on the performance of specific cognitive functions.

In summary, the proposed model represented the development of emotion perception through learning of caregivers' visual, audio, and tactile expressions during interactions. The experimental results demonstrated that both tactile dominance and perceptual development have a combined role in facilitating the development of emotional perception; however, this influence appears only when both functions are integrated into the model together. Connecting our results to behavioral studies in physiology helps to elucidate the neural and social mechanisms of the development of emotion perception. On the other hand, human emotions are affected not only by external senses but also by human behavior and internal, physiological systems (e.g., the endocrine system). It is important that we take such systems into account as in other studies in cognitive developmental robotics [41, 44, 52, 70, 72, 73] when modeling the development of emotion perception. Additionally, we intend to examine the development of emotion perception by integrating physiological indices using nonparametric Bayesian models [106]. To address these future issues, we will combine the results of our prior analyses for more accurate modeling of emotion development.

# Chapter 5

# Emotion Expression and Estimation by Mental Simulation

Humans can express their own emotion and estimate the emotional states of others during communication. This section proposes a unified model that can estimate the emotional states of others and generate emotional self-expressions. The proposed model utilizes a multimodal RBM | a type of stochastic neural network. RBMs can abstract latent information from input signals and reconstruct the signals from it. We use these two characteristics to rectify issues affecting previously proposed emotion models: constructing an emotional representation for estimation and generation of emotion instead of using heuristic features, and actualizing mental simulation to infer the emotions of others from their ambiguous signals. Our experimental results indicate that the proposed model can extract features representing the distribution of categories of emotion via self-organized learning. Imitation experiments demonstrate that using our model, a robot can generate expressions better than with a direct mapping mechanism when the expressions of others contain emotional inconsistencies. Moreover, our model can improve the estimated belief in the emotional states of others through the generation of imaginary sensory signals from defective multimodal signals (i.e., mental simulation). These results suggest that these abilities of the proposed model can facilitate emotional human{robot communication in more complex situations.

## 5.1 Background

Sharing emotion with communication partners is an important attribute in human{ human and human{robot interaction. In attempts to achieve emotional communication, several empathic robots have been developed [34{46]. A study on emotion categorization has been reported that can rectify the first issue mentioned above. Kim et al. [80] proposed deep neural networks that learn to extract features for emotional categorization from audiovisual signals. In their system, deep belief networks (DBNs) comprising RBM were used as unsupervised learning mechanisms. The RBM can abstract input signals and reconstruct the signals therefrom. In experiments, their model extracted emotion-specific features from general ones that are not always important for the classification of emotion. Their results showed that deep neural networks can obtain useful features to represent emotion.

This section proposes a unified model that can estimate human emotion and generate its own emotional expressions to imitate human expressions based on the estimation of the partner's emotion in human–robot interaction. Note that we define "imitation" as the generation of expressions of the robot considering the emotional state of the interaction partner in this study. The model overcomes two issues confronting previous emotional models: constructing an emotional representation of multimodal signals for estimation and generation of emotion instead of using heuristic features, and actualizing mental simulation to infer the emotions of others from their ambiguous multimodal signals. We employed RBMs to address these two issues, as they are able to abstract input signals and recall the signals from their abstract representations. The abstraction capability of RBMs enables us to overcome the first limitation by reducing the dimensions of multimodal signals and associate the multimodal signals. Moreover, our model carries out mental simulation by exploiting the ability to generate sensorimotor signals. The mental simulation mechanism enables the model to estimate the emotional states of others from partial multimodal expressions based on its own experiences.

We examine the abilities of the proposed model via three experiments: 1) construction of an emotional representation from human multimodal expressions using the self-organizing feature extraction capability of RBMs to enable the robot to imitate

human expressions based on the estimation of the human's emotion, 2) confirmation of the advantage of the unified model over the direct mapping model via an imitation experiment when human expressions conflict in modalities, and 3) demonstration of the advantage of mental simulation in the imitation interaction with partial human expressions. Finally, we discuss the relationships among the capabilities of our model and MNS, the limitations of our model, and future issues to be addressed.

## 5.2 Target tasks and the proposed method

This section introduces our target tasks in human{robot interaction and the proposed model that enables the robot to imitate multimodal expressions of others based on the estimation of their emotional states.

### 5.2.1 Challenges and requirements for emotion-based imitation

We considered face-to-face interactions between a human and a robot as the target situation in our study (Fig. 4.1). The human and the robot used multimodal signals, such as facial expressions, hand movements, and speech to communicate with each other. The aim of this interaction is for the human to express his/her emotion by using multimodal signals, and for the robot to imitate the partner's expressions as its own multimodal expressions based on estimation of the partner's emotion. For imitating the emotional expressions of others, it is important that the robot not directly copy expressions of others but instead generate its own expressions through emotion estimation of others because human expressions are sometimes ambiguous. For example, human multimodal expressions are inconsistent between modalities (i.e., a tearful face when crying with delight), or the robot does not perceive complete human expressions in interaction. To copy expressions of others only is not enough to imitate the emotional expressions of others.

To achieve these goals by the robot, our computational model requires mechanisms for estimation of emotion of others and generation of its own expressions based on the perceived emotion. Moreover, the model requires that the estimation mechanism

be robust against ambiguous observations. The proposed model composed of RBMs can resolve these issues by representing two processes: estimation process of the emotional states of others, and the generation process of its own expressions based on the estimated emotion of others. To represent these processes in the same model enables us to actualize the mental simulation mechanism to update the estimation belief of emotion of others. In later sections, we explain the architecture of the proposed model and the method to estimate and generate emotion.

### 5.2.2 Proposed model

Our model consists of two parts: stacked RBMs for each modality in the lower layers and an RBM at the top layer that integrates signals from the lower RBMs (Fig. 5.1). The model first receives the multimodal expressions of the partner as input signals at each modality RBM and estimates the emotional state of the partner through forward sampling from the lower layers to the top layer (red arrows in Fig. 5.1). Following the estimation, the model uses the estimated emotion to generate its own expressions in imitation of the expressions of the partner based on the partner's emotion through backward sampling from the top layer to the lower layers (green arrows in Fig. 5.1). That is, the lower layers of our model represent the multimodal signals of human and robot, and the top layer represents the emotional state inferred from the signals.

We first provide the architecture of our model, called a multimodal DBN in Section 5.2.2. Following this, we explain the estimation and generation mechanisms of emotion as well as an approach to actualize the mental simulation that updates the beliefs concerned with emotion estimation by reconstructing others' ambiguous signals.

#### Multimodal deep belief network

We considered an association of multimodal signals (i.e., facial expression, hand movement, and speech) in our experiments to acquire the representations of human emotional expressions. To model this emotional representation, we use multiple RBMs to first extract features from each sensory signal and then integrate the lower RBM

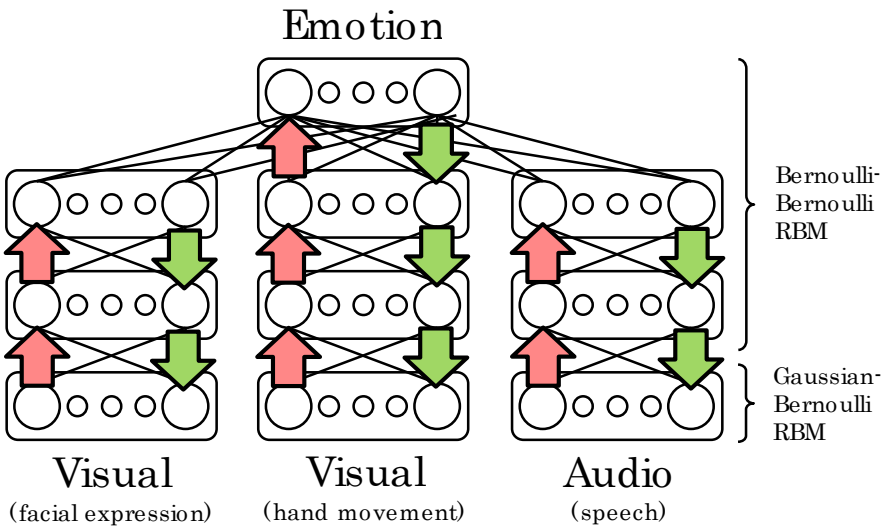



Figure 5.1: Overview of our proposed model. Each two layers (connected black lines) represent an RBM. The red arrows signify forward sampling (estimation of the partner's emotion) and the green arrows signify backward sampling (generation of own emotional expressions).

outputs using a single RBM. First, two different RBMs were stacked to abstract low-dimensional features from each sensory signal, with the lower RBM being Gaussian{Bernoulli due to the continuous sensory signals, and the upper RBM being Bernoulli{Bernoulli. The reason for using different RBMs for each layer of the sensory networks is that using the Bernoulli{Bernoulli RBM to process sensory signals induces information loss because we must discretize the continuous sensory values to be able to encode them. On the other hand, using the Gaussian{Bernoulli RBM for discrete values (i.e., encoded signals using Gaussian{Bernoulli RBMs) induces an increasing learning cost and creates more local solutions for most model parameters than the Bernoulli{Bernoulli RBM. Then, an additional Bernoulli{Bernoulli RBM, called an association RBM, uses the top layers of each stacked RBM as its visible layer to integrate entire modalities (see Fig. 5.1). This structure is called a multimodal DBN [83].

Let $h^{F}_{t;i} \in \{0; 1\}$, $h^{H}_{t;j} \in \{0; 1\}$, and $h^{S}_{t;k} \in \{0; 1\}$ denote the top hidden units of facial expression, hand movement, and speech networks, respectively. The activation

probability of the s-th hidden unit of the association RBM $h^A_s \in \{0; 1\}$ is given by

$$p(h^A_s = 1 | h^F_t; h^H_t; h^S_t) = \sigma\left(\sum_i h^F_{t;i} w_{is} + \sum_j h^H_{t;j} w_{js} + \sum_k h^S_{t;k} w_{ks} + b^A_s\right); \quad (5.1)$$

where $w_{\cdot s}$ is the connection weight between each top hidden unit of the sensory RBMs and the s-th hidden unit of the association RBM, and $b^A_s$ is a bias parameter.

Each RBM network of this model was separately trained and staked from a lower layer to higher layer. The Gaussian{Bernoulli and the Bernoulli{Bernoulli RBMs used Eq. (3.6){Eq. (3.8) and Eq. (3.12){Eq. (3.15) to learn the model parameters, respectively. For details of the training method, see [58].

### 5.2.3 Mechanism of mental simulation in the proposed model

It is known that MNS generates motor signals from observation signals based on self-experience [31]. This simulation mechanism of MNS, called mental simulation, facilitates comprehension of others' actions. Research in recent years has shown that the mechanism also relates to the understanding of others' internal mental states, including emotion [32, 33]. Humans estimate the emotional states of others based not only on the perception of others' expressions, but also their own experiences of expression and personal knowledge. Oberman et al. [107] reported that when subjects stopped the mimicking of another person's facial expressions, the recognition of emotional states of others was also disrupted. This result shows that use of self-generated information improves the accuracy of the estimation of the emotional states of others.

Our model demonstrates this capability for mental simulation by using the generative ability of a multimodal deep belief net. The model receives multimodal signals (i.e., visual signals and audio signals, as shown in Fig. 4.1) from a partner during human{robot interaction. The model can estimate others' emotional states from the multimodal signals through forward sampling from each stacked RBM to an association RBM (represented by red arrows in Fig. 5.1). The model can also generate its own multimodal expressions using the same network structure through backward sampling from the top layer to each visible layer of the stacked RBMs (represented

by green arrows). Based on these mechanisms, the model can reconstruct the absence of the partner's multimodal expressions as virtual observation signals from observed signals through the forward and backward samplings. For instance, assume that facial signals are absent (i.e., $v^F = 0$) in the interaction. The model reconstructs unobserved facial signals (i.e., $v^F = 0 \rightarrow v^F = \hat{v}^F$) from other modalities (i.e., hand movements and speech) via the incompletely estimated state of the other. Then, the model is able to update its belief of estimation of another's emotion via repeated samplings by using the reconstructed signals with the observed signals. We propose an energy-based sampling method to update estimation of others' emotions by reconstructing imaginary signals from inputted signals. RBMs acquire the energy function of the combination of visible and hidden activations (Eq. 3.4 and Eq. 3.11). This energy corresponds to the frequency of data, and frequent combinations in the RBM produce low energies (Eq. 3.3). We utilize this energy for updating estimation of others' emotions. When a new sampled emotion representation in the top RBM of the proposed model shows lower energy than the previous representation, the model accepts the new representation. Algorithm 1 provides the details of the our proposed sampling method, where $v_o$, $v_u$, and $\hat{v}_u$ are observed signals, unobservable signals, and reconstructed signals for the lowest visible units of the network, respectively, and $v_s$ and $h_s$ are visible and hidden activations of the top RBM of the network, respectively.

### 5.2.4 Support vector machine for quantitative evaluation of estimate emotion

Our model learns the relationships between emotional states and multimodal expressions of others without use of emotional labels in an unsupervised manner. The model is able to estimate emotion of others as activations of the top RBM and is able to compare them to each other. However, a supervised classifier is necessary for recognition of actual emotional state expressions (e.g., happiness, neutral, and anger). For quantitative evaluation of estimated emotion in the proposed model, we employ a support vector machine (SVM) as was done in [79, 80]. The SVM uses the activations of the top RBM in our model as the input and assigns emotional categories from a

Algorithm 1 Energy-based sampling

$v_s \sim p(v_s|v_o; v_u = 0)$
$h_s \sim p(h_s|v_s)$
$E \leftarrow \text{Energy}(v_s; h_s; \theta)$
$\hat{v}_u \sim p(v_u|h_s)$
while $0 < N$ do
  $\hat{v}_s \sim p(\hat{v}_s|v_o; \hat{v}_u)$
  $\hat{h}_s \sim p(\hat{h}_s|\hat{v}_s)$
  $\hat{E} \leftarrow \text{Energy}(\hat{v}_s; \hat{h}_s; \theta)$
  if $\hat{E} < E$ then
    $h_s \leftarrow \hat{h}_s$
  end if
  $\hat{v}_u \sim p(v_u|h_s)$
  $N \leftarrow N - 1$
end whilereturn $h_s$

choice of four emotions. We use the default parameter of LIBSVM [108] with linear kernels for training and evaluation.

## 5.3 Experiments and results

In this section, we introduce our experimental settings and report the experimental results. The main objective of the proposed model is that the robot with our model imitates the expressions of the interaction partner based on estimation of his/her emotion even when the partner's expressions are ambiguous or absent in the interaction.

We evaluate the abilities of our model for imitation of human emotional expressions via three experiments: 1) constructing an emotional representation from multimodal signals of humans, 2) comparing imitated expressions of our model with those of the direct mapping method, and 3) evaluation of the mental simulation capability in emotion estimation of a human from ambiguous signals. We examine the feature extraction capability of our model, which enables the robot to imitate the expressions of humans based on the estimation of their emotion in the first experiment. In the second experiment, we compare the imitation capabilities between our model and the direct mapping system when human expressions exhibit conflicts between

modalities. We examine the mental simulation capability for updating the emotion estimation from partial observation of human expressions in the imitation task in the third experiment.

We first explain the dataset used in our experiments as well as its multimodal features and parameter settings for the proposed model. Following this, we report the results of three experiments.

### 5.3.1 Experimental settings

#### IEMOCAP dataset as interaction data

We used the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [2] as interaction data for our experiments. The IEMOCAP database contains audiovisual data from ten actors who acted in both scripted and improvised scenarios. Facial expressions and hand movements were recorded with a motion-capture system. Fig. 6.2 shows a sample of the motion-capture markers. Subjects attached 53 markers to their faces and six markers on their hands. Conversations were also recorded using video cameras.

All recorded data were evaluated using categorical labels and continuous three-dimensional values. More than three evaluators annotated each utterance of the interactions. We employed the data corresponding to four specific emotions: happiness, neutral, anger, and sadness (as done by [79, 80]). We selected only the majority voted emotional category as the ground truth for each utterance if the annotations were not agreed to among the evaluators. For instance, when two evaluators voted the sad category whereas one evaluator voted the neutral category for a focused utterance, we set the emotional category as sadness for the ground truth. We selected three actors' data as the training dataset and one actor's data as the test dataset. The total number of training datasets was 634 (happiness:106, neutral:171, anger:164, sadness:193) and that of test datasets was 202 (happiness:31, neutral:70, anger:40, sadness:61).

### Feature extraction from IEMOCAP dataset

We calculated the basic audiovisual features from the IEMOCAP database for experiments in two steps. First, we extracted modality dependent features. Facial features included 34 distances between markers and their difference from the prior time step. Each green line in Fig. 6.2 corresponds to a feature. Each distance is represented in two-dimensional space (i.e., x, z) because markers do not conspicuously change regarding the y-coordinate. Audio features include pitch, intensity, 13-dimensional Mel-frequency filter banks (MFBs), and their difference from the prior time step.

