A Study on
Multimodal Human-Machine Communication

- Realization of Physical, Intelligent, and Kansei Interaction -

A Dissertation Presented by

Kenji Suzuki

Department of Pure and Applied Physics
Graduate School of Science and Engineering
Waseda University

March 2003

Are people machines?

- The Society of Mind, Marvin Minsky

Contents

Chapter 1

Introduction

Over the past two decades, the application of robotic and machine technology has expanded from industrial use to residential use. Advances in hardware and software computing have also enabled sophisticated implementation, that covers not only the motion control and the dynamic planning strategy but also the learning and the sensorimotor coordination for autonomous robots. Future machines and robots will be required to interact with people in a dynamic and uncertain environment. Therefore, research is specifically concerned with cognition in the context of man-machine-environment interaction.

Regarding the information sharing among people and robots, natural and intuitive multimodal communication, and the active function of thinking and understanding people's feelings are key subjects. On the other side, regarding the sharing of physical action space, robots must ensure the safety and precision of their motion, assuming physical contact with people.

Many researchers have emphasized the importance of studying human-machine communication. Multimodal communication between people and machines by means of speech, gestures and haptics are firmly focused on the key issue for the residential use of future machines and robots. The term "multimodal communication" refers to the integrated way of communication with various modalities of information such as images, acoustics, languages, gestures and facial expression. They are transmitted through human senses typified by the sense of sight, hearing, touch, smell and taste.

Multimodal communication is an extremely complicated process to which every aspect of human information processing contributes. Researchers have investigated not only medias and multimedia information but also psychology and social science that are the human-related and biologically inspired studies, for example, natural language, cognitive science, and artificial intelligence. Thus it is necessary to thoroughly conduct an interdisciplinary study of multimodal communication.

An approach that gains insights coming from human communication can indeed be used in the domain of multimodal human-machine communication. The conventional method is, however, based largely on the informational process by systems. Consequently, this means that the logical communication is the focal point. Up to today, many studies about human-computer interaction have been reported. Most of these approaches are thought of as computer-aided interaction. For example, regarding impedance control aimed at the soft and flexible control for realizing force and task fulfillment, physical interaction is virtually realized by inverse and forward kinematic calculations. Whereas, it is absolutely imperative to construct a human-machine communication system taking into account the qualitative property of communication.

Information processing in multimodal communication can be classified into three properties: that is physical, intelligent and Kansei[1] information processing. It is regarded that physical, intelligent and Kansei interactions, respectively, are caused by these types of processing.

Physical information processing has long been studied, which deals with physical data from the environment. The laws of nature underlying multimodal human-machine communication governs the behavior of both humans and machines like the rest of nature. This information processing is based on "signal processing" with special attention paid to the physical interpretation of the phenomenon.

With the improvements in information technology, intelligent information processing has become a main issue in many fields. Classical artificial intelligence has emerged as the issue. The causality of implementation is described as the logical rule. Symbols, signs, and language are used as the explanation of knowledge. Modeling is the process of describing the system in terms of mathematical equations.

Recently, the research phase has entered a new stage, that is, Kansei information processing. Kansei is a Japanese word that means something like "sensitivity", "intuitiveness" and "feeling". People have a certain feel about the use of the word "Kansei", although the scale is not measured in a quantitative way, and not visible like the feelings of people. Kansei information processing thus deals with the subjectivity of people's perceptions. The Kansei system is described in terms of emotional resonance, comfort and satisfaction.

The measurement of common human Kansei in a qualitative way, i.e., questionnaires, have been widely investigated in the engineering and psychological fields. These results are applied to product development and have achieved some positive results. Also, the measurement of psychophysical quantities by statistical methods has been significantly undertaken for understanding human Kansei, affectivity and emotion in psychology. On the

---

[1] "Kansei" as described by capitalization of the first letter is used throughout this dissertation. It can be also described by capital letters in some articles and books.

basis of these results, a modeling of the subjectivity and individual Kansei in a significant way is definitely required. This approach significantly differs from a conventional evaluation based on objectivity and logic. The subjectivity and individual Kansei should be treated as psycho-physiological interrelationship.

In this dissertation, the development of systems and robots that can interact with people in a natural, cooperative and intuitive manner is mainly focused on. In particular, an attempt at differentiating the ways of communication between humans and machines through multimodal channels will be discussed by physical, intelligent and Kansei interactions. It can be seen that this enables the system to do different and effective processing in a real, dynamic and uncertain environment. The aim is that the developed systems or robots can have the ability to provide an output in accordance with the type of input through an appropriate modality for the communication. The study of system embodiment is also the issue. The proper concepts of situatedness and embodiment are also used in different ways in these studies.