Moreover, we used hand movements as features in contrast to previous studies [79, 80], as it is known that humans consider visual expressions to a greater extent than audio expressions when the interacting partner's expressions conflict [109]. We assumed that hand movements also represent emotional states, for instance, very rapid hand movements may represent strong angry states or happy states. Hand movement features consist of the velocities of four markers (LH2, LH3, RH2, and RH3 in Fig. 6.2) and their difference from the prior time step. Each velocity is expressed in three dimensions (i.e., x, y, and z).

Second, we calculated statistical features from all modality dependent features in each utterance. The statistics included the mean, variance, range, maximum, and minimum. All features were normalized using z-normalization. The mean and standard deviation for normalization were calculated over all of the training data. We used the same mean and standard deviation values to normalize both the training dataset and the test dataset. Finally, the number of facial expression features, hand movement features, and audio features obtained represented 680, 120, and 150 dimensions and the total number of extracted features was 950.

### Parameter settings of the proposed model

Our model structure has already been shown in Fig. 5.1 and described in Section. 5.2.2. Each modality network has three layers and one RBM combines all modality networks. For the facial expression network, we set the number of visible, first-hidden, and second-hidden units to 680, 300, and 100, respectively. The number of visible units for the hand movement and the audio network were set to 120 and 150,

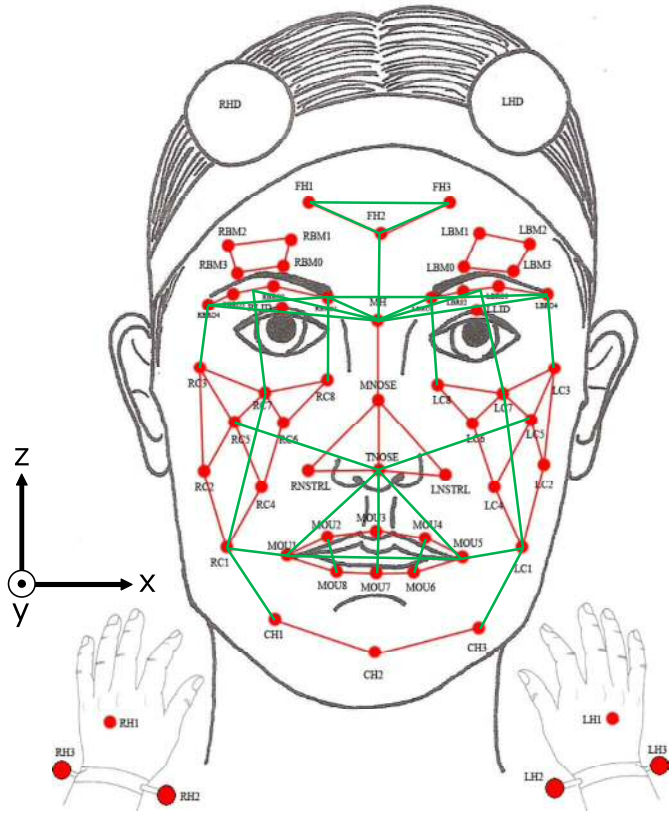



Figure 5.2: Motion capture marker-set of the IEMOCAP database. The vertical axis represents the x direction and the horizontal axis shows the z direction (modified from [2]).

respectively. We set both the number of first- and second- hidden units for both the hand movement and the audio network to 100. The number of visible and hidden nodes of the highest RBM were 300 and 50, respectively. The parameters for model training, $\epsilon_w$, $\epsilon_a$, $\epsilon_b$, and $\epsilon_z$ were set to 0.001 for Gaussian{Bernoulli RBMs, and all parameters except $\epsilon_z$ were set to 0.01 for Bernoulli{Bernoulli RBMs.

### 5.3.2 Experiment I: Constructing emotional representation of multimodal signals by self-organization

To imitate the emotional expressions of others based on the estimation of their emotional states, it is important to extract low-dimensional features from high-dimensional multimodal signals. We assumed that each modality signal contained distinctive features corresponding to emotion, and that humans' emotional states can be represented by combinations of these distinctive features. To examine the capability of the RBMs to extract relevant features from multimodal signals, we examined the emotional representations acquired in our model.

First, each modality-specific network was trained separately. Subsequently, their outputs were concatenated as input data for the association RBM, which was then trained. The maximum number of training steps was 10,000 for all RBMs. Following the training of all RBMs, we calculated the activation in the hidden layer of the association RBM through forward sampling (red arrows in Fig. 5.1) by using the training dataset, and carried out PCA to visualize the activation in low-dimensional space.

Fig. 5.3 shows (a) the first and second PCs space and (b) the first and third PCs space of highest layer of our model. The contribution ratio of each component was 25.0%, 13.2%, and 9.37%, respectively. The color and shape of the markers in Fig. 5.3(a) represent emotional states of each activation calculated from training data. Note that the model did not use these labeled categories of emotion of data for training. The distribution of each emotional data item overlapped with those of others. In particular, neutral and happiness data spread across the entire PC space; however, anger and sadness data formed a corresponding emotional cluster in this space, and the distribution of happiness data was biased toward that of data related to anger. We confirmed that the distribution of the emotional representation showed gradual change, which related to the intensity of the emotional states from the upper-left corner (i.e., sadness, which usually has low intensity) to the lower-right corner (i.e., anger, which usually contains high-intensity features). In Fig. 5.3(b), the color and shape correspond to individual data. This figure shows that personal expressions were clearly differentiated in the third principal component. This feature can assist in

considering social characteristics (e.g., cultural differences and different personalities evincing emotional expressions) for emotion estimation.

These results indicate that the proposed model can extract low-dimensional representations of multimodal emotional expressions at the highest layer (50 dimensions) from high-dimensional multimodal signals (a total of 950 dimensions) in an unsupervised manner. The two main components of the emotional representation (i.e., the first and the second PCs) showed that data for angry and sad expressions were concentrated around each other and formed clusters. The distributions of both clusters represented the variation in the intensity of the relevant emotion. Moreover, PC3 represented the personalities of multimodal expressions. These features of the acquired representation can help to imitate a partner's emotional expression from the estimation of his/her emotional state.

### 5.3.3 Experiment II: Comparing imitated expressions via our model with the direct mapping method

In this experiment, we examined the capability of the proposed model to imitate emotional expressions of others and compared it with that of the direct mapping system, which directly copies others' expressions to those of a robot. Our motivation for this experiment was to examine this capability when the same person's expressions from different modalities were in conflict. The difference between these two systems is that the proposed model generated its own expressions through the estimation of other's emotional states.

Our model first estimates the emotional state of others in the top layer from the multimodal expressions of others as input signals through forward sampling (red arrows in Fig. 5.1). The model then generates multimodal signals by backward sampling (green arrows in Fig. 5.1) to imitate others' expressions based on the estimated emotion. There was no guarantee that the generated multimodal expressions from the model would be identical to the corresponding human expressions, as the multimodal signals interacted with other modalities through the sampling conducted by the association RBM. However, the generated multimodal expressions were based on the same emotional state, which was estimated by the model even when human

expressions exhibited emotional inconsistency in modalities. By contrast, the direct mapping system only copied the multimodal expressions of others as its own emotional expressions. The multimodal expressions of the robot were always the same as those of the humans in all modalities because the system operated the multimodal signals individually.

We used a humanoid robot iCub to express emotions in this experiment. The reconstructed multimodal signals through backward sampling were converted into expressions for the iCub. Regarding facial expression, the iCub has two components expressing emotional states: eyebrows and mouth. Twenty patterns can be expressed by combining four eyebrow patterns and five mouth expression patterns. Each expression pattern was selected by using the corresponding feature values. Because iCub can only express discrete patterns, facial features were quantified using thresholds. For instance, the open or closed mouth pattern was selected via the distance between the center marker of the upper lip (MOU3) and the lower lip (MOU7) and the angle of the mouth was decided by the angle between the edge of the mouth (MOU5) and the immediate marker of the cheek (LC1) in Fig. 6.2. For hand movement and the audio expression, we prepared prototype patterns (e.g., saying "hello" and moving the hand in a cyclic manner) and only modulated their parameters to imitate the partner. For example, we used the mean of the pitch and intensity from the auditory features to modify the prosody of the iCub's speech. The speed of the iCub's hand movement was modulated by the mean velocity of the partners corresponding hand movement features. In the imitation experiments with our model, the iCub used sensory features generated by backward sampling as its own expressions. On the other hand, the iCub used extracted sensory features from other's expressions directory as own expressions for the experiments with the direct mapping system.

Fig. 5.4 shows one example of our experimental results using the test dataset. Fig. 5.4(a) is a screenshot of a human emotional expression from the IEMOCAP dataset. The subject's expression for this utterance in the interaction was labeled as an angry state in the dataset. However, recognizing the emotional state of the subject as anger from this figure is difficult. Considering only the visual image, the state appears to be surprise or happiness. Fig. 5.4(b) shows the resultant expression of the direct mapping system. The iCub expressed a surprised face with an opening

mouth instead of an angry face. Because the direct mapping method was not able to consider inconsistencies between modalities, there was a mismatch between the iCub's expression and the other's actual emotional state. In contrast, Fig. 5.4(c) shows the imitation result via estimation of the other's emotion by our model. In this figure, the iCub shows a different facial expression from that of the direct mapping experiment, even though the input signal was the same. The estimated state of the other's multimodal expressions in PC1-2 space is presented as a cross marker in Fig. 5.5. The figure shows that the estimation was close to the anger distribution, achieved by considering a combination of the multimodal signals. These results show that the proposed model can imitate expressions of others that correspond to their emotional state even when their expressions include inconsistencies between sensory signals.

### 5.3.4 Experiment III: Evaluation of mental simulation capability in emotion estimation from partial multimodal signals of partners

We conducted a third experiment to examine the mental simulation capability in our model for imitation of multimodal expressions of others from partial observations. The experimental settings and data used were similar to those in previous experiments; however, the input signals lacked portions of their modalities. For example, the robot could not perceive the partner's facial expressions and/or hand movements for emotion estimation because the robot attended to other people or objects. We assumed that the model did not receive any signals for the absent modalities, and set signals for these modalities as zero vectors 0.

We first evaluated the performance of emotion estimation by the proposed model with a linear SVM using complete multimodal signals. Here, the complete multimodal signal means that no modalities signals are lacking in the dataset (i.e., the entire set of multimodal signals was input). The SVM was trained with the training dataset and corresponding emotional labels in a supervised manner. Training of the proposed model and the SVM and the evaluation of emotion estimation were carried out ten times each for the validation test.

Table 5.1 presents the confusion matrix of emotion estimation. The column labels

indicate the actual emotional state of input data, and row labels show estimated emotion labels. The correct estimation rates exceeded the random chance rate (25:0%), and the general estimation accuracy achieved an average of 59:6%. The accuracy in previous studies [79, 80] was approximately 65% for the general estimation. Our results are similar to previous studies; nevertheless we did not conduct parameter tuning for the proposed model and the SVM.

The correct estimations for happiness and neutral states are lower than for anger and sadness. This result may reflect the overlap of happiness and neutral states in the PC space (Figure 5.3).

Table 5.1: Estimation accuracy (%) of others' emotion from complete multimodal expressions.

| | Happiness | Neutral | Anger | Sadness |
|---|---|---|---|---|
| Happiness | 52.6 | 14.5 | 20.0 | 12.9 |
| Neutral | 5.00 | 33.6 | 29.6 | 31.9 |
| Anger | 10.8 | 5.00 | 83.6 | 0.75 |
| Sadness | 4.26 | 21.5 | 5.4 | 68.9 |

In a second experiment, the proposed model received partial signals from others (e.g., only other's speech for example). Our model reconstructed unobservable signals from other observed expressions to refine accuracy of the estimate based on energy-based sampling (Algorithm 1). For instance, under a condition without the signals for facial expression, the proposed model reconstructed the signals for hand movements and speech. Subsequently, the model updated the estimate of another's emotion using the actual signals with the reconstructed signals by mental simulation for N iterations. We set a maximum number of sampling iterations N and unobservable signals as 100 and 0, respectively. Experiments under all conditions were executed ten times by changing initial parameters.

Table 5.2 shows the correct answer rate for each emotion, and Figure 5.6 shows the average estimation rate under each condition, where the labels denote the signals presented to the model. For example, the "Han.–Fac." condition signified that the model received hand movements and facial expressions from the dataset, and the

"Aud." condition signified that the model received only audio signals. Two-modality conditions (i.e., Fac.{Han., Han.{Aud. and Fac.{Aud.) provided fewer errors than single-modality conditions (i.e., Fac., Han., and Aud. conditions). In the single-modality conditions, the audio signal produced the lowest error. Subsequently, the error increased from the Fac. condition to the Han. condition. This result indicates that auditory signals contained information more relevant to emotion in the current experiment. From these results, we noticed that hand signals did not correspond to emotional states directly because they relate to the context of the interaction (e.g., passing objects). In the IEMOCAP dataset, some interactions simulated routine office situations, which is why they were omitted in previous studies [79, 80]. Facial signals also relate less to the scenarios of interaction less. Mouth movements as sole inputs are more strongly affected by the contents of speech rather emotion. Under the two-modality conditions, the errors in the Han.-And. and Fac.-Aud. conditions had similar values, and were lower than those of the Fac.-Han. condition. These results did not contradict previous results [79]. Lack of auditory signals led to noisier estimation than in other conditions because facial and hand features contain large variances for emotion estimation. These results indicated that auditory signals contribute more than other modalities to the emotion estimation in these settings.

We compared the results with the update of emotion estimation (ES) to without the update (BL: baseline) to assess the role of the signal sampling presented in Section 5.2.3. The student's t-test was conducted between BL and ES under each condition individually. Figure 5.6 indicates that the estimation rates under the conditions except for the Fac. condition increased from those of the baseline. Aud., Han.{ Aud., and Fac.–Aud. conditions exhibit statistically significant results: $t(18) = 3{:}36$, $p < 0{:}01$, $t(18) = 3{:}67$, $p < 0{:}01$, and $t(18) = 3{:}18$, $p < 0{:}01$, respectively. Han., Han.{ Fac., and Complete conditions exhibit statistically significant results: $t(18) = 2{:}57$, $p < 0{:}05$, $t(18) = 2{:}69$, $p < 0{:}05$, and $t(18) = 2{:}16$, $p < 0{:}05$, respectively, and the Fac. condition exhibited a non-significant result: $t(18) = 1{:}00$, $p = 0{:}33$.

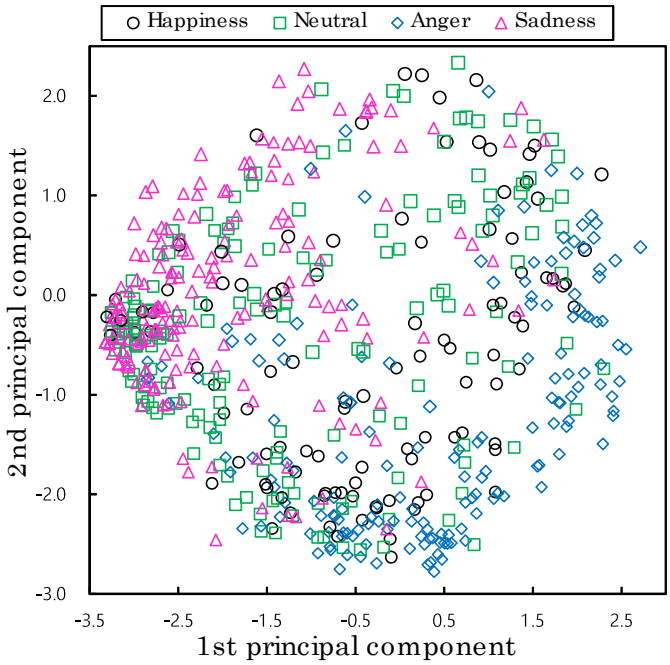
Most estimation performances of sadness were increased by the reconstruction of unobservable expressions (Table 5.2). However, the estimation accuracy of other states with the proposed sampling method sometimes decreased compared with the accuracy of the baseline because energy-based sampling considers only minimization

of the energy of the top RBM's activations without considering lower layer activations. This phenomenon is similar to the preconception of estimation based on self-experience in humans. When the received expression pattern is not similar to learned patterns that belong to the same emotional state category of received signals, the estimation may decrease using energy-based sampling. However, when the patterns of others' expressions are close to one's self-knowledge of expressions, the estimation will increase by using the reconstruction of unobservable signals.