The main issue is divided into the following three parts:

1. Physical and intelligent interaction: A developmental study of an autonomous Humanoid[2] robot
2. Kansei interaction: An applied study of the construction of music-based human-robot communication
3. Kansei measurement: A basis study of the measurement of Kansei in a quantitative way

The author first introduces the modeling of a hierarchical structure of multimodal communication between humans and machine. The implications of this structure for the construction of computational and mechanical models will also be described. Afterwards, each style of interaction, i.e., physical, intelligent and Kansei will be described with some case studies. The discussion and conclusion of this dissertation will then follow.

Most systems dealing with physical interaction are carried out by the same method of intelligent interaction at the processing level. In contrast, here the author proposes a hierarchical architecture that has two independent layers in order to clearly differentiate the physical and intelligent interactions. The types of behavior caused by the computational procedures of the system are constrained by the procedures of physical interaction. The behavior of the developed robot can reflect continuous inputs from a complicated external environment so that the robot could behave in a natural and intuitive manner.

---

[2] The term "Humanoid" denotes an anthropomorphic robot designed to behave like and interact with people. The use of the term covers either or both "having human-like form" and "having human characteristics."

As an example of Kansei interaction, music-based human-robot communication has been investigated. A substantial robotic interface is constructed for the realization of an interactive musical environment for collaborative work between people and machine. The robot can be effectively used for musical performances with motion by the exploitation of the embodiment. The "moving instrument" can display the refractive motion on stage while producing sound and music by embedded stereo speakers according to the context of the performance.

Sound and music are typical channels of non-verbal communication that humans often use to express their mind. To date, many studies about the musical interaction between humans and machines have been proposed. However, few studies have described the autonomous mobile robot for musical performance although humans often accompany music with body motion. In such a multimodal musical environment, assume that the style of interaction can be classified along two axes: the robot's autonomy and direct/indirect contact. Throughout this study, an interaction paradigm of establishing virtual and real world connections by a robotic interface was investigated.

Regarding the issue of Kansei measurement, a novel artificial neural network model is proposed, which can obtain a nonlinear mapping to associate the physical features of an object with its impression. This is a new measurement method of Kansei information, that aims at embedding given objects into an arbitrary space, namely description space, under the condition where a difference or similarity (distance) between two objects is given.

Multivariate analysis is effectively used in the field of data analysis, which is a statistical method that can effectively explain and illustrate a general trend in data. The evaluation and visualization of Kansei information have been conducted by such a statistical method. Also, an artificial neural network (ANN) is often used for data analysis, as the extension of the statistical methods based on a linear model. ANN is applied to various problems, which are difficult to be conducted by conventional computers due to a nonlinear property.

The proposed model realizes learning from such a relationship. This approach will be described by comparing it to the related statistical methods and other models of neural networks.

Dissertation Organization

This dissertation is divided into eight chapters, which is organized as follows.

Chapter 2 gives an introduction of the literature and related research on multimodal communication between human and machine, and also presents a proposal for a hierarchical structure of multimodal human-machine communication. The superiority of the process by differentiating the style of communication in physical intelligent and Kansei interaction will be described. The author makes clear the features and aim of this study, and will compare the study to other works in related research areas.

Chapter 3 gives an explanation and interpretation of physical and intelligent interaction. The implications of the proposed structure will also be presented.

Chapter 4 presents the development of an autonomous humanoid robot with a double-layered hierarchical architecture. In this architecture, a signal processing layer and computational layer are hierarchized for the differentiation between physical and intelligent communications. The originality and advantage of the implemented robotic architecture will be discussed. In addition, some examples of physical and intelligent interactions will be given such as force following, motion by grasping, object tracking and reaching, reaction to environmental sound, and speech conversation. In regard to each behavior of the robot, the internal process of the system and the sensing data from external environment will be explained. Moreover, note that the network-based architecture works for the achievement of diverse and different types of communication in a simultaneous way. This allows the system to effectively execute various types of behavior in parallel.

Chapter 5 provides an introduction of Kansei interaction. Synthetic and analytical methods to understand the mechanism of Kansei interaction will be employed in the following two chapters. The traditional view of the Kansei interaction and the problem underlying the topics will be described.

Chapter 6 describes human-robot communication through music, which is considered a synthetic approach for understanding the Kansei interaction. Sound and music are typical channels of non-verbal communication in which Kansei plays an important role. The proposed approach to equip musical instruments with an autonomous mobile ability will provide for a new computer music performance in the real world. The construction of three mobile robots for music-based human-robot interaction is described along with a model of the human-machine-environment.

Based upon the proposed model, these robots have been developed in collaboration with composers and choreographers. They performed not only in the experimental laboratory but also at a public exhibition and demonstrations. This explains the credibility and advantage of practicability. Although it is difficult to evaluate and assess the effects of these robots, the author believes that the development of these robots that can perform in the real world is a worthwhile subject.