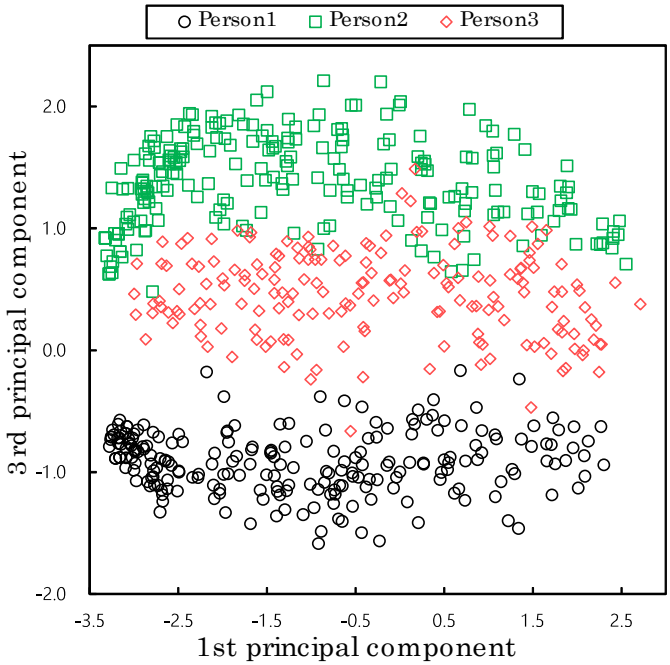
Figure 5.7 illustrates an example of the transition of the emotion representation over the energy-based sampling under the Aud. condition. In addition, Figure 5.9 shows changes in the distance between the sampled emotion representation and the ground truth, and the energy of the top RBM. The ground truth was calculated as the mean value of 100 sampled representations when the model received the entire set of multimodal signals from the test dataset. The original emotion label of the example was anger; however, the extracted emotion representation from the data was close to the distribution of sadness and neutral at the first sampling step. Throughout the energy-based sampling, the emotion representation came closer to the ground truth and the distribution of the anger state. Moreover the energy of the top RBM was reduced by reconstructing the unobserved signals from the observed ones.

Additionally, Fig. 5.8 shows the trajectory of estimation in 20 steps through mental simulation. The original emotion label is anger as for the previous example. The change in distance from the ground truth over the mental simulation is illustrated in Fig. 5.10. Fig. 5.11 depicts reconstructed facial expressions through the mental simulation, and shows expressions in the initial, fifth, seventh, and 15th steps corresponding to Figs. 5.8 and 5.10. In this interaction, the subject expressed a very strong angry face and a loud voice; therefore, he mostly opened his mouth (Fig. 5.11(a)). Our model reconstructed an expression that looks like a weak anger face from audio and hand signals in the first step, and then through mental simulation, the model reconstructed an angry face in the fifth step. Through additional mental simulation, the facial expression became an angry face with open mouth, and the estimation became even closer to the ground truth. In the 15th step, the mouth had opened wider than the reconstruction in seventh step, and the estimated expression was closer to the ground truth. These results demonstrate that our model can imitate

expressions of others by updating the emotion estimation of others based on mental simulation even when some expressions of partners were absent.




(a) First and second PC space




(b) First and third PC space

Figure 5.3: Principal component space of the acquired representation of multimodal signals in the highest hidden layer by self-organized learning in our model. (a) Each color and shape of the markers indicate the emotional state from interaction data (b) Each marker in the space corresponds to humans who expressed the data.

(a) Human expression

(b) iCub's expression generated by the direct mapping system

(c) iCub's expression generated by the proposed model

Figure 5.4: Example of expressions obtained in imitation experiments from the direct mapping system or the proposed model. In (a), the human expressed an angry emotion in the interaction. Note that the model did not learn the expressions of the human.

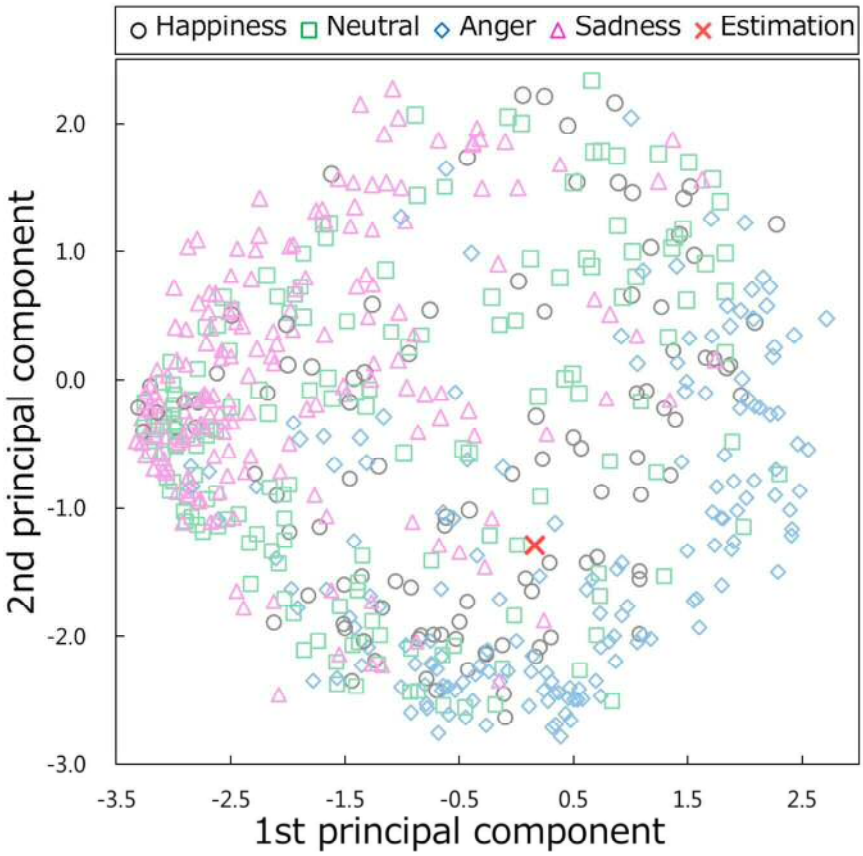



Figure 5.5: Generated emotional representation of multimodal expressions from the example data (Fig. 5.4(a)) in the PC1-2 space, which was acquired by Experiment I (Fig. 5.3(a)). The estimation is highlighted by the red cross.

Table 5.2: Comparison of estimation rates between baseline and energy based sampling conditions (%). Bold numbers indicate best results.

| Observation | Han. | | Fac. | | Aud. | | Han.{Fac. | | Han.{Aud. | | Fac.{Aud. | | Complete | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | BL | ES | BL | ES | BL | ES | BL | ES | BL | ES | BL | ES | BL | ES |
| Happiness | 1.29 | 8.06 | 41.9 | 44.2 | 6.78 | 21.6 | 35.5 | 31.9 | 3.23 | 11.9 | 48.1 | 48.7 | 52.6 | 50.6 |
| Neutral | 48.7 | 26.4 | 34.0 | 28.7 | 48.0 | 31.7 | 38.2 | 21.6 | 53.3 | 39.7 | 34.4 | 34.6 | 33.6 | 35.4 |
| Anger | 3.00 | 4.25 | 50.3 | 19.3 | 55.6 | 84.5 | 20.0 | 39.7 | 53.5 | 78.3 | 80.0 | 87.0 | 83.6 | 87.3 |
| Sadness | 31.3 | 54.0 | 42.3 | 77.7 | 59.0 | 53.3 | 54.1 | 78.3 | 59.7 | 55.1 | 64.8 | 71.1 | 68.9 | 73.9 |
| Average | 21.1 | 23.2 | 42.1 | 42.5 | 42.3 | 47.8 | 37.0 | 55.0 | 42.3 | 46.2 | 56.8 | 60.4 | 59.6 | 61.8 |

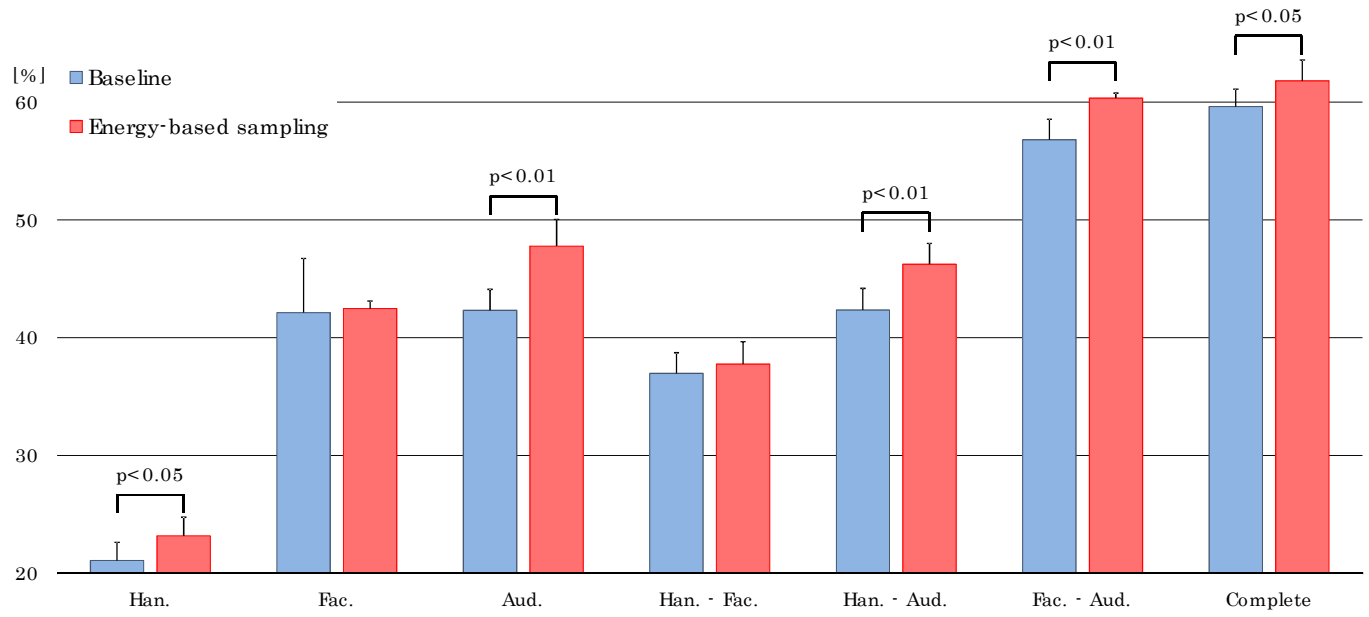



Figure 5.6: General estimation rate under each condition corresponding to observed data.

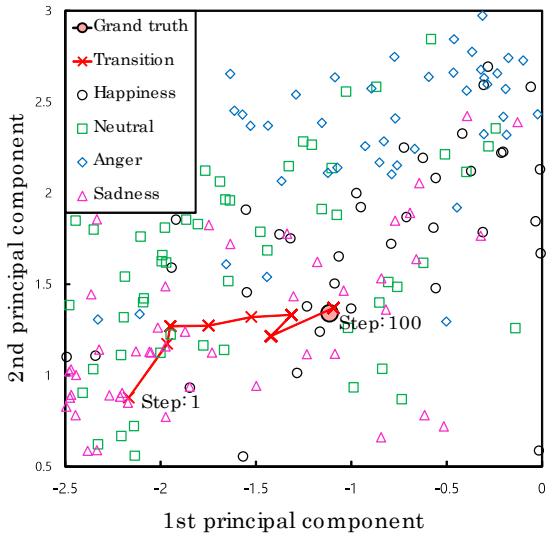



Figure 5.7: Example trajectory of the estimation in the PC1-2 space through mental simulation under the Aud. condition.

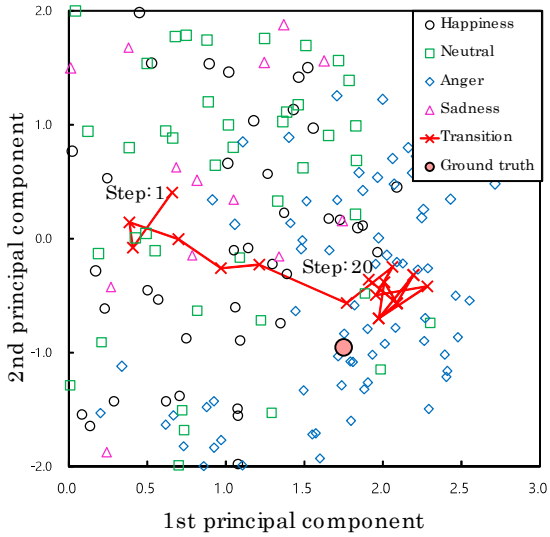



Figure 5.8: Example trajectory of the estimation in the PC1-2 space through mental simulation under the Han.-Aud condition.

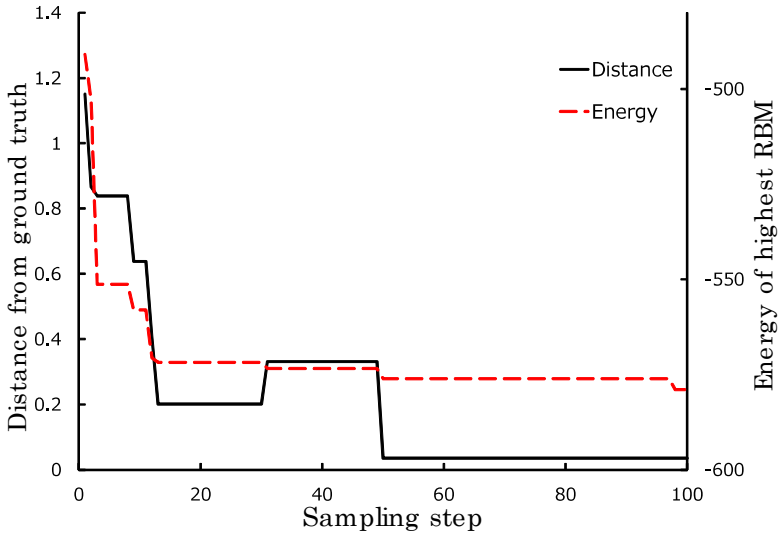



Figure 5.9: Example of change in distance between ground truth and the energy of the top RBM of the proposed model.

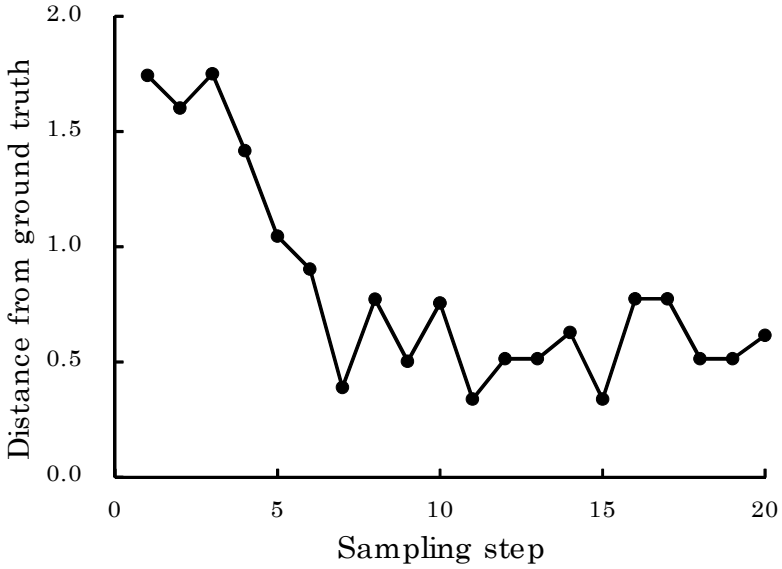



Figure 5.10: Change of distance between ground truth and the estimation corresponding to Fig. 5.7.

(a) Actual human expression (b) 1st step (c) 5th step (d) 7th step (e) 15th step

Figure 5.11: Example of the partner's facial expression and reconstructed facial expressions by the mental simulation. (a) A screenshot of the actual human expression in this interaction. (b)-(e) Robot's facial expressions reconstructed by the mental simulation mechanism from the partner's hand movements and audio expressions at initial, fifth, seventh, and fifteenth steps in the simulation.

## 5.4 Discussion and Future Work

### 5.4.1 Acquiring emotional representation of multimodal expressions

The proposed model was able to extract a set of features from multimodal general features using the capability of RBMs. This is in contrast to Lim's model [44], which uses ad hoc features. It is known that the STS | a region of the human brain that belongs to MNS | processes feature selection and multimodal integration for emotional categorization [22{24]. We did not consider the exact structure and mechanisms of these areas in our model; however, our model could reproduce similar functions using the abilities of the RBMs.

On the other hand, the categories of emotion in the PC1-2 space did not differentiate from one another, although the space represents a gradual change in the distribution of emotion. In a psychology study, Russell [61] proposed a circumplex model of affect that represents emotional states in two-dimensional space comprising a positive/ negative axis and an arousal/ sleep axis. This representation of emotion in low-dimensional space is useful for not only the analysis of human emotion but also for modeling robots emotional states for human{robot interaction (HRI). The gradual change in our model relates to the intensity of emotion in the PC1-2 space (see Section 5.3.2 and Fig. 5.3(a)). It appears that our model acquired the arousal/ sleep axis of Russell's model as the distribution in the PCs space in an unsupervised manner. This result suggests that there is a possibility that communication robots can acquire the relationship between emotions from only observing human emotional expressions. We used emotion labels to visualize the distribution of the representations, but the model did not require labeled input data. On the other hand, the other axis of Russell's model, the positive/ negative one, did not appear to exist in the PC space. To obtain this axis, the model has to perceive the emotional valence of expressions (e.g., smiling face is a positive value, or a loud sound is a negative value for robots). Our previous study showed that the proposed model was able to obtain a feature space, which could isolate emotional states, similar to Russell's emotional model, by considering emotional valences based on tactile stimuli in interaction. That model used

emotional valences as low-order emotion labels (i.e., positive, neutral, or negative) in a semi-supervised manner. Including the information from sensory signals with the current features in real HRI may aid in acquiring clear emotional representation of multimodal expressions and it may enhance the performance of emotion estimation and expression.

### 5.4.2 Limitations of the mental simulation mechanism in our model

The results of Experiment 3 indicate that the performance of estimation after reconstructing the unobservable signals increased after the first estimation of other's emotion from partial expressions. However, the mental simulation mechanism does not always lead to accurate estimation of others' emotions. The estimation of neutral expressions is confused as different emotion by using reconstructed signals. The reason for such results is that the neutral data includes considerable variation in expressions and their distribution spreads over the entire representation space. It suggests that robots with our model estimate others' emotions more emotionally (i.e., extremely) by putting themselves in another's shoes.

The energy-based sampling method corresponds to that of selection based on proximity to self-experience, similar to prejudice. Humans do not employ mental simulation data if the observed signals are sufficient for estimation. Humans execute mental simulations when the belief of estimation changes saliently. From our results, the enhancement of belief is not directly linked to accurate estimation, and induces incorrect estimation, such as prejudice would. However, neither the model nor humans can determine the absolute emotions of others, so the ground truth does not actually exist in human{human or human{robot interaction.

# Chapter 6

# Active Perception based on Energy Minimization

Humans use various types of modalities to express their internal states. If a robot interacting with humans can pay attention to limited signals, it should select the more informative ones to estimate the partner's states. We propose an active perception method that controls the robot's attention based on an energy minimization criterion. An energy-based model, which has learned to estimate the latent state from sensory signals, calculates energy values corresponding to occurrence probabilities of the signals; the lower the energy is, the higher the likelihood of a signal being correctly interpreted. Our method therefore selects the modality that provides the lowest expectation energy among available ones to exploit more frequent experiences. We employ a multimodal deep belief network to represent relationships between humans' states and expressions. Our method demonstrated better performance for the modality selection than other methods in a task involving emotion estimation. We discuss the potential of our method to advance human{robot interaction.

## 6.1 Background

Object properties such as categories and/or states are observed and recognized from various types of modality signals. For instance, visual information tells us what a

bottle is, and sound signals enable us to distinguish empty bottles from filled bottles. Similarly, humans' internal states (e.g., intentions and emotions) are expressed through many types of signals, such as facial expressions, gestures, vocalization, etc., (Fig. 4.1). During communication, we humans try to estimate others' internal states by observing their expressions. However, it is not always true that every modality signal links to the actual state of the target. Some signals might be noisy or ambiguous. If a robot manipulating objects and/ or interacting with humans has limited resources to access multimodal signals (e.g., time for estimation and cost for perception), it should select the more informative ones among the available ones to estimate the target states. For instance, if the robot can obtain only a single modality signal by acting on a target, it must properly select the next action to induce more effective information about the target.

In the robotics field, the issue of attention selection is formulated as active perception [110, 111]. Those studies assumed that one action allows the system to perceive only a single sensory signal (i.e., one to one mapping) and the system cannot execute more than one action at the same time. Therefore, the system should select actions one by one. For instance, Sakaguchi [110] proposed a haptic sensing system that estimated object categories based on sensory integration and the active perception method using tactile sensors. His active perception method used mutual information between the object category and the i-th sensory signal to select the next perception from the set of tactile sensors (e.g., pressure-sensor, thermo-sensor, and vibration-sensor). The mutual information tells the system by how much the entropy of the object category probabilities is reduced when the system observes the object using the i-th sensor. His system selects the sensor that indicates the maximum mutual information. The proposed method demonstrated better performance than a random selection strategy; the recognition accuracy was improved and the number of observation times was reduced by the method. Taniguchi et al. [111] focused on active perception in an object recognition task based on multimodal sensory signals. They represented the relationship between sensory perceptions of the robot and object categories by using a Bayesian probabilistic model called the multimodal hierarchical Dirichlet process and proposed an active perception method from the viewpoint of

information theory. Their method selects the next perception modality that maximizes the information gain between the present estimation and the expectation of the estimation for unobserved signals. Their experimental results showed that the proposed method achieved faster and more accurate recognition of objects.

Our research aim is to apply an active perception method to the estimation of others' internal states in multimodal human{robot interactions, as shown in Fig. 4.1. This section proposes an active perception method based on an energy minimization strategy using an energy-based model. The energy values of the model correspond to the probabilities of the data. The proposed method selects the "attention" to the modality that minimizes the expectation of the energy values among all selectable modalities. The reason why the proposed method uses the energy minimization criterion is that a lower energy indicates a higher likelihood of the data in the energy-based model. We employed a multimodal deep belief network (MDBN), which is a type of energy-based model and used in the previous chapter, to represent the relationship between humans' internal states and their expressions by abstracting and integrating multimodal signals. Our model was applied to a task to estimate others' emotional states in multimodal interactions. As a first step, we utilized the IEMOCAP dataset [2], which is a multimodal human{human interaction dataset, for training the human{robot interaction model (i.e., MDBN) and evaluating the proposed active perception method. Experimental results showed that the proposed method achieved higher estimation accuracy than other methods. Finally, we discuss the relationship between the energy minimization method and the information gain maximization method, which has been used in other studies.

## 6.2 Energy-based Model for Modeling Multimodal Interactions

This section introduces the energy-based model, which is a RBM [58, 60] and the MDBN [83], to represent the relationship between humans' internal states and their multimodal expressions. Figs 3.3 shows the structure of the RBM, which is a key mechanism in our artificial emotion model. The structure of the MDBN is illustrated

in 5.1. The MDBN consists of two parts: stacked RBMs for each modality in the lower layers and an RBM at the top layer that associates the lower RBMs' outputs. First, we provide details regarding the MDBN, and then we indicate the characteristics of the energy-based model and the energy function for the proposed active perception method.

### 6.2.1 Multimodal deep belief network

We consider an association of multimodal signals to acquire the relationships between humans' internal states and their own expressions. We assume that the interaction partners use a set of N modalities (i.e., $M = \{m_1; \cdots ; m_n; \cdots ; m_N\}; |M| = N$) such as facial expressions, vocalization, gestures, etc. during interactions to express their own internal states (Fig. 4.1). First, two different types of RBMs are stacked to extract sensory features from input signals of each modality (i.e., from $m_1$ to $m_N$): the lower RBM being Gaussian{Bernoulli and the upper being Bernoulli{Bernoulli. Then, the top layers of all stacked RBMs are connected to an additional Bernoulli{ Bernoulli RBM (see Figure 5.1).

Let $h_n^2 \in \{0; 1\}^{J_n}$ denotes the second hidden layer (i.e., the top layer of the stacked RBM) activations of the n-th modality ($m_n$) network. The activation probability of the k-th hidden unit $h_k \in \{0; 1\}$ of the association RBM is given by replacing v of the Bernoulli{Bernoulli RBM with $h^2 = \{h_1^2; \cdots ; h_N^2\}$. We acquire the probability by modifying Eq. (3.1), resulting in Eq.(6.1),

$$\begin{aligned} &p(h_k = 1 | h^2 = \{h_1^2; \cdots ; h_N^2\}) = \\ &sig\Big(\sum_j^{J_1} h_{1;j}^2 w_{jk} + \cdots + \sum_j^{J_N} h_{N;j}^2 w_{jk} + c_k\Big); \end{aligned} \tag{6.1}$$

where $w_{jk}$ is the connection weight between the k-th unit of the highest layer and the j-th unit of each top layer of the sensory RBMs, and $c_k$ is a bias parameter.

### 6.2.2 Energy Function for the proposed method

The data probability of the energy-based model such as the RBM is represented by the Boltzmann distribution, which is given as

$$p(X;\theta) = \frac{1}{\mathcal{Z}(\theta)} \exp\bigl(-E(X;\theta)\bigr); \tag{6.2}$$

where $X$ is data, and $\mathcal{Z}(\theta)$ is a partition function. $E(X;\theta)$ is the energy function, which assigns the energy value for corresponding data $X$ based on the parameter $\theta$. Eq. (6.2) states that the occurrence probability of the data indicating the lower energy is higher than the data indicating the higher energy. The energy-based model modulates the parameter $\theta$ to minimize the energy values of training data, so that the energy of experienced data (i.e., training data) becomes small, and their probabilities become high to relate the above characteristic.

The energy function of the association RBM of our model is defined as

$$\begin{aligned} &E(X = \{h_1^2;\cdots;h_N^2;h\};\theta) = \\ &-\sum_j^{J_1} b_j h_{1;j}^2 - \sum_j^{J_1}\sum_k h_{1;j}^2 h_k w_{jk} - \cdots \\ &-\sum_j^{J_N} b_j h_{N;j}^2 - \sum_j^{J_N}\sum_k h_{N;j}^2 h_k w_{jk} \\ &-\sum_k c_k h_k: \end{aligned} \tag{6.3}$$

We use this energy function value as the criterion for the proposed active perception method.

## 6.3 Active Perception based on Energy Minimization

In this section, we describe our active perception method based on energy minimization on the MDBN. The important point of our proposed method is that the model selects the next modality that minimizes the expectation energy to the greatest extent in the entire unobserved modality. As described above, the energy value of the

energy-based model represents the frequency of the corresponding data. In other words, the modality signal, which leads to the lowest energy in the set of modality, has the highest likelihood for the current estimation. To select the most efficient modality signal for estimation of another's state, the proposed model performs active perception based on the above characteristics.

The interaction partners express their own states by using a modality set M. Let $M_r \subseteq M$ denote the subset of observed modality signals by the robot, the active perception method defined as an attention selection from a subset of unobserved signals $M \backslash M_r$.

Fig. 6.1 shows an example of the energy field, which is enclosed by the components of hidden activations of the association RBM. The blue manifold and color density denote the distribution of the network energy given by Eq. (6.3). The darker color indicates the lower energy; namely, each darker color distribution represents clusters of the training data. The small circles represent variations of the hidden activation h as the estimation of the other's state. The white nodes are hidden activations from the observed modality signals $M_r$. The gray nodes are expected hidden activations from the observed signals and the reconstructed modality signals $m_n^0$, and the red node is a truth activation from all modality signals. We introduce the outline of the proposed active perception method by using this figure. Here, the interaction partner uses four modality expressions (i.e., $M = \{m_1; m_2; m_3; m_4\}; |M| = 4$), and the model executes the active perception T times ($T \ 5 \ N - 1$). The active perception method performs the following steps.

Step 1

The model perceives the $m_{init}$ modality signal as the initial perception and adds the signal to the set of observed modality signals $M_r$ (here, $m_{init} = m_1$) and estimates the other's state $h[M_r]$ from the observed modality signals (e.g., $h[m_1]$ in Fig. 6.1(a)).

Step 2

The model reconstructs unobserved modality signals as imaginary observations $m_n^0$ (i.e., here $m_2^0$, $m_3^0$, and $m_4^0$) from the hidden activation $h[M_r]$. Then the model estimates the other's state again by using the observed modality signal

and each imaginary observation one by one (i.e., each gray node connected with the first white node: $h[m_1 + m_2^0]$, $h[m_1 + m_3^0]$, and $h[m_1 + m_4^0]$) and calculates the energy values $E_n$.

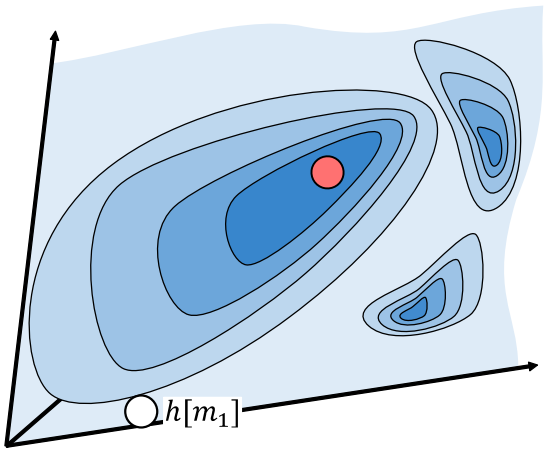
Step 3

The model selects the n-th modality, which minimizes the network energy of hidden activation $h[m_1 + m_n^0]$ most in all imaginary observations, as the next perception (e.g., $m_4$ is selected), and perceives the actual n-th modality signal. Then the model estimates the other's state $h[M_r]$ again (here $h[m_1 + m_4]$ in Fig. 6.1(c)).

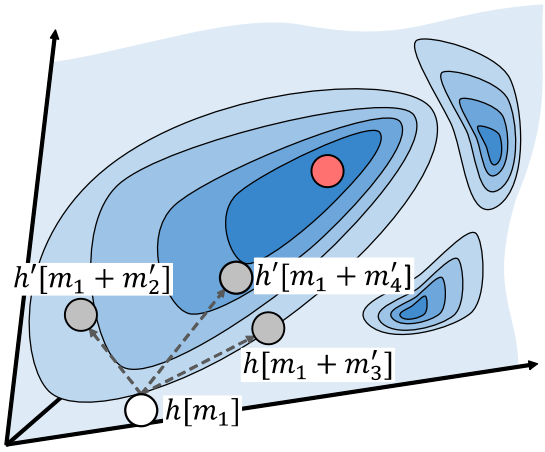
Step 4

Back to Step 2 until T iterations are reached.

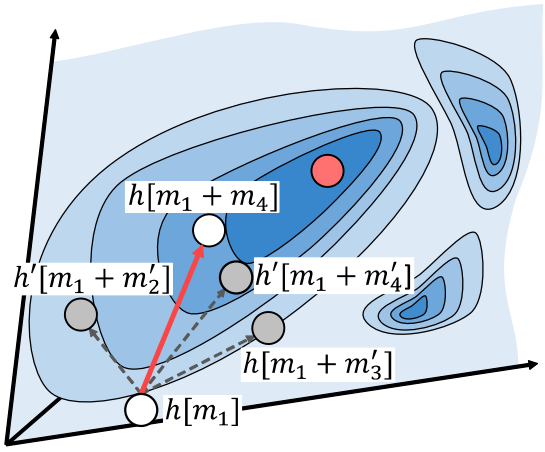
Algorithm 1 provides the details of the proposed active perception method. Here, $v_n$ and $v_n^0$ are n-th observed signals and reconstructed imaginary signals, $h_n^2$, $h$, $h_n^{2^0}$, and $h^0$ are hidden activations of the top layer of the n-th modality network, hidden activations of the association RBM, and their imaginary activations, respectively. T is the maximum number of executions of the active perception. We introduce S as the number of samples of the expectation energy used to calculate mean values.




(a) Step 1




(b) Step 2




(c) Step 3

Figure 6.1: Outline of the active perception method based on the energy minimization

Algorithm 1 Active perception based on energy minimization on MDBN

Require: T 5 N − 1

select initial modality signals $m_{init}$, and adding to set $M_r$

for t = 1 to T do

for s = 1 to S do

for n = 0 to N do

if $m_n \in M_r$ then

$h_n^2 \sim p(h_n^2|v_n)$

else

$h_n^2 = 0$

end if

end for

$h \sim p(h|h_1^2;\cdots;h_n^2;\cdots;h_N^2)$

for n = 1 to N do

if $m_n \in M_r$ then

$h_n^{2\,0} \sim p(h_n^{2\,0}|h)$

$v_n^{\,0} \sim p(v_n^{\,0}|h_n^{2\,0})$

$h_n^{2\,0} \sim p(h_n^2|v_n^{\,0})$

$h^0 \sim p(h^0|h_1^2;\cdots;h_n^{2\,0};\cdots;h_N^2)$

$E_{n;s} \leftarrow E(h_1^2;\cdots;h_n^{2\,0};\cdots;h_N^2;h^0;\theta)$

end if

end for

end for

$E_n = \frac{1}{S}\sum_s^S E_{n;s}$

$n = \text{argmin}\, E_n$

$M_r \leftarrow M_r \cup m_n$

end for

for n = 1 to N do

if $m_n \in M_r$ then

$h_n^2 \sim p(h_n^2|v_n)$

else

$h_n^2 = 0$

end if



end for

$h \sim p(h|h_1^2;\cdots;h_n^2;\cdots;h_N^2)$

return h

## 6.4 Experimental settings

This section introduces the experimental settings for evaluation of the proposed model and method in human{robot interactions. We focus on emotional states as modeling targets of humans' internal states in our experiments. First, we discuss the multimodal interaction dataset and extracted features from the dataset used in our experiments. Then we provide detailed information on our proposed model.

### 6.4.1 Multimodal Interaction Dataset: IEMOCAP

As a first step to address our research aim, we utilized the IEMOCAP dataset [2], which is a multimodal human{human interaction dataset, for a training the human{robot interaction model (i.e., MDBN) and for evaluation of the proposed active perception method in our experiments. Fig. 6.2 shows an example of interaction data of the IEMOCAP dataset. One actor, who has the attached motion capture markers, (left side) interacts with another actor. The IEMOCAP database contains audiovisual data from ten actors who performed scripted and improvised emotional scenarios in face-to-face interactions. The facial expressions and hand movements were recorded with a motion-capture system. The subjects mounted 53 markers on their faces and six markers on their hands. In addition, their conversations were recorded using video cameras.