Chapter 7 presents a new model for the measurement of Kansei. Regarding the proposed model of artificial neural network, the algorithm, the structure, and the mathematical formulations will be given. By comparing the Multi-Dimensional Scaling method (MDS) and Hayashi's Quantization Theory IV (QT-IV), the mathematical proof for applying the related problem will be followed. Also, the proposed model is applied to a learning environment that is composed of learning objects with physical features along with its impression as an example of Kansei information. It is proved that the algorithm can effectively associate between physical features of an object and its Kansei information.

In the proposed method, if the distance between two input patterns is given as a teacher signal, the network can obtain a nonlinear mapping from an input space to an output space under the condition so that the given distance is preserved. This is a new method of Kansei quantization by mapping from a physical pattern space to another space based on the psychological patterns. In addition, new data, which are not used in the training in the network, can also be evaluated by the generalization ability of the network. Also, by applying an individual data set on impression to the system, visualization of the nonlinear mapping is used at an early stage to measure differences among individuality and characteristics common to all.

Chapter 8 summarizes the contributions made in this dissertation. The discussion throughout these basic and applied studies will be also given. The conclusion of this dissertation and suggestions for further researches are also included.

Chapter 2

Objectives and Approach

2.1 Multimodal Communication

Graphical User Interface (GUI) is recognized as a confirmed technology for people to interact with computers. When someone uses a computer - consumer use by normal means - he/she works by hitting the keyboard key with their fingers, and moving the mouse using their arm and hand, while looking at the monitor and hearing the reacted sounds such as beeps, key clicks from the speakers, etc. These can be carried out with the aid of GUI technology. Usability[1] is improved at a rapid rate with an improvement in the related technology. This shows that people use multimodality for their interaction with a computer, and the interaction with multimodality has enhanced the paradigm of Human-Computer Interaction.

The term *modality* refers to the input and output channels of humans. In the human view, the communication channel is composed of sense organs. The modality and sense organ is tightly coupled with perception. Of course, the processing is carried out in the nervous system, and muscle-brain circuit. **Table 2.1** shows the different senses and their corresponding modalities and sense organs as defined in physiology [Charwat, 1992]. As noted by Shepherd [Shepherd, 1998], the notion of the human sensory modality can be divided into seven groups including its internal/external chemical reactions at the neurophysiological level as shown in **Table 2.2**.

The understanding of visual and acoustical modality has received much more attention than the other modalities. Some of the sensory modalities are not taken up due to their controversial aspects and other reasons. The sense of balance does not have a cortical representation. Taste is not a very useful channel of man-machine interaction. People usually do not want to taste or "bite" a machine or an interface.

---

[1] In fact, it is not easy to describe the usability of an interface in a quantitative way. This is also a subject in which Kansei plays an important role.

Sensory perception	Modality	Sense organ
sense of sight	visual	eyes
sense of hearing	auditive	ears
sense of touch	tactile	skin
sense of smell	olfactory	nose
sense of taste	gustatory	tongue
sense of balance	vestibular	organ of equilibrium

Table 2.1. Different senses and their corresponding modalities and sense organs as defined by physiology. This table is originally quoted from [Silbernagel, 1979], which is simplified and modified in [Schomaker et al., 1995].

2.1.1 Communication in interaction

Here, the term *Human-Machine Communication* is used as against *Human Human-Robot Interaction.* The communication aspects are thus focused on, that is, the context of physical interaction such as cooperative handling of pieces and task fulfillment will not be considered.

In the information exchanged between people and machine, the semantics are different according to the direction. The user - usually a human - first configures the robot by assembling and programming, and specifies a particular task. Furthermore, the user, as a partner, may supervise the robot to achieve the given task and provides an evaluation of the robot's performance.

In the residential environment, it is indispensable to design interfaces which allows untrained people to make efficient, intuitive and safe use of a robot. People must be provided with an interface that allows him/her to intuitively interact with the robot. The need for enhancing Human-Machine Communication is closely related to the idea of allowing humans to make use of robots.

An ecological view to communication is a more basic attitude. Lindström et al. [Lindström et al., 1999] stated that communication is the process whereby individuals send and receive information (information exchanged) about each other and their surroundings. Communication is achieved through the use of signals, i.e., traits that have specially evolved to transfer information between one individual (the signaler) to another (the signal receiver). The robot does not need to construct its own symbols for communication purposes, but utilizes the user-defined symbols for its own perceptions and actions.

The purpose of communication is tightly related to the level of communication. Consequently, the following two keywords are considered.