All the recorded data were evaluated using categorical labels. Three evaluators annotated an emotion label for each utterance of the interactions. We selected the majority voted emotional category as the ground truth of each utterance if the annotations did not agree among the evaluators. For instance, when two evaluators voted the sad category whereas one evaluator voted the neutral for a focused utterance, we set the emotional category as sadness for the dataset. If some categories had the same number of votes (one vote each), we set the category of the data as an "ambiguous state". The set of emotional labels contains happiness, excitement, surprise, neutral, frustration, anger, sadness, fear, disgust, and ambiguous state. We selected eight actors' data as the training dataset and two actors' data as the test dataset. The total number of training dataset was 3,993 (happiness: 247, excitement: 408, surprise: 26, neutral: 467, frustration: 792, anger: 525, sadness: 513, fear: 18,

Figure 6.2: Example interaction data of IEMOCAP dataset

disgust: 1, ambiguous state: 996) and that of the test dataset was 992 (happiness: 50, excitement: 141, surprise: 5, neutral: 139, frustration: 206, anger: 96, sadness: 140, fear: 2, disgust: 0, ambiguous state: 213). In the experiments, our model was trained with the training data in an unsupervised manner (i.e., the model did not use the emotional labels for learning). We use the emotional labels to represent our experimental result only.

### 6.4.2 Feature Extraction from Audiovisual Signals

We considered three modality signals (i.e., $|M| = 3$): facial expressions, hand movements, and vocalization, and calculated the basic audio{visual features from the IEMOCAP database for the experiments in two steps. First, we extracted modality dependent features. The features of facial expressions included nine patterns of distance between markers and their differences from the prior time step. Each distance is represented in two-dimensional space (i.e., x and z components of the distance), which represents the surface of a full face. The hand movement features are composed of the velocity of four markers and their differences to the prior time step. Each velocity is expressed in three dimensions. The audio features include pitch, intensity, 13-dimensional MFB, and their differences from the prior time step.

Second, we calculated statistical features from all modality dependent features

for each utterance. The statistics included the mean, variance, range, maximum, and minimum of modality dependent features. All the features were normalized by Z-score. Ultimately, the number of facial expression features (or dimensions), hand movement features, and audio features obtained were 180, 120, and 150, respectively.

### 6.4.3 Network Structure

Each modality-specific network has three layers (i.e., visible, first-hidden, and second-hidden layers) and one association RBM connected to the hidden layers of all the modality-specific (stacked) networks (Figure 5.1). We set the number of visible units to 180 for the facial expression network, 120 for the hand movement network, and 150 for the audio network, respectively. The number of first and second layer hidden units for all networks were both set to 100. The number of visible and hidden nodes of the association RBM were 300 and 50, respectively.

## 6.5 Results

We demonstrate our experimental results in this section. The proposed model was trained by using the training dataset discussed in Section 6.4.1. Each modality-specific network was trained separately, and their outputs were concatenated as inputs to the association RBM, which was then trained. Note that we did not use any labeled data for the model training (i.e., unsupervised learning).

After training, we selected one modality signal of test data as the initial modality signal $m_{init}$ for the model to evaluate our proposed method. All types of modalities were selected as the initial input and examined in the experiments. We set $T = 1$ and $S = 100$ in the experiments. First, we demonstrate the experimental results of our active perception method and the method of Taniguchi et al. [111] in emotional interactions, and then provide a detailed analysis of the results based on the type of initial modality.

### 6.5.1 Evaluation of the Proposed Active Perception Method

In the experiment, our model executed the active perception based on Algorithm 1. All modality signals of the test dataset were used as initial modality signals $m_{init}$ (i.e., 2,976 data). We evaluate how close the hidden activation after the active perception $h[m_{init} + m_n]$ comes to the activation generated by using a complete set of modality signals $h[M]$. Kulback-Leibler (KL) divergence was employed to evaluate the closeness.

$$KL(p(h[m_{init} + m_n]); p(h[M])) = \sum p(h[m_{init} + m_n]) \log \frac{p(h[m_{init} + m_n])}{p(h[M])}; \tag{6.4}$$

In Eq. (6.4), $p(h[m_{init} + m_n])$ and $p(h[M])$ are the probability of hidden activation of $h[m_{init} + m_n]$ and $h[M]$, respectively. The smaller the KL divergence is, the higher the performance result for the evaluation.

Fig. 6.3 shows the experimental results. Here, "EM.max", "EM.min", "Random", and "IG.max" show the results for different active perception types while "Initial" indicates the result before the active perception (as a baseline). Under the EM.max strategy, the model selected $n = \text{argmin}\, E[E_n]$ according to Algorithm 1 (i.e., the proposed method). Under the EM.min strategy, the model set $n = \text{argmax}\, E[E_n]$. Under the Random strategy, the model selected $m_n$ from $M \backslash M_r$ at random. IG.max indicates the result of the active perception based on the method proposed by Taniguchi et al. [111].

The KL divergence of the baseline was largest in all activations, namely, the estimated states were not accurate from only the single modality. In contrast, the EM.max, the EM.min, the Random, and the IG.max strategies exhibited less divergence than the initial activations. This result shows that the second modality signals can help to estimate another's internal state. The divergence of the EM.max strategy is the smallest for all strategies. A one-way repeated-measures analysis of variance was conducted for all results. The different strategies exhibited significant differences: $F(2; 8925) = 826{:}32; p < 0{:}001$. The experimental results demonstrate the superiority of the proposed active perception method. We will present the detailed analysis of this result for all strategies in Sec. 6.5.2 and discuss the relationship between our

proposed method and the method of Taniguchi et al. [111].

We performed a PCA for the 50-dimensional output of the network to visualize the energy field in a 2D space such as Fig. 6.1. Fig. 6.4 shows the first and second PCs space of the hidden activations of the association RBM. Each marker indicates the hidden activation from the training data, and the color and shape of the markers correspond to emotional categories of the dataset. The background color presents the energy values of corresponding activations. The region where the data was crowded (i.e., PC1 is greater than zero, and PC2 is between minus one and one) indicates a low energy distribution. The energy distribution is not smooth in this PC space because many dimensional features are compressed by the analysis method; however, the energy distribution in the original space may be smooth. We illustrate two examples of our experimental results. Fig. 6.5 illustrates the 3D space, which is composed of the two-dimensional PC space in Fig. 6.4, and a vertical axis indicating the energy value of the data. This figure corresponds to the three-dimensional view of Fig. 6.1. The shapes of the markers indicate the hidden activations of the association RBM under each strategy, and color indicates each group of the experimental results. Both results demonstrate that the proposed method (red arrows in Fig. 6.5) selects the modality that minimizes the actual network energy to a greater extent than another modality. The hidden activations obtained by the proposed method are closer together than those obtained by the EM.min strategy in the PC space.

### 6.5.2 Detailed Analysis of Active Perception by Types of Initial Modality

We analyzed the experimental results based on the types of initial modality signals. Fig. 6.6 shows the KL divergence from ground truth activations under each initial modality condition. The graph labels: Hand only, Face only, and Audio only represent the result when the model selects only the corresponding modality. Namely, under the Hand only strategy, the model always selected hand signals as the second modality signals regardless of the change in network energy.

The EM.max strategy (i.e., proposed method) under all initial conditions performed more accurate estimations than the EM.min strategy. These results suggest

that our active perception method works well regardless of initial modality signals. However, the KL divergence of the EM.max strategy under the Audio initial condition is larger than for other conditions, and the Hand only strategy shows higher performance than the EM.max strategy. By analyzing the results based on the modality types, it was revealed that the KL divergence between initial hidden activation and the ground truth was not even, and the Hand initial condition represented the smallest value in all conditions. These results indicate that each modality contains a different amount of information (i.e., different entropy) for representing hidden activations. Fig. 6.7 shows the selection rates of the modality in the EM.max strategy under each condition, and each number indicates the rate of the most selected modality. Each experimental result tended to select the modality whose KL divergence is lower than other modalities in the initial setting as the next attention modality. Especially, the Face initial condition always selected the hand modality signals (i.e., same as the Hand only strategy). The experimental results suggest that the hand modality signals contain more information for the estimation of others' internal states than other modalities in our experimental setting and that the proposed method makes a good selection of highly informative signals (i.e., the signal has large information gain) for active perception. The IG.max strategy [111] similarly selected highly informative signals; however, under the Audio initial condition, the method has lower performance than EM.max. We suppose that each method has different limitations based on the characteristics of the task.

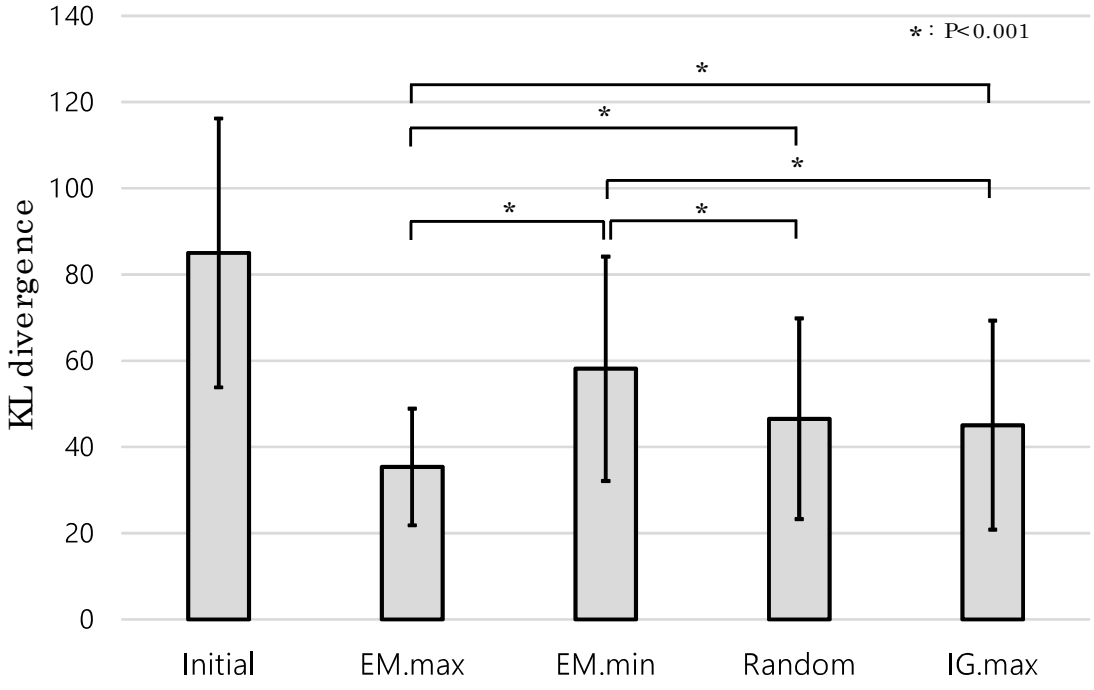



Figure 6.3: KL divergence from the ground truth

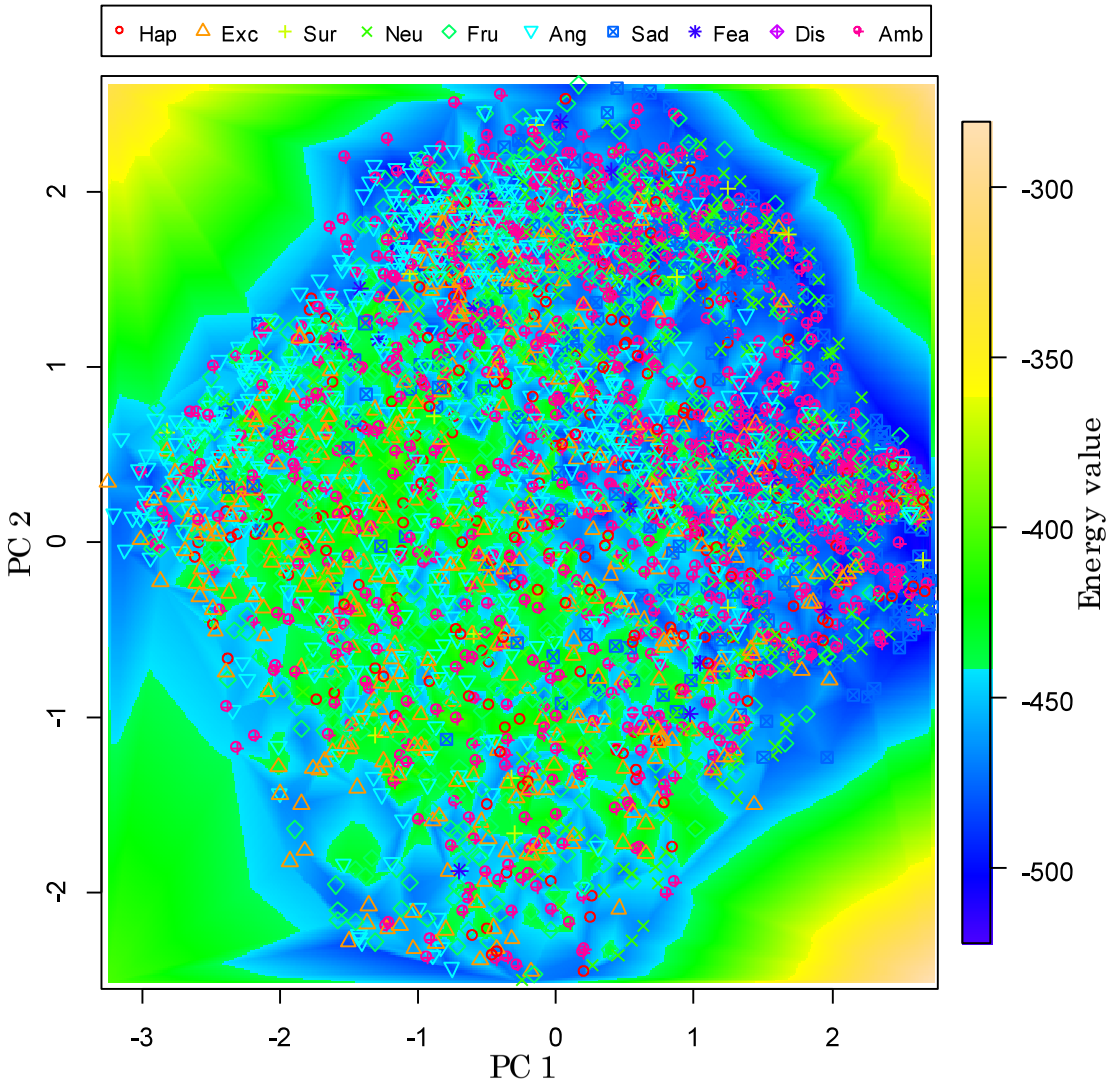



Figure 6.4: Energy field and the data distribution in the first and second PC space