1. The path: the information exchanged in which modality and how it

Sensory modality	Form of energy	Receptor organ	Receptor cell
Chemical (internal)			
blood oxygen	O_2 tension	carotid body	nerve endings
glucose	carbohydrate oxidation	hypothalamus	gluco-receptors
pH (cerebrospinal fluid)	ions	medulla	ventricle cells
Chemical (external)			
taste	ions & molecules	tongue & pharynx	taste bud cells
smell	molecules	nose	olfactory receptors
Somatic senses			
touch	mechanical	skin	nerve terminals
pressure	mechanical	skin & deep tissue	encapsulated nerve endings
temperature	thermal	skin, hypothalamus	peripheral & central
pain	various	skin & various organs	nerve terminals
Muscle sense[*1]			
muscle stretch	mechanical	muscle spindles	nerve terminals
muscle tension	mechanical	tendon organs	nerve terminals
joint position	mechanical	joint capsule & ligaments	nerve terminals
Sense of balance			
linear acceleration	mechanical	sacculus/utriculus	hair cells
angular acceleration	mechanical	semicircular canal	hair cells
Sense of hearing			
hearing	mechanical	cochlea	hair cells
Sense of vision			
sight	mechanical	retina	photoreceptors

Table 2.2. Human sensory modality at the neurophysiological level. [*1]Muscle sense, *Kinesthesia*, means perception of body movements in physiology and psychology. It is the perception that enables one person to perceive movements of the own body. [Shepherd, 1998]

is transmitted.

2. The causality: an event is caused by what kind of input stimulus.

2.1.2 Harmonized human-machine environment

With the aid of the growth of multimedia technology, human-machine communication by means of speech, gestures and haptics has been implemented in various scenes of our lives. Nowadays, the improvements in technology have enabled us to develop human-like robots (for example, [Hirai et al., 1998] [Brooks et al., 1999] [Hashimoto et al., 2002]). Advances in hardware and software computing have also enabled sophisticated implementation such as motion control and dynamic planning strategy. Recently, human-cooperative robots have become more widespread all over the world. The research on a humanoid robot [Lim et al., 1999][Miwa et al., 2001] [Ogata et al., 2000] [Iwata et al., 2001][Tojo et al., 2000] has also been widely extended from the mechanical realization to the biological analysis of human beings.

The aim is to build a harmonized environment, where people and machines can "live" together and interact with each other. This "harmonization" is defined as the naturalness and intuitiveness of communication. The machine is required to make its own decision according to the precise selection of communication channels. In this kind of communication, multimodality is one of the key issues. It provides natural, seamless and intuitive communication between humans and machines. The robot is a machine with a mobility and high redundancy. It allows humans to interact with it in various ways. The machine also should be able to achieve a given task in various ways.

In recent years, pet-type robots have become commercially available. They can exhibit attractive and devoted behavior so that the audience is satisfied with its performance. These are examples of an advanced interface that has a substantial body with multimodality. Such multimodal human-machine interactions typified by non-verbal communication have been widely investigated. In communication among people, non-verbal communication plays an important role, sometimes more important than verbal communication. A classic psychological work [Mehrabian, 1972] remarked that only 7% of the meaning of a message is communicated through verbal exchange. On the other side, 55% of the meaning of a message is expressed through non-verbal ways, such as facial expression, posture, and gesture. The remaining 38% is dependent upon the voice quality such as tone and intonation.

In the sketched harmonized human-machine environment, the robot would behave in response to given stimuli and its internal state in a real environment. The robot can continue to interact with the humans who do collaboratively work and play together, even in situations where unexpected inputs, disturbances and interruptions occur.

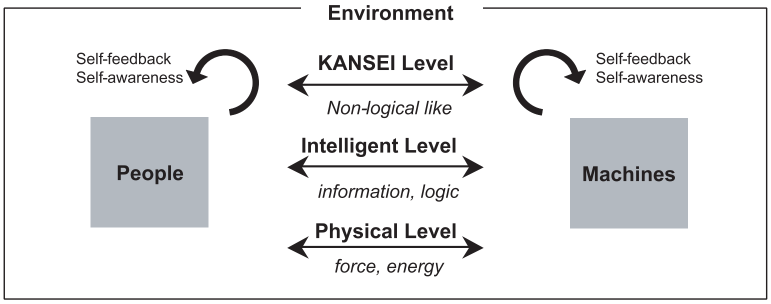


Figure 2.1. A layered communication model: there are three levels of communication according to the style of interaction; physical, intelligent and Kansei interaction.

2.2 Overview of the Approach

As described in the previous section, assume that a communication between human and machine is modeled with a hierarchy as illustrated in **Figure 2.1**. There exist three styles of communication from a qualitative point of view: physical, intelligent and Kansei interaction.

A humanoid robot "*iSHA*" (interactive Systems for Humanoid Agent) have been developed, which is designed to behave like and interact with humans. An intelligent robotic architecture is implemented, which integrates goal-oriented subsystems by taking the flexibility and scalability of the system into consideration.