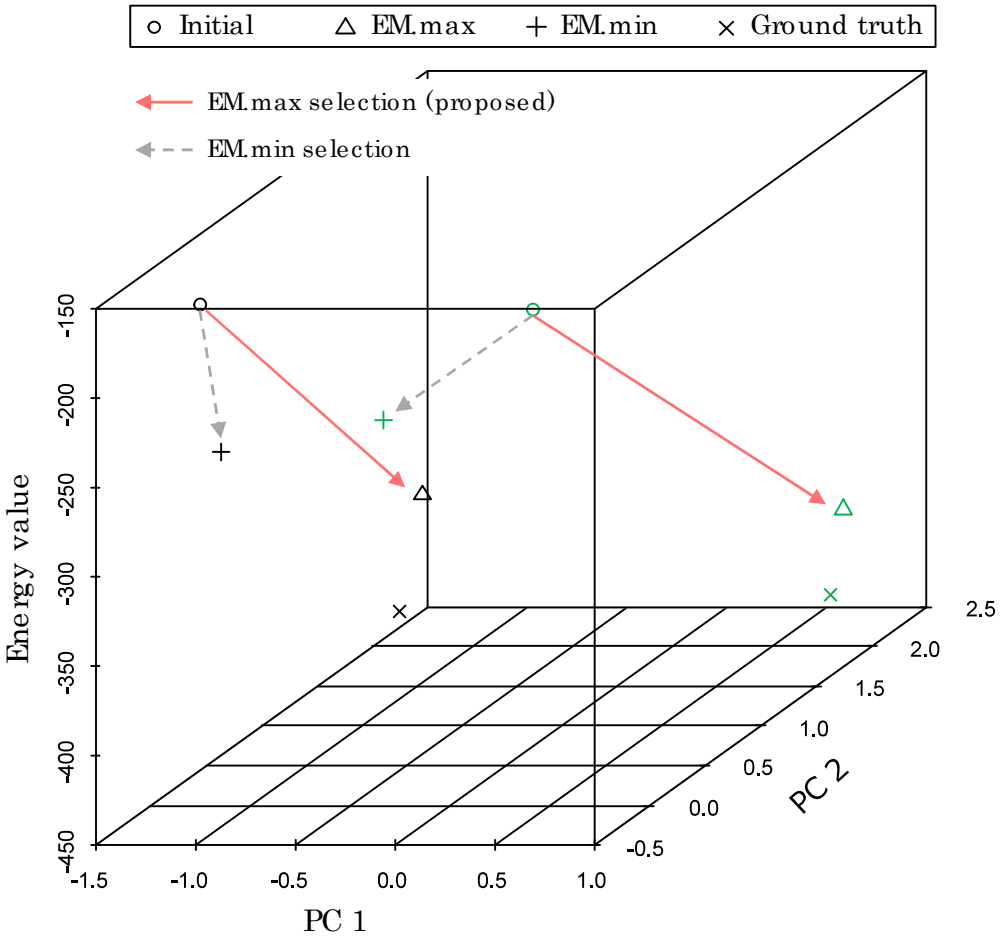



Figure 6.5: Example transitions of the hidden activations in the PC space

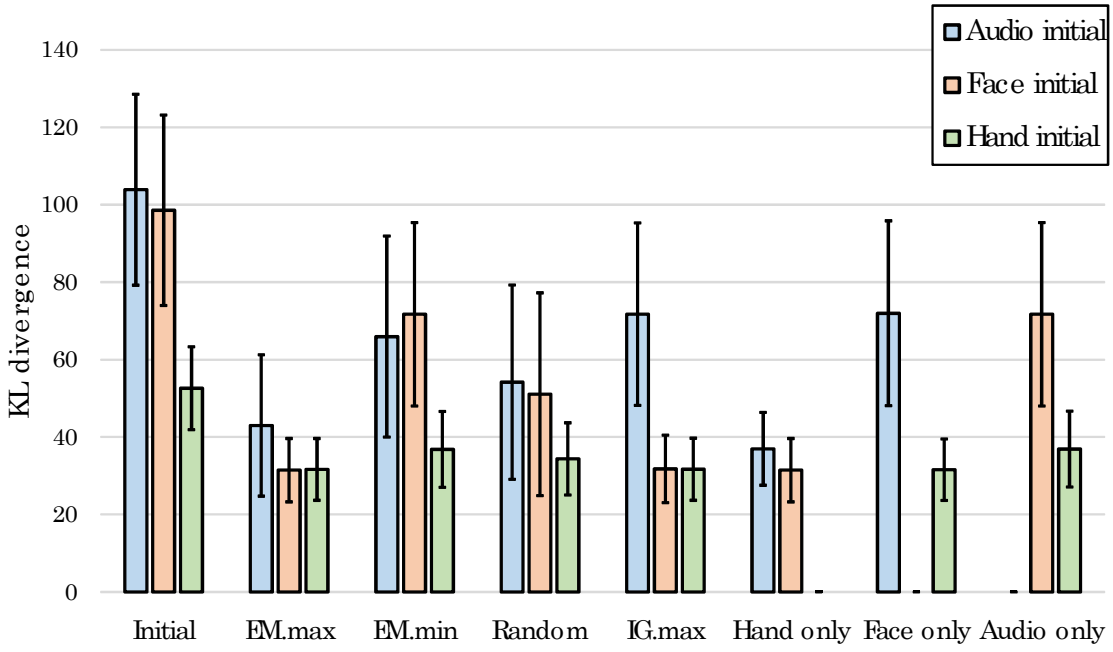



Figure 6.6: KL divergence from the ground truth under each initial condition

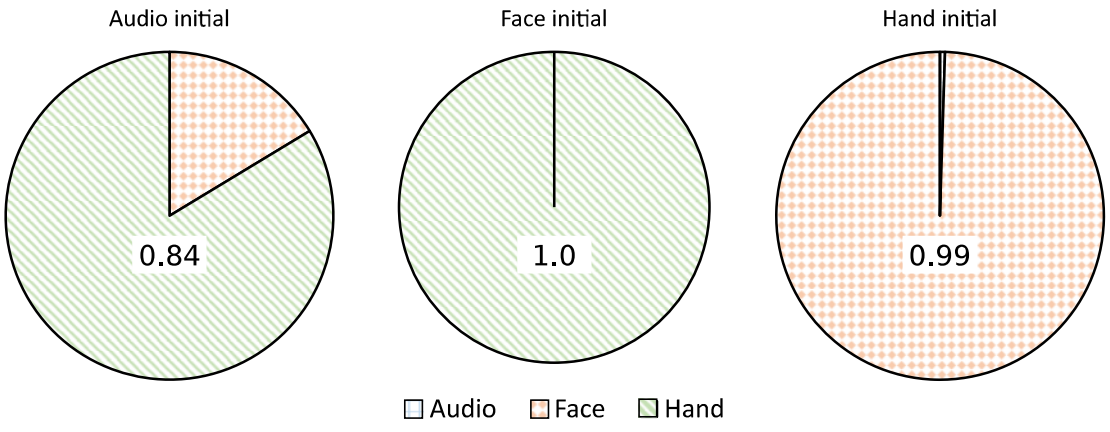



Figure 6.7: Rate of selected modality under each initial condition

## 6.6 Discussion and Future Work

Our model was trained using three different modality signals of emotional expressions in an unsupervised manner, so there is no guarantee that the hidden activations represent emotional categories clearly. Fig. 6.4 shows hidden activations of all trained data labeled with emotional categories. There are no clear emotional clusters in the space; however, the low-dimensional feature, which corresponds to the intensity or activity of the multimodal expressions, is represented from the lower left corner to the upper right corner based on the emotional categories. In the audiovisual features, which were extracted from the IEMOCAP dataset, the hand signal's features represent the highest correspondence to emotion expressions among all modalities because the signals vary extremely in the interactions, as the features include the velocity and the acceleration characteristics of the hand signals. As a result, the hidden activations under the Hand initial condition showed the lowest KL divergence in all modality conditions in the experiments, as depicted in Fig. 6.6.

The previous studies for estimating object categories [110, 111] used mutual information between the current estimation and unobserved modality signals as a criterion for active perception. Their methods select the next modality whose expectation of the mutual information is highest among all modalities. This selection method corresponds to the minimization of the uncertainty of the category estimation because the mutual information represents the amount of information between two variables. In contrast, our active perception method uses the energy value of the network model and selects the next modality whose expectation energy is lowest among all modalities. This corresponds to the maximization of the amount of information between modality signals and hidden activations for active perception because lower energy indicates a higher likelihood of the data. We suppose that there are complementary relationships between our method and previous methods [110, 111]. Basically, our proposed method and the previous method (i.e., the EM.max and the IG.max strategies) showed similarly good performance under the Face initial condition and the Hand initial condition. However, the KL divergence under the Audio initial condition was worse than our result. To assess this supposition, we will compare our active perception method with the mutual information maximization criterion by

mathematically formulating them.

The proposed active perception method may be able to use other energy-based models such as RBMs containing softmax nodes [112], the reinforcement learning model [113], or the energy-based generative model [114]. We will attempt to apply our method to these models to evaluate more complex scenarios, for example, the human emotion changes over time and/ or by the robot's actions during interactions.

For future issues, we will formulate the reason(s) why our method showed better performance than other methods and compare the results with different criteria such as mutual information maximization proposed in [110, 111]. Furthermore, we will apply the proposed method in actual human{ robot interactions.

# Chapter 7

# Conclusion

The work presented in this dissertation addressed how to reproduce emotion development and four essential capabilities for emotional interaction. For these aims, we proposed artificial emotion models based on the idea of predictive coding and evaluated the models in our intended interaction situations.

We feel emotions in everyday life. Humans recognize their own physical and mental conditions as emotions and express them in communication with others. Although how humans acquire various types of emotions is still a topic of debate. Several observational and analytical studies in psychology, neuroscience, and cognitive science have given evidence that human emotion and its functions are acquired in the developmental process during infancy. However, the developmental mechanisms and factors remain unclear. On the other hand, emotion and its functions have attracted not only psychological studies but also engineering studies. To elucidate those mechanisms and realize the functions of emotion in communication play essential roles in building empathic robots. Actually, several robots focused on the mechanisms of human emotion have been developed for interaction with humans. Regarding these approaches, cognitive developmental robotics and affective developmental robotics have been proposed, which relate emotion, by involving findings in psychology, neuroscience, cognitive science, and robotics.

This chapter summarizes our approaches for modeling the development of emotion perception and essential capabilities for emotional interaction and describes the knowledge acquired through the experiments. Subsequently, research issues that

should be solved in the future toward a deeper understanding of emotion and realizing emotional communication between humans and robots are discussed.

## 7.1 Summary of Our Approaches

In this dissertation, we proposed a general artificial emotion model based on the idea of predictive coding by employing a RBM as a key component. The proposed model has been used for the study of emotion development in infant–caregiver interactions in Chapter 4. A modified version of the model has been also used for solving issues in emotion human–robot interaction in Chapters 5 and 6. The following sections describe the detail of the artificial emotion model and summarize each experimental result and the obtained findings.

### 7.1.1 Modeling Emotion Development by Predictive Coding with Findings of Neuroscience and Developmental Psychology

In Chapter 4, we have presented the model of development for emotion perception based on the idea of predictive coding with our two hypotheses from neuroscience and developmental psychology findings: tactile dominance and perceptual improvement. The proposed model was exposed to multimodal emotional expressions of humans who imitate caregivers through the virtual infant–caregiver interaction. We evaluated how tactile dominance and perceptual improvement contribute to the development of emotion perception, and their roles were examined by comparing among four conditions (with and without each mechanism in the hypotheses).

From the experimental results, it was suggested that tactile dominance and perceptual development facilitated the development of emotion perception only when both mechanisms were integrated into the model. We conclude that the ability of the human skin to perceive the emotional valence of interaction stimuli and improvement in perceptual acuity play essential roles in the development of emotion perception. Connecting these findings to psychological and neuroscience studies, we expect to elucidate on the neural and social mechanisms of the development of emotion.

### 7.1.2 Reproducing Essential Ability in emotional interaction by Predictive Coding

In Chapter 5 and 6, we have presented a unified model that can express emotional signals, estimate others' emotions, and execute active perception in emotional interaction between humans and robots. The model was constructed of RBMs with some modifications to the model of Chapter 4. However, the functions of tactile dominance and perceptual improvement were not considered.

In Chapter 5, we focused on the first of two abilities: emotion expression and emotion estimation. Our model was expected to rectify two issues noted in previous emotional models by employing the idea of predictive coding. The first is acquiring the emotional representation of multimodal expressions for emotion generation and estimation. The second is the update of estimation belief of the partner's emotion from ambiguous signals via mental simulation. In Chapter 6, we focused on the last ability in emotional interaction, active perception. The key idea of execution of active perception involves minimizing the network energy of the proposed model based on the idea of the free energy principle [8, 9], because energy values represent the likelihood of the corresponding data. So the criterion is to select the most plausible signals based on the current estimation. Those abilities were evaluated through assumed face-to-face human{robot emotional interaction using the IEMOCAP dataset.

From the experimental results, the followings were confirmed.

- The model can obtain low-dimensional emotional representation of multimodal signals through interaction with humans and generate its own multimodal expressions based on the perceived emotional representation.

- The model can update the estimated belief in the interaction partner's emotional state from his/ her ambiguous expressions based on the mental simulation mechanism.

- The model can execute active perception to estimate a partner's emotion by selecting the next attention for his/ her multimodal expressions based on the mental simulation using energy minimization.

The first result verifies that the proposed model based on the idea of predictive coding by utilizing an RBM network can represent the relationship between humans' and robots' multimodal signals and emotional states as latent signals. The second and the third results show that the prediction of unobserved signals from observed ones in the model can be used for reproducing essential abilities in emotional interaction with humans. We concluded that the characteristics of the model based on the idea of predictive coding enable the robot to improve emotional interactions in human–robot communication.

## 7.2 Toward Deeper Understanding

The specific future issues applicable to each study were described in each chapter. This section discusses extensions of our approaches for better understanding human emotion and constructing artificial emotion for robots.

### 7.2.1 Other Generative Models for Time and Spatial Extension

In this dissertation, we employed the RBM as a key mechanism that captures the idea of predictive coding in our artificial emotion model. The RBM has the potential for applying the free energy principle proposed Friston [8, 9] that has attracted attention in neuroscience, because the learning algorithm of the RBM is related to free energy minimization. On the other hand, the RBM, especially the Bernoulli–Bernoulli RBM described in Chapter 3, can handle limited information that represents only binary states. Furthermore, the RBM cannot operate time-series information.

Various generative models have been proposed in the field of machine learning. For example, the hidden Markov model [115] is a popular model used with time-series data, and a Gaussian process hidden semi-Markov model [116] has been proposed as an extension of the hidden Markov model to represent high-dimensional complex time-series data. Multimodal latent Dirichlet allocation [117] and the multimodal hidden Markov model [118] can handle multiple signals for representing concepts from various modality signals such as object and motion categories. Moreover, some energy-based

generative models, which can represent the uncertainty of information (similar to the RBM), have also been proposed [114, 119]. It is important to adopt the above models for flexibility of representation of spatiotemporal information applied to robots in a real environment and to simulate the more complex information processing conducted in the human brain.

### 7.2.2 Considering Interoceptive and Embodiment Signals

We can measure the emotional states of humans as body and internal organ reactions such as heart rate, sweating, and changes in hormones as these reactions are known to be related to emotional perception in humans. Seth [14{16] explains emotion based on the free energy principle. His idea is that emotion is acquired as a result of prediction for exteroceptive, proprioception, and interoceptive signals (i.e., embodiment signals). In other words, the prediction for ourselves and its error are perceived as a specific state, emotion.

On the other hand, many engineering approaches to construct emotion models have neglected considering the embodiment signals. Few studies have employed a robots' battery level, the temperature of its motor(s), or signals of artificial hormonal systems as embodiment signals for modeling emotion. Internal organs are not necessary functions for robots. However, these functions can represent different dynamics with external stimuli. We believe that such dynamics can be an essential key idea for representing different time scales of "emotion" such as affection, feeling, mood, etc. Additionally, considering similar embodiment to humans is also important to truly understand emotion and emotional communication. As mentioned in the above section, adopting generative models that can process spatiotemporal information may enable us to challenge these issues.

### 7.2.3 Action, Evaluation, and Artificial Empathy

The word "emotion" contains "motion". It is important to consider the motions and actions of robots in emotion studies. In recent years, emotional studies in cognitive science have considered that emotion is the evaluation mechanism of stimuli based on one's own actions. This mechanism allows humans to compare experiences of self and

others in a common dimension via emotion. It plays important roles in estimating and sharing emotions with others and leads to empathy. In Chapter 6, the model executed active perception as the robot's action; however, it is necessary to take actions more deeply into account in terms of effects on others and on the environment. We expect that we can realize artificial empathy applied as affective developmental robotics by focusing on physical communication with humans and on the issues stated so far.

# Bibliography


[1] K.M.B. Bridges. A genetic theory of the emotions. The Pedagogical Seminary and Journal of Genetic Psychology, Vol. 37, pp. 514{527, 1930.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, Vol. 42, No. 4, pp. 335{359, 2008.

[3] Joseph LeDoux. The emotional brain, fear, and the amygdala. Cellular and molecular neurobiology, Vol. 23, No. 4-5, pp. 727{738, 2003.

[4] William James. What is an emotion? Mind, Vol. 9, No. 34, pp. 188{205, 1884.

[5] A.R. Damasio. Descartes' error: Emotion, rationality and the human brain. New York: Putnam, p. 352, 1994.

[6] Keith Oatley and PN Johnson-Laird. Cognitive approaches to emotions. Trends in cognitive sciences, Vol. 18, No. 3, pp. 134{140, 2014.

[7] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, Vol. 2, No. 1, p. 79, 1999.

[8] Karl Friston, James Kilner, and Lee Harrison. A free energy principle for the brain. Journal of Physiology-Paris, Vol. 100, No. 1-3, pp. 70{87, 2006.

[9] Karl Friston. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, Vol. 11, No. 2, pp. 127{138, 2010.

[10] Giovanni Pezzulo, Francesco Rigoli, and Karl Friston. Active inference, homeostatic regulation and adaptive behavioural control. Progress in neurobiology, Vol. 134, pp. 17{35, 2015.

[11] Yukie Nagai and Minoru Asada. Predictive learning of sensorimotor information as a key for cognitive development. In Proc. of the IROS 2015 Workshop on Sensorimotor Contingencies for Robotics, 2015.

[12] Jun Tani. Exploring robotic minds: actions, symbols, and consciousness as self-organizing dynamic phenomena. Oxford University Press, 2016.

[13] Jun Tani. On the interactions between top-down anticipation and bottom-up regression. Frontiers in Neurorobotics, Vol. 1, p. 2, 2007.

[14] Anil K Seth, Keisuke Suzuki, and Hugo D Critchley. An interoceptive predictive coding model of conscious presence. Frontiers in psychology, Vol. 2, p. 395, 2012.

[15] Anil K Seth. Interoceptive inference, emotion, and the embodied self. Trends in cognitive sciences, Vol. 17, No. 11, pp. 565{573, 2013.

[16] Anil K Seth and Karl J Friston. Active interoceptive inference and the emotional brain. Phil. Trans. R. Soc. B, Vol. 371, No. 1708, p. 20160007, 2016.

[17] S.S. Tomkins. Affect Imagery Consciousness: Volume I: The Positive Affects. Springer publishing company, 1962.

[18] C.E. Izard. The psychology of emotions. Springer, 1991.

[19] L.A. Sroufe. The coherence of individual development: Early care, attachment, and subsequent developmental issues. American Psychologist, Vol. 34, No. 10, pp. 834{841, 1979.

[20] M. Lewis. The self in self-conscious emotions. Annals of the New York Academy of Sciences, Vol. 818, No. 1, pp. 119{142, 1997.

[21] A.D. Craig. How do you feel? interoception: the sense of the physiological condition of the body. Nature Reviews Neuroscience, Vol. 3, No. 8, pp. 655{666, 2002.

[22] Michael S Beauchamp, Nafi E Yasar, Richard E Frye, and Tony Ro. Touch, sound and vision in human superior temporal sulcus. Neuroimage, Vol. 41, No. 3, pp. 1011{1020, 2008.

[23] Salvatore Campanella and Pascal Belin. Integrating face and voice in person perception. Trends in cognitive sciences, Vol. 11, No. 12, pp. 535{543, 2007.

[24] Rebecca Watson, Marianne Latinus, Takao Noguchi, Oliver Garrod, Frances Crabbe, and Pascal Belin. Crossmodal adaptation in right posterior superior temporal sulcus during face{voice emotional integration. The Journal of Neuroscience, Vol. 34, No. 20, pp. 6813{6821, 2014.

[25] Benjamin Kreifelts, Thomas Ethofer, Wolfgang Grodd, Michael Erb, and Dirk Wildgruber. Audiovisual integration of emotional signals in voice and face: an event-related fmri study. Neuroimage, Vol. 37, No. 4, pp. 1445{1456, 2007.

[26] Tobias Grossmann, Tricia Striano, and Angela D Friederici. Crossmodal integration of emotional information from face and voice in the infant brain. Developmental Science, Vol. 9, No. 3, pp. 309{315, 2006.

[27] Vittorio Gallese, Luciano Fadiga, Leonardo Fogassi, and Giacomo Rizzolatti. Action recognition in the premotor cortex. Brain, Vol. 119, No. 2, pp. 593{609, 1996.

[28] Giacomo Rizzolatti. The mirror neuron system and its function in humans. Anatomy and embryology, Vol. 210, No. 5-6, pp. 419{421, 2005.

[29] Giacomo Rizzolatti and Corrado Sinigaglia. Mirrors in the brain: How our minds share actions and emotions. Oxford University Press, USA, 2008.

[30] Giuseppe Di Pellegrino, Luciano Fadiga, Leonardo Fogassi, Vittorio Gallese, and Giacomo Rizzolatti. Understanding motor events: a neurophysiological study. Experimental brain research, Vol. 91, No. 1, pp. 176{180, 1992.

[31] Marco Iacoboni. Imitation, empathy, and mirror neurons. Annual review of psychology, Vol. 60, pp. 653{670, 2009.

[32] Vittorio Gallese and Alvin Goldman. Mirror neurons and the simulation theory of mind-reading. Trends in cognitive sciences, Vol. 2, No. 12, pp. 493{501, 1998.

[33] Frank Van Overwalle and Kris Baetens. Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. Neuroimage, Vol. 48, No. 3, pp. 564{584, 2009.

[34] Andra Adams and Peter Robinson. An android head for social-emotional intervention for children with autism spectrum conditions. In Affective Computing and Intelligent Interaction, pp. 183{190. Springer, 2011.

[35] Yuki Matsui, Masayoshi Kanoh, Shohei Kato, Tsuyoshi Nakamura, and Hidenori Itoh. A model for generating facial expressions using virtual emotion based on simple recurrent network. Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 14, No. 5, pp. 453{463, 2010.

[36] Masayoshi Kanoh, Shohei Kato, and Hidenori Itoh. Facial expressions using emotional space in sensitivity communication robot" ifbot". In IEEE/RSJ International Conference on Intelligent Robots and Systems, Vol. 2, pp. 1586{1591. IEEE, 2004.

[37] Masayoshi Kanoh, Susumu Iwata, Shohei Kato, and Hidenori Itoh. Emotive facial expressions of sensitivity communication robot "ifbot". Kansei Engineering International, Vol. 5, No. 3, pp. 35{42, 2005.

[38] Gabriele Trovato, Massimiliano Zecca, Tatsuhiro Kishi, Nobutsuna Endo, Kenji Hashimoto, and Atsuo Takanishi. Generation of humanoid robot's facial expressions for context-aware communication. International Journal of Humanoid Robotics, Vol. 10, No. 01, p. 1350013, 2013.

[39] Tatsuhiro Kishi, Takuya Kojima, Nobutsuna Endo, Matthieu Destephe, Takuya Otani, Lorenzo Jamone, Przemyslaw Kryczka, Gabriele Trovato, Kenji Hashimoto, Sarah Cosentino, et al. Impression survey of the emotion expression humanoid robot with mental model based dynamic emotions. In IEEE International Conference on Robotics and Automation, pp. 1663{1668. IEEE, 2013.

[40] Cynthia Breazeal, Daphna Buchsbaum, Jesse Gray, David Gatenby, and Bruce Blumberg. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. Artificial life, Vol. 11, No. 1-2, pp. 31{62, 2005.

[41] Cynthia Breazeal and Lijin Aryananda. Recognition of affective communicative intent in robot-directed speech. Autonomous robots, Vol. 12, No. 1, pp. 83{104, 2002.

[42] Cynthia Breazeal. Emotion and sociable humanoid robots. International Journal of Human-Computer Studies, Vol. 59, No. 1, pp. 119{155, 2003.

[43] Frank Hegel, Torsten Spexard, Britta Wrede, Gernot Horstmann, and Thurid Vogt. Playing a different imitation game: Interaction with an empathic android robot. In IEEE-RAS International Conference on Humanoid Robots, pp. 56{61. IEEE, 2006.

[44] Angelica Lim and Hiroshi G Okuno. The mei robot: towards using motherese to develop multimodal emotional intelligence. IEEE Transactions on Autonomous Mental Development, Vol. 6, No. 2, pp. 126{138, 2014.

[45] Angelica Lim and Hiroshi G Okuno. A recipe for empathy. International Journal of Social Robotics, Vol. 7, No. 1, pp. 35{49, 2015.

[46] Ingo Lütkebohle, Frank Hegel, Simon Schulz, Matthias Hackel, Britta Wrede, Sven Wachsmuth, and Gerhard Sagerer. The bielefeld anthropomorphic robot head" flobi". In IEEE International Conference on Robotics and Automation, Vol. 3, pp. 3384{3391, 2010.

[47] Iolanda Leite, Carlos Martinho, and Ana Paiva. Social robots for long-term interaction: a survey. International Journal of Social Robotics, Vol. 5, No. 2, pp. 291{308, 2013.

[48] Minoru Asada. Towards artificial empathy. International Journal of Social Robotics, Vol. 7, No. 1, pp. 19{33, 2015.

[49] Minoru Asada. Development of artificial empathy. Neuroscience research, Vol. 90, pp. 41{50, 2015.

[50] A. Watanabe, M. Ogino, and M. Asada. Mapping facial expression to internal states based on intuitive parenting. Journal of Robotics and Mechatronics, Vol. 19, No. 3, p. 315, 2007.

[51] Arlene S Walker-Andrews. Infants' perception of expressive behaviors: differentiation of multimodal information. Psychological bulletin, Vol. 121, No. 3, p. 437, 1997.

[52] Jacqueline Nadel and Darwin Muir. Emotional development: Recent research advances. Oxford University Press, USA, 2005.

[53] Michael Lewis, Jeannette M Haviland-Jones, and Lisa Feldman Barrett. Handbook of emotions. Guilford Press, 2008.

[54] Tobias Grossmann. The development of emotion perception in face and voice during infancy. Restorative neurology and neuroscience, Vol. 28, No. 2, pp. 219{236, 2010.

[55] P. Ekman. An argument for basic emotions. Cognition & Emotion, Vol. 6, No. 3-4, pp. 169{200, 1992.

[56] Vivien Ainley, Matthew AJ Apps, Aikaterini Fotopoulou, and Manos Tsakiris. 'bodily precision' : a predictive coding account of individual differences in interoceptive accuracy. Phil. Trans. R. Soc. B, Vol. 371, No. 1708, p. 20160003, 2016.

[57] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, Vol. 3, No. Jan, pp. 993{1022, 2003.

[58] G. E. Hinton. A practical guide to training restricted boltzmann machines. Technical report, Department of Computer Science University of Toronto, 2010.

[59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, pp. 2672{2680, 2014.

[60] G. E. Hinton and R Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, Vol. 313, No. 5786, pp. 504{507, 2006.

[61] J.A. Russell. A circumplex model of affect. Journal of personality and social psychology, Vol. 39, No. 6, p. 1161, 1980.

[62] J.A. Russell. Core affect and the psychological construction of emotion. Psychological review, Vol. 110, No. 1, p. 145, 2003.

[63] James A Russell. Emotion, core affect, and psychological construction. Cognition and emotion, Vol. 23, No. 7, pp. 1259{1283, 2009.

[64] Nico H Frijda, Peter Kuipers, and Elisabeth Ter Schure. Relations among emotion, appraisal, and emotional action readiness. Journal of personality and social psychology, Vol. 57, No. 2, p. 212, 1989.

[65] Martha Peláez-Nogueras, Jacob L. Gewirtz, Tiffany Field, Maricel Cigales, Julie Malphurs, Sara Clasky, and Aida Sanchez. Infants' preference for touch stimulation in face-to-face interactions. Journal of Applied Developmental Psychology, Vol. 17, pp. 199{213, 1996.

[66] Martha Peláez-Nogueras, Tiffany Field, Ziarat Hossain, and Jeffrey Pickens. Depressed mother's touching increases infants' positive affect and attention in still-face interactions. Child Development, Vol. 67, pp. 1780{1792, 1996.

[67] Matthew J. Hertenstein. Touch: Its communicative functions in infancy. Human Development, Vol. 45, pp. 79{94, 2002.

[68] Minoru Asada, Koh Hosoda, Yasuo Kuniyoshi, Hiroshi Ishiguro, Toshio Inui, Yuichiro Yoshikawa, Masaki Ogino, and Chisato Yoshida. Cognitive developmental robotics: a survey. Autonomous Mental Development, IEEE Transactions on, Vol. 1, No. 1, pp. 12{34, 2009.

[69] Tetsuya Ogata and Shigeki Sugano. Emotional communication between humans and the autonomous robot wamoeba-2 (waseda amoeba) which has the emotion model. JSME International Journal Series C Mechanical Systems, Machine Elements and Manufacturing, Vol. 43, No. 3, pp. 568{574, 2000.

[70] A Blanchard and Lola Canamero. From imprinting to adaptation: Building a history of affective interaction. In Procs 5th Int Workshop on Epigenetic Robotics. Lund University, 2005.

[71] Cyril Hasson, Philippe Gaussier, and Sofiane Boucenna. Emotions as a dynamical system: the interplay between the meta-control and communication function of emotions. Paladyn, Vol. 2, No. 3, pp. 111{125, 2011.

[72] Antoine Hiolle, Matthew Lewis, and Lola Canamero. Arousal regulation and affective adaptation to human responsiveness by a robot that explores and learns a novel environment. Frontiers in neurorobotics, Vol. 8, , 2014.

[73] John Lones, Matthew Lewis, and Lola Canamero. From sensorimotor experiences to cognitive development: investigating the influence of experiential diversity on the development of an epigenetic robot. Frontiers in Robotics and AI, Vol. 3, p. 44, 2016.

[74] Paul Ekman. Facial action coding system (facs). A human face, 2002.

[75] Laurel D Riek, Philip C Paul, and Peter Robinson. When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry. Journal on Multimodal User Interfaces, Vol. 3, No. 1-2, pp. 99{108, 2010.

[76] Laurel D Riek and Peter Robinson. Real-time empathy: Facial mimicry on a robot. In Workshop on Affective Interaction in Natural Environments (AFFINE) at the International ACM Conference on Multimodal Interfaces (ICMI 08). ACM. Citeseer, 2008.

[77] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions.

IEEE transactions on pattern analysis and machine intelligence, Vol. 31, No. 1, pp. 39{58, 2009.

[78] Yongjin Wang and Ling Guan. Recognizing human emotional state from audio-visual signals. IEEE Transactions on Multimedia, Vol. 10, No. 5, pp. 936{946, 2008.

[79] Emily Mower, Maja J Mataric, and Shrikanth Narayanan. A framework for automatic human emotion classification using emotion profiles. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 5, pp. 1057{1070, 2011.

[80] Yelin Kim, Honglak Lee, and Emily Mower Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3687{3691. IEEE, 2013.

[81] Sainbayar Sukhbaatar, Takaki Makino, Kazuyuki Aihara, and Takashi Chikayama. Robust generation of dynamical patterns in human motion by a deep belief nets. In Asian Conference on Machine Learning, pp. 231{246, 2011.

[82] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning, pp. 689{696, 2011.

[83] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In International Conference on Machine Learning Workshop, 2012.

[84] Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. Journal of Machine Learning Research, Vol. 15, No. 1, pp. 2949{2980, 2014.

[85] KyungHyun Cho, Alexander Ilin, and Tapani Raiko. Improved learning of gaussian-bernoulli restricted boltzmann machines. In Artificial Neural Networks and Machine Learning{ICANN 2011, pp. 10{17. Springer, 2011.

[86] L Alan Sroufe. Emotional development: The organization of emotional life in the early years. Cambridge University Press, 1997.

[87] N. Danziger, K.M. Prkachin, and J.C. Willer. Is pain the price of empathy? the perception of others' pain in patients with congenital insensitivity to pain. Brain, Vol. 129, No. 9, pp. 2494{2507, 2006.

[88] Matthew J Hertenstein, Rachel Holmes, Margaret McCullough, and Dacher Keltner. The communication of emotion via touch. Emotion, Vol. 9, No. 4, p. 566, 2009.

[89] A.D.L. Jean, D.M. Stack, and A. Fogel. A longitudinal investigation of maternal touching across the first 6 months of life: Age and context effects. Infant Behavior and Development, Vol. 32, No. 3, pp. 344{349, 2009.

[90] Richard J Traub and Lorne M Mendell. The spinal projection of individual identified a-delta-and c-fibers. Journal of neurophysiology, Vol. 59, No. 1, pp. 41{55, 1988.

[91] L. Fabrizi, R. Slater, A. Worley, J. Meek, S. Boyd, S. Olhede, and M. Fitzgerald. A shift in sensory processing that enables the developing human brain to discriminate touch from pain. Current Biology, 2011.

[92] M. Björnsdotter, I. Morrison, and H. Olausson. Feeling good: on the role of c fiber mediated touch in interoception. Experimental brain research, Vol. 207, No. 3, pp. 149{155, 2010.

[93] M. Shibata, Y. Fuchino, N. Naoi, S. Kohno, M. Kawai, K. Okanoya, and M. Myowa-Yamakoshi. Broad cortical activation in response to tactile stimulation in newborns. NeuroReport, Vol. 23, No. 6, p. 373, 2012.

[94] Robert M. Bradley. Fetal sensory receptors. Physiological Reviews, Vol. 55, No. 3, pp. 352{382, July 1975.

[95] Velma Dobson and Davida Y Teller. Visual acuity in human infants: a review and comparison of behavioral and electrophysiological studies. Vision Research, Vol. 18, No. 11, pp. 1469{1483, 1978.

[96] Lynne Werner Olsho, Elizabeth G Koch, and Christopher F Halpin. Level and age effects in infant frequency discrimination. The Journal of the Acoustical Society of America, Vol. 82, p. 454, 1987.

[97] Linda B Smith and Deborah G Kemler. Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. Journal of Experimental Child Psychology, Vol. 24, No. 2, pp. 279{298, 1977.

[98] Linda B Smith. Perceptual development and category generalization. Child Development, pp. 705{715, 1979.

[99] Yukie Nagai, Minoru Asada, and Koh Hosoda. Learning for joint attention helped by functional development. Advanced Robotics, Vol. 20, No. 10, pp. 1165{1181, 2006.

[100] Roland Schmidt, M Schmelz, HE Torebjörk, and HO Handwerker. Mechano-insensitive nociceptors encode pain evoked by tonic pressure to human skin. Neuroscience, Vol. 98, No. 4, pp. 793{800, 2000.

[101] Anne Fernald. Intonation and communicative intent in mothers' speech to infants: Is the melody the message? Child development, pp. 1497{1510, 1989.

[102] Maria Spinelli, Mirco Fasolo, and Judi Mesman. Does prosody make the difference? a meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. Developmental Review, Vol. 44, pp. 1{18, 2017.

[103] Y. Tada and K. Hosoda. Acquisition of multi-modal expression of slip through pick-up experiences. Advanced Robotics, Vol. 21, No. 5-6, pp. 601{617, 2007.