So far, most robotic systems have been designed for achieving a particular task. In this conventional view of system design, one module or one function performs one task. On the contrary, a complicated system such as a humanoid robot should be designed from an integrated point of view. For instance, some integrated systems for a humanoid robot have been reported [Cheng et al., 2000a][Cheng et al., 2000b][Imai et al., 1999]. These systems embed a mechanism through channeling all inputs into an integrated system in a competitive and cooperative manner.

iSHA has an upper body resembling a human in shape and a mobile base with two wheels. The upper body with a head and two arms has 24 degrees-of-freedom. Two wheels situated under the body provide a safe and robust locomotion. Each eye equipped with a small CCD camera, small microphones embedded in the head, and touch sensory devices on the body provide binocular vision, auditory and touch sensing abilities to the robot, respectively.

Regarding the issue of Kansei interaction, a substantial robotic interface is proposed for the realization of an interactive musical environment for collaborative work between humans and machines. This is a new paradigm of a human-machine Kansei interaction through sound and music. The robot can be effectively used for musical performances with motion by exploitation of the embodiment. The "moving instrument" can display the refractive motion on stage while producing sound and music by embedded stereo speakers according to the context of the performance.

Sound and music are typical channels of non-verbal communication that humans often use to express their mind in which Kansei plays an important role. So far, many studies about musical interaction between humans and machine have been proposed. However, few studies have described the autonomous mobile robot for musical performance although humans often accompany music with body motion. In such a multimodal musical environment, the style of interaction can be classified along two axes: the robot's autonomy and direct/indirect contact. Throughout this study, an interaction paradigm has been investigated for establishing virtual and real world connections by a robotic interface.

The moving instrument works as a sort of reflector to create an acoustic and visual space in the multimodal environment. The proposed approach to equip musical instruments with an autonomous mobile ability is providing new computer music performances in the real world.

Regarding the issue of Kansei measurement, a new method of Kansei quantization using an improved artificial neural network model will be described. The mechanisms and the process of human perception have been widely discussed in many research fields. In the process of perception, one possible explanation is that physical features in a high-dimensional space perceived by humans are mapped onto another low-dimensional space in the human mind which contains semantic parameters.

The purpose of this method is to construct a non-linear mapping that associates between physical features of an object and its impression. The network can obtain the non-linear mapping between the input objects and the outputs by providing the desired distance between the objects, not the desired output. The desired distance represents the similarity between the input objects. By applying this method to real world problems such as a modeling of emotional facial expression and multi-class classification, the verification of the method is discussed.

2.3 Related Work

The term *Artificial Intelligence* has attracted many researchers in many fields over the past few decades. Up to now, many influential AI approaches have been found not only in the engineering field but also psychological and philosophical fields. For example, Minsky [Minsky, 1986] has portrayed the mind as a society of tiny components, namely *agent*, that do not themselves have a mind, as illustrated in **Figure 2.2** Intelligence is explained as a

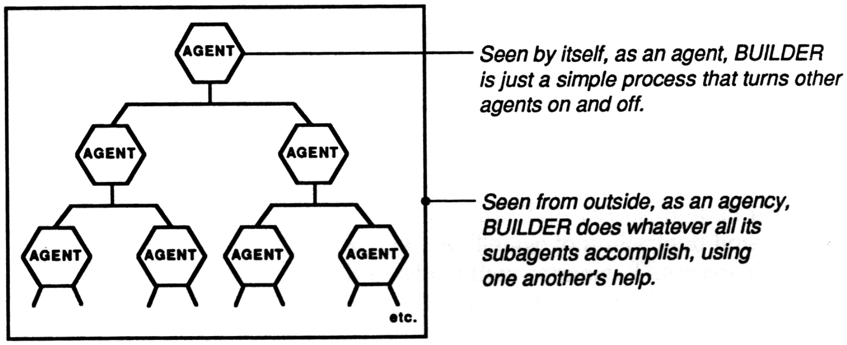

AGENT
AGENT
AGENT
AGENT
AGENT
AGENT
AGENT
etc.
Seen by itself, as an agent, BUILDER is just a simple process that turns other agents on and off.
Seen from outside, as an agency, BUILDER does whatever all its subagents accomplish, using one another's help.

Figure 2.2. *The Society of Mind*: a figure describing an agent and an agency, quoted from [Minsky, 1986, p.23]. This book explains how minds work and how can intelligence emerge from nonintelligence.

combination of simpler things. It is not enough to explain only what each separate agent does, but should understand how a group of agents, namely *agency*, does - that is, those parts are interrelated.

Many different approaches to AI must be pursued. A definition of Artificial Intelligence is "the field of research concerned with making machines do things that people consider to require intelligence. There is no clear boundary between psychology and Artificial Intelligence because the brain itself is a kind of machine." [Minsky, 1986] This is quoted from the glossary in the back of the book.

On the other hand, the terms artificial life and evolutionary computation have become some of the key issues in robotics research. For instance, applications of evolutionary robotics (for example, [Brooks, 1997, Fogel, 1999, Harvey et al., 1996, Watson et al., 1999]) have often been conducted. With the improvement in large-capacity batteries and high-power inverters, autonomous and self-subsistent ability can be integrated into the robot. These characteristics are also essential factors in order for the robot to coexist with humans.

The survival ability of creatures has lately received attention in artificial life discussions [Steels, 1993]. With the aim of true and complete autonomy, survival robots have developed to enable a small mobile robot to learn and obtain survival strategy on how and when to recharge its own battery [Birk, 1996][Tabe et al., 2001].

The study of man-machine symbiotic systems have been emphasized with the aim of investigating how people and machines interact with one another through action and perception. Klingspor et al. [Klingspor, 1997] remarked that the three most important factors for the Human-Robot Communication (HRC) are: purpose of communication (i.e., the purpose of the exchanged

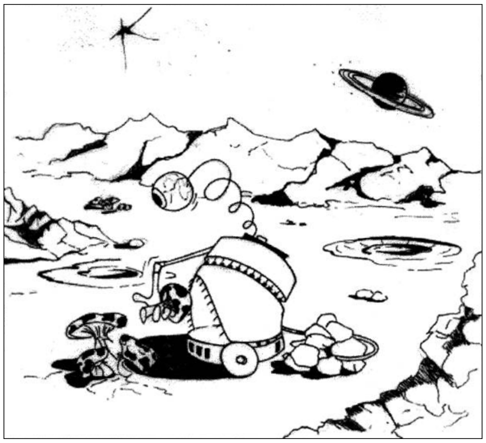


Figure 2.3. Fungus Eaters, which are robots to live on a far planet, to collect ore for a reward, and to survive itself for the maintenance, rather to live there. This figure is quoted from [Toda, 1962]

information and the level of abstraction), communication media (verbally, via gestures, via explicit interaction) and direction of communication (vise versa or in both directions). This thesis study are much paid attention to the level of communication.

Graefe et al. developed a humanoid service robot *HERMES* focusing on dependability [Bischoff et al., 2002]. They argue that dependability is a consequence of fundamental design decisions in the context of intelligent experimental robots. The developed humanoid robot *HERMES* was presented to the general public every day for 6 months in a special exhibition with only 3 failures: one motor controller, one motor driver and one audio amplifier. An improved dependability of intelligent robots are explained by:

1. Learning from nature how to design reliable, robust and safe systems
2. Providing natural and intuitive communication and interaction between the robot and its environment
3. Designing for ease of maintenance
4. Striving for a tidy appearance

The dependability directly contributes the credibility of a robot. It will be much required for future machines, especially human cooperative robots. A long-term dependability test can be conducted at a public exhibition.

Pfeifer et al. emphasized the importance of the embodiment and situatedness of the intelligent robot [Pfeifer et al., 1999]. He created the first Learning Fungus Eater as a physical material object. The Fungus Eater learned to avoid obstacles in order to evade them and run along walls. This behavior left the impression of emotion with observers. Goals are awarded and motivations emerged - being observed -, although it simply interacted with its environment. However, it could not sufficiently supply itself with food, since the corresponding motivation was not inserted into it. This circumstance of the autonomy was then considered in the Self-sufficient Fungus Eater, with the help of only one logical rule, that is, the motivation to eat mushrooms. Also, by observers despite a relatively simply structured behavior, a high emotional intensity was endowed to this model.

Fungus Eaters were sketched by Masanao Toda, who is a psychologist in the 60's: The Fungus Eaters are robots, whose main task is it to collect uranium ore on a far planet for which a reward is paid (**Figure 2.3**). In order to maintain their energy level, they can feed themselves with mushrooms, which are planted and grown. Since all activities of the robots use energy, including thinking, they are constantly forced to ponder the decision that occurs between ore collecting and mushroom consumption. Besides there are mechanical obstacles and changing environments. From these relatively simple basic assumptions, complex decision conflicts result. This model is sketched for the experiences with these natures, the psyche of people and their emotions.

Although the survival ability will not be discussed in this thesis study, the author considers that the ability is a fundamental factor of an autonomous robot. Four different types of robots that appear in this dissertation are all battery-operated. This concept underlies the development of an autonomous robot.

Chapter 3

Physical and Intelligent Interaction

Communication among humans consists of two types: physical and intellectual interactions. The former is the communication under physical constraint and interaction in accordance with direct/indirect contact. The latter is informational interaction through intellectual ability of a high order with each other. Humans can do either of these interactions or a combination of both types of interaction. For instance, speech conversation is one of the most effective and intellectual interactions between humans but is not physical interaction. In contrast, the hand shaking has two aspects: people express a sign of goodwill with this motion and also gain force interaction according to the motion at the same time. These interactions are caused by different demands and processing. Physical interaction is composed mainly of immediate responses and simple types of behavior. Intellectual interaction is composed of intelligence and sophisticated types of behavior.