[104] Ravinder S Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensing—from humans to humanoids. Robotics, IEEE Transactions on, Vol. 26, No. 1, pp. 1{20, 2010.

[105] Y. Indo. Nerve growth factor and the physiology of pain: lessons from congenital insensitivity to pain with anhidrosis. Clinical Genetics, 2012.

[106] Takato Horii, Yukie Nagai, and Minoru Asada. Toward analysis of emotional development using physiological and behavioral data. In Proceedings of the HRI 2014 Workshop on HRI: A Bridge between Robotics and Neuroscience, pp. 47{48, 2014.

[107] Lindsay M Oberman, Piotr Winkielman, and Vilayanur S Ramachandran. Face to face: Blocking facial mimicry can selectively impair recognition of emotional expressions. Social neuroscience, Vol. 2, No. 3-4, pp. 167{178, 2007.

[108] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), Vol. 2, No. 3, p. 27, 2011.

[109] Albert Mehrabian, et al. Silent messages, Vol. 8. Wadsworth Belmont, CA, 1971.

[110] Yutaka Sakaguchi. Haptic sensing system with active perception. Advanced robotics, Vol. 8, No. 3, pp. 263{283, 1993.

[111] Tadahiro Taniguchi, Toshiaki Takano, and Ryo Yoshino. Multimodal hierarchical dirichlet process-based active perception. arXiv preprint arXiv:1510.00331, 2015.

[112] Nitish Srivastava, Ruslan R Salakhutdinov, and Geoffrey E Hinton. Modeling documents with deep boltzmann machines. arXiv preprint arXiv:1309.6865, 2013.

[113] Brian Sallans and Geoffrey E Hinton. Reinforcement learning with factored states and actions. Journal of Machine Learning Research, Vol. 5, No. Aug, pp. 1063{1088, 2004.

[114] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. arXiv preprint arXiv:1606.03439, 2016.

[115] Emily B Fox, Erik B Sudderth, Michael I Jordan, and Alan S Willsky. Bayesian nonparametric methods for learning markov switching processes. IEEE Signal Processing Magazine, Vol. 27, No. 6, pp. 43{54, 2010.

[116] Tomoaki Nakamura, Takayuki Nagai, Daichi Mochihashi, Ichiro Kobayashi, Hideki Asoh, and Masahide Kaneko. Segmenting continuous motions with hidden semi-markov models and gaussian processes. Frontiers in neurorobotics, Vol. 11, p. 67, 2017.

[117] Tomoaki Nakamura, Takaya Araki, Takayuki Nagai, and Naoto Iwahashi. Grounding of word meanings in latent dirichlet allocation-based multimodal concepts. Advanced Robotics, Vol. 25, No. 17, pp. 2189{2206, 2011.

[118] Muhammad Attamimi, Yusuke Katakami, Kasumi Abe, Takayuki Nagai, and Tomoaki Nakamura. Modeling of honest signals for human robot interaction. In Human-Robot Interaction (HRI), 2016 11th ACM/ IEEE International Conference on, pp. 415{416. IEEE, 2016.

[119] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. arXiv preprint arXiv:1609.03126, 2016.

# Published Papers by the Author

## Articles in Journal

1. Takato Horii, Yukie Nagai, and Minoru Asada, “Modeling emotional development guided by tactile dominance and perceptual improvement during infancy”. IEEE Transaction on Cognitive and Developmental Systems, 2018.2.

2. Takumi Kawasetsu, Takato Horii, Hisashi Ishihara, and Minoru Asada, “Mexican-Hat-Like Response in a Flexible Tactile Sensor Using a Magnetorheological Elastomer”. Sensors, Vol.18, No.2, pp.587, 2018.2.

3. Takato Horii, Yukie Nagai, and Minoru Asada, “Imitation of human expressions based on emotion estimation by mental simulation”. Paladyn, Journal of Behavioral Robotics, Vol.7, No.1, pp.40-54, 2016.12.

4. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “磁性・非磁性エラストマを積層した磁気式触覚センサの基礎特性解析”. 日本AEM学会誌, Vol.24, No.3, pp.204-209, 2016.12.

5. 長井志江, 堀井隆斗. “予測学習に基づく情動の計算論的モデル”. 人工知能学会誌, Vol.31, No.5, pp.694-701, 2016.9.

## Papers in Proceedings of International Conferences

1. Takumi Kawasetsu, Takato Horii, Hisashi Ishiara, and Minoru Asada. "Size dependency in spatial response property of elastomeric tactile sensor laminated on inductor coil". In Proceedings of the IEEE Sensors Conference, 2017.

2. Takato Horii, Yukie Nagai, Minoru Asada. "Active perception based on energy minimization in multimodal human-robot interaction". In Proceedings of the 5th International Conference on Human-Agent Interaction, 2017.

3. Takumi Kawasetsu, Takato Horii, Hisashi Ishiara, and Minoru Asada. "Magnetorheological elastomer-gel tactile sensor with an electromagnet". In Proceedings of the ICRA 2017 Workshop on The Robotic Sense of Touch, 2017.

4. Takumi Kawasetsu, Takato Horii, Hisashi Ishiara, and Minoru Asada. "Object Classification with Magnetorheological Elastomer-Gel Tactile Sensor that shows difference of Gaussian like spatial response". In Proceedings of the Humanoids 2016 Workshop on Tactile sensing for manipulation: new progress and challenges, 2016.

5. Takumi Kawasetsu, Takato Horii, Hisashi Ishiara, and Minoru Asada. "Towards rich physical human-robot interaction: A novel magnetic-type flexible tactile sensor that detects its surface deformation". In Proceedings of the ICRA 2016 Workshop on human-robot interfaces for enhanced physical interactions, 2016.

6. Takato Horii, Yukie Nagai, and Minoru Asada. "Emotion Recognition and Generation through Multimodal Restricted Boltzmann Machines". In Proceedings of the IROS 2015 Workshop on Grounding robot autonomy: Emotional and social interaction in robot behaviour, 2015.

7. Takato Horii, Yukie Nagai, Lorenzo Natale, Francesco Giovannini, Giorgio Metta, and Minoru Asada. "Compensation for Tactile Hysteresis Using Gaussian Process with Sensory Markov Property". In Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots, 2014.

8. Nobutsuna Endo, Tomohiro Kojima, Hisashi Ishihara, Takato Horii, and Minoru Asada. Design and Preliminary Evaluation of the Vocal Cords and Articulator of an Infant-like Vocal Robot "Lingua". In Proceedings of the IEEE-RAS International Conference on Humanoid Robots, 2014.

9. Hideyuki Takahashi, Nobutsuna Endo, Hiroki Yokoyama, Takato Horii, Tomoyo Morita, and Minoru Asada. "How does emphatic emotion emerge via human-robot rhythmic interaction?". In Proceedings of the 2nd International Conference on Human-agent interaction, 2014.

10. Takato Horii, Francesco Giovannini, Yukie Nagai, Lorenzo Natale, Giorgio Metta, and Minoru Asada. "Contact force estimation from flexible tactile sensor values considering hysteresis by Gaussian process". In Proceedings of the 4th IEEE International Conference on Development and Learning and on Epigenetic Robotics, 2014.

11. Nobutsuna Endo, Tomohiro Kojima, Yuki Sasamoto, Hisashi Ishihara, Takato Horii, and Minoru Asada. "Design of an Articulation Mechanism for an Infant-like Vocal Robot "Lingua"". In Proceeding of the 3rd Conference on Biomimetic and Biohybrid Systems, 2014.

12. Takato Horii, Yukie Nagai, and Minoru Asada. "Toward analysis of emotional development using physiological and behavioral data". In Proceedings of the 9th ACM/IEEE International Conference on Human-Robot Interaction Workshop on HRI: A Bridge between Robotics and Neuroscience, 2014.

13. Takato Horii, Yukie Nagai, and Minoru Asada. "Touch and Emotion: Modeling of developmental differentiation of emotion lead by tactile dominance". In Proceedings of the 3rd IEEE International Conference on Development and Learning and on Epigenetic Robotics, 2013.

14. Yuji Kawai, Jihoon Park, Takato Horii, Yuji Oshima, Kazuaki Tanaka, Hiroki Mori, Takashi Takuma, Minoru Asada, Yukie Nagai. "Throwing Skill Optimization through Synchronization and Desynchronization of Degree of Freedom". In Proceedings of the RoboCup Symposium 2012, 2012.

15. Takato Horii, Yukie Nagai, and Minoru Asada. "Ancestral sampling based on energy minimization in multimodal DBN for accurate emotion estimation from ambiguous human expressions". In Proceedings of the 3rd International symposium on Cognitive Neuroscience Robotics, 2016.

16. Takumi Kawasetsu, Takato Horii, Hisashi, Ishiara and Minoru Asada. “A magnetic type tactile sensor that detects deformation of its surface made of dual layer elastomer”. In Proceedings of the Workshop “From Robotics to Cognitive Interaction and Beyond”, 2016.

17. Takato Horii, Yukie Nagai, and Minoru Asada. “Emotional Human-Robot Interaction via Multimodal Deep Belief Network”. In Proceedings of the Workshop “From Robotics to Cognitive Interaction and Beyond”, 2016.

18. Takato Horii, Yukie Nagai, and Minoru Asada. “Emotional Human-Robot Interaction through Multimodal Restricted Boltzmann Machines”. In Proceedings of the 2nd International Workshop on Cognitive Neuroscience Robotics, 2016.

19. Takato Horii, Yukie Nagai, and Minoru Asada. “Emotion Recognition and Generation through Multimodal Restricted Boltzmann Machines”. In Proceedings of the International Workshop on Cognitive Development for Friendly Robots and Rehabilitation, 2015.

20. Takato Horii, Yukie Nagai, and Minoru Asada. “A Probabilistic Approach to Analyze Temporal Change of Emotion”. International Conference on Infant Studies 2014 Pre-Conference on Computational Models of Infant Development, 2014.

21. Kiril Kiryazov, Vikram Narayan, Christian Becker-Asano, Giorgio Metta, Robert Lowe, Tom Ziemke, Takato Horii, Hector Barron. “Multimodal Emotion Recognition and Behaviour Modulation via Apprasial Model of Affect with iCub Robot”. In Proceedings of the EUCogIII Conference, 2012.

## Papers in Proceedings of Japanese Conferences

1. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “磁性エラストマと平面コイルを用いたインダクタンス測定に基づく柔軟触覚センサの開発”. 第 26 回 MAGDA コンファレンス, 金沢, 2017.

2. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “磁性エラストマを用いた柔軟な 3 軸力覚センサ”. 第 35 回日本ロボット学会学術講演会, 埼玉, 2017.

3. 堀井隆斗. “人-ロボット感情コミュニケーションにおける能動的知覚 (注意) モデル”. 日本赤ちゃん学会若手部会第 5 回研究合宿, 静岡, 2017.

4. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “Difference of Gaussian 様空間応答を示す磁気式柔軟触覚センサ”. 第 34 回日本ロボット学会学術講演会, 山形, 2016.

5. 堀井隆斗, 長井志江, 浅田稔. “Restricted Boltzmann Machine を用いた多感覚情動コミュニケーション”. 第 30 回人工知能学会全国大会, 福岡, 2016.6.

6. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “磁性・非磁性エラストマを用いた磁気式触覚センサの磁場解析”. ロボティクス・メカトロニクス講演会 2016, 横浜, 2016.

7. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “磁性エラストマを用いた磁気式触覚センサの特性評価”. 第 16 回計測自動制御学会システムインテグレーション部門講演会予稿集, 愛知, 2015.

8. 川節拓実, 堀井隆斗, 石原尚, 浅田稔. “磁性・非磁性エラストマを積層した磁気式触覚センサの基礎特性解析”. 第 24 回 MAGDA コンファレンス, 宮城, 2015.

9. 堀井隆斗, 長井志江, 浅田稔. “ノンパラメトリックベイズモデルを用いた情動変容解析”. 発達神経科学学会第 4 回学術集会, 大阪, 2015.

10. 堀井隆斗. “ノンパラメトリックベイズモデルを用いた情動変容解析”. 身体性認知科学と実世界応用に関する若手研究専門委員会第 13 回研究会, 東京, 2015.

11. 堀井隆斗, 長井志江, 浅田稔. “ノンパラメトリックベイズモデルを用いた生理指標の時系列解析に基づく情動変化のダイナミクス推定”. 2015 年度人工知能学会全国大会, 北海道, 2015.

12. 黒木隆大, 長井志江, 堀井隆斗, 池田尊司, 熊谷晋一郎, 浅田稔. “色ー情動間の対応関係を用いた自閉症スペクトラム障害者の他者情動認識支援”. 2015 年度人工知能学会全国大会, 北海道, 2015.

13. 堀井隆斗, 長井志江, Francesco Giovannini, Lorenzo Natale, Giorgio Metta, 浅田稔. “柔軟触覚センサ情報のマルコフ性を考慮したガウス過程回帰による接触力推定”, 第 32 回日本ロボット学会学術講演会, 福岡, 2014.

14. 堀井隆斗. “ガウス過程回帰による柔軟触覚センサのヒステリシス補償”. 身体性認知科学と実世界応用に関する若手研究専門委員会第 12 回研究会, 宮城, 2014.

15. 堀井隆斗, 長井志江, 浅田稔. “ノンパラメトリックベイズモデルを用いた時系列生理指標解析に基づく情動推定の試み”. 日本赤ちゃん学会第 14 回学術集会, 神奈川, 2014.

16. 堀井隆斗, 長井志江, 浅田稔. “磁性エラストマーを用いた磁気式触覚センサ”. 日本機械学会ロボティクス・メカトロニクス講演会講演論文集, 富山, 2014.

17. 小島友裕, 遠藤信綱, 笹本勇輝, 石原尚, 堀井隆斗, 浅田稔. “音声発達過程の構成的理解のための乳児様発声プラットフォームの開発”. 日本機械学会ロボティクス・メカトロニクス講演会講演論文集, 富山, 2014.

18. Nobutsuna Endo, Tomohiro Kojima, Yuki Sasamoto, Hisashi Ishihara, Takato Horii, Minoru Asada. “Design and Preliminary Experiments of an Articulation Mechanism for an Infant-like Vocal Robot “Lingua” towards natural conversation with people@home”. 第 40 回人工知能学会 AI チャレンジ研究会, 福岡, 2014.

19. 堀井隆斗, 長井志江, 浅田稔. “乳児期の触覚優位性を利用した複数感覚情報の統合による情動分化モデル”. 第 31 回日本ロボット学会学術講演会, 東京, 2013.

20. 堀井隆斗. “磁性エラストマーを用いた磁気式触覚センサの提案”. 身体性認知科学と実世界応用に関する若手研究専門委員会第 11 回研究会, 大阪, 2013.

21. 堀井隆斗, 長井志江, 浅田稔. “乳児期の触覚優位性に基づく情動分化モデル”. 日本赤ちゃん学会第 13 回学術集会, 福岡, 2013.

22. 堀井隆斗. “乳児期の触覚インタラクションに基づく情動分化モデルの構築”. 身体性認知科学と実世界応用に関する若手研究専門委員会第 10 回研究会, 東京, 2012.

23. Jihoon Park, Yuji Kawai, Takato Horii, Yuji Oshima, Kazuaki Tanaka, Hiroki Mori, Yukie Nagai, Takashi Takuma, Minoru Asada. “Differentiation within

Coordination in Acquisition of Skilled Throwing”. 第 35 回人工知能学会 AI チャレンジ研究会, 大阪, 2012.

## Book Chapter

1. Yuji Kawai, Jihoon Park, Takato Horii, Yuji Oshima, Kazuaki Tanaka, Hiroki Mori, Yukie Nagai, Takashi Takuma, and Minoru Asada. “Throwing Skill Optimization through Synchronization and Desynchronization of Degree of Freedom”. RoboCup 2012: Robot Soccer World Cup XVI Lecture Notes in Computer Science, vol. 7500, pp. 178-189, 2013.

## Patent

1. 特許出願 2017-170627. 川節拓実, 堀井隆斗, 石原尚. 触覚センサ, 大阪大学, 出願日 2017 年 9 月 5 日.
2. 特許出願 2016-146070. 川節拓実, 堀井隆斗, 石原尚, 仲田好宏, 浅田稔, 細田耕, 宮下敬宏. 変形測定装置, 株式会社国際電気通信基礎技術研究所, 出願日 2016 年 7 月 26 日.

## Award

1. Best Student Paper Award of the 5th International Conference on Human-Agent Interaction, 2017.
2. ロボカップジャパンオープン 2017 日本ロボット学会賞, 2017.
3. 第 24 回 MAGDA コンファレンス 優秀ポスター講演論文賞, 2015.
4. Honorable Mention Poster Award of the 2nd International Conference on Human-Agent Interaction, 2014.
5. 人工知能学会 研究会優秀賞, 2013.
6. ロボカップ日本委員会 ロボカップ研究賞, 2013.
7. ロボカップジャパンオープン 2013 日本ロボット学会賞, 2013.

8. 第 35 回人工知能学会 AI チャレンジ研究会 人工知能学会賞, 2012.