	Physical Interaction	Intelligent Interaction
Demand	Reflexive motion (Real-time) Steady-continuous	Intellectual motion (Intelligent) On-and-off
Processing	Simple	Complicated

Table 3.1. The characteristics of physical and intelligent interactions. The details of the processing will be described in the following sections.

Consequently, an architecture that can clearly separate the physical and intelligent interaction not only by the computational framework but also at the processing level for providing multiple operating systems to the robot.

Style Modality	Physical Interaction	Intelligent Interaction
Sight	Camera-based sensing	Tracking/reaching objects
Hearing	Stereo microphone Sound sensing	Sound localization Speech Conversation Dancing movements Pitch detection
Touch	Body touch Handshakes	Various types of Behavior Reaction to touch sensing Force Following Handshakes
Smell	(by Olfactory organ)	(Distinguishing flavors)
Taste	(by Gustatory organ)	(Distinguishing foods)

Table 3.2. The implemented style of interaction between people and the robot. The details of the processing will be described in the following sections.

Most of the systems dealing with physical interaction, including computational compliance control, are carried out by the same method as intellectual interaction at the processing level. In contrast, this is a new approach for processing method based on a hierarchical architecture that has two independent layers in order to clearly separate the physical and intellectual interactions. The types of behavior caused by the intellectual procedures of the system are constrained by the physical procedures. Assume that the mixture of processing can thus provide safety and credibility for communication to the machine.

In the next chapter, as an example of physical and intelligent interactions, a development of a humanoid robot platform is described, which integrates a number of agents such as image and speech processing, and a haptic interface. Also, a robotic architecture by taking into consideration a physically grounded approach is proposed. The architecture allows various types of behavior executed in parallel. The characteristics of the robotic design are 1) autonomous and self-subsistent ability 2) system plug-in and behavior plug-in architecture, and 3) human-like modalities.

Chapter 4

Development of a Humanoid Robot

In this chapter, an overview of a developed humanoid robot and some experimental results is presented. Illustrating the novelty and effectiveness of the proposed approach with examples of both physical and of intelligent interaction. The developed robot has several distributed agents that can work independently. Each agent has channels of communication between human and machine in a multimodal environment.

4.1 Robotics System Design

An architectural framework for sensing and reasoning processes should allow the robot to display goal-oriented behavior and should preserve the ability to respond to critical situations in a real-time environment.

Some key points to be considered in the design of a planning and control architecture are that the robotic architecture should be distributed, allow both reactive and deliberative reasoning, and involve a method for dealing with information from multiple sources.

The architecture developed for control of Shakey the Robot is well-known as a centralized architecture [Nilsson, 1980]. The robot operated by gathering all available sensory data and creating a unified representation of its environment. Although the centralized architecture has the advantage of enabling the robot to behave autonomously in a coherent fashion and with multiple goals, it is not appropriate for a real-time system in a dynamic and uncertain environment. In contrast to the centralized architecture, subsumption architecture ([Brooks, 1986] as illustrated in **Figure 4.1**) that employs priority-based arbitration is one of the representative instances of behavior-based architecture [Arkin, 1998]. In the architecture as typically described, simple types of behavior are hierarchically organized so that more complex types of behavior emerge. A robot control system should be de-

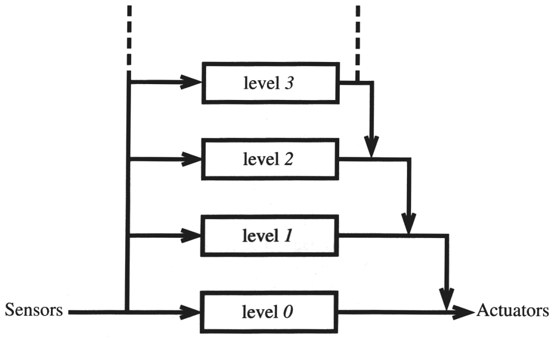


Figure 4.1. Subsumption Architecture: Control is layered with higher level layers subsuming the roles of lower level layers when they wish to take control. The system can be partitioned at any level, and the layers below form a complete operational control system. [Brooks, 1986]

composed according to not the structure of the internal functions but the desired behavior of the machine in response to the external environment. Such behavior-based architecture for the supervision of mobile robots is recently in wide usage as intelligent robotics architecture. For example, an approach to build a sociable robot with the subsumption architecture has been reported [Breazeal et al., 2000]. It is, however, not a physically grounded architecture but a computational one based on intelligent processing.

4.1.1 Double-layered hierarchical architecture

In the developed robot, a double-layered structure is adopted, as illustrated in **Figure 4.2**. It should be noted that a double-layered hierarchical processing is implemented in order to clarify the stimulus difference.

There exist two layers: signal processing layer and computation layer. As for the signal processing layer, data from touch sensing devices are fed to the behavior coordinator. The cells of the output layer correspond by signals to the actuators.

As for the computation layer, sensing data are given to input cells from a sensing module installed in the robot (e.g., low-level robot sensor data, equipped stereo cameras, microphones and tactile sensors), and each cell has a unique source from an input channel. In the internal procedure in the computation layer, there are two types of behavior coordinator in which each one receives signals independently from all input cells. These coordinators are based on short-time and long-term memory. Each behavior coordinator is connected to and has influence on the others by activation and inhibition. A weighted sum of the input signals is fed to these coordinators. Each one corresponds to a style of behavior of the robot (e.g., dancing movements, binocular object tracking, response to tactile sensing). Specifically, a signal

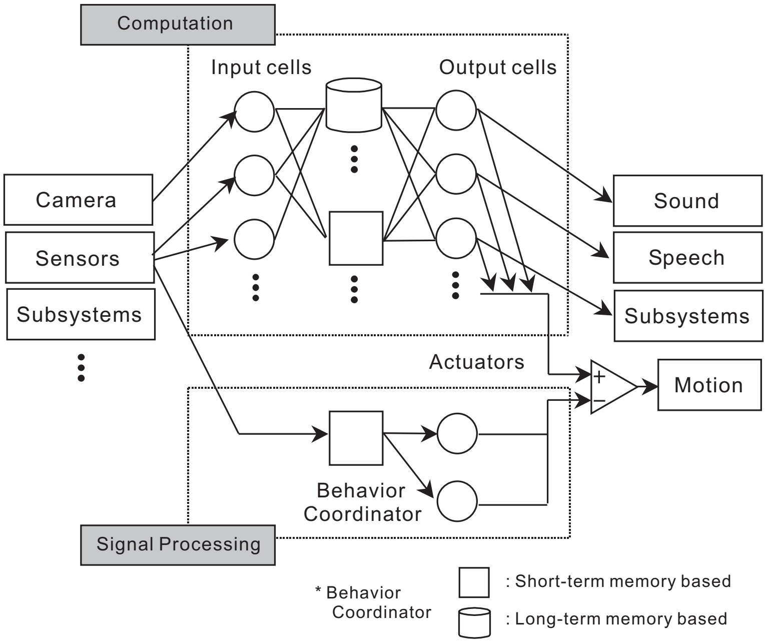


Figure 4.2. *iSHA* - Robotic Architecture: consisting of two layers, signal processing layer and computation layer that have a role of performing physical interaction and intelligent interaction, respectively.

in the input cells is transmitted to the behavior coordinator. The output of each coordinator is then multiplied by certain (fixed) weight parameters and transmitted to the output cells. Each cell has a unique connection to output channels such as actuators, sound and visual outputs. The operations through the connection constitute a linear combiner.

The signal to actuators from both layers is simply summed by the analog adder and is fed to each actuator. That is, the behavior of the robot depends upon the balance of the signals from both physical and intellectual processing layers. That is, the signal to actuators represented by the target angle for each joint $\hat{u}_i$ (i=1,2, ..., 26) is simply described as:

$$\hat{u}_i = \sum_{j=0}^{N} w_{ij} b_j \qquad (u_{imin} < \hat{u}_i < u_{imax}) \tag{4.1}$$

where N denotes the number of behavior module, and b_j represents the output of jth behavior coordinator.

In the present work, the connection between behavior coordinator and the weight parameters are predetermined and fixed. Consequently, the meaning of each type of behavior is predetermined by the system designer. The multi-goal tasks are achieved because each behavior coordinator produces an action independently.

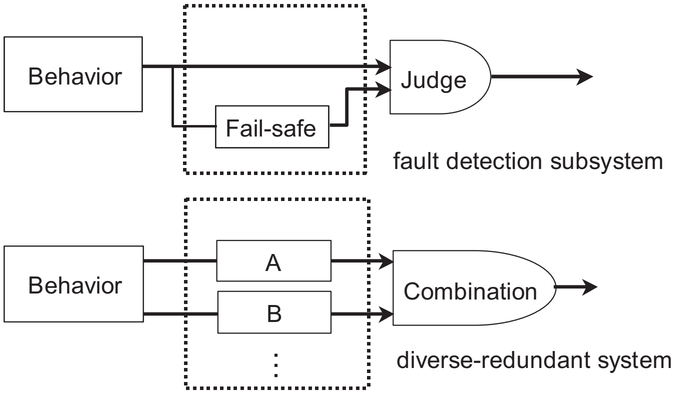


Figure 4.3. Fault detection system and diverse redundant system: upper figure illustrates the conventional system that contains fail-safe subsystem. While, lower figure illustrates a system that holds two (or more) complete operating subsystems.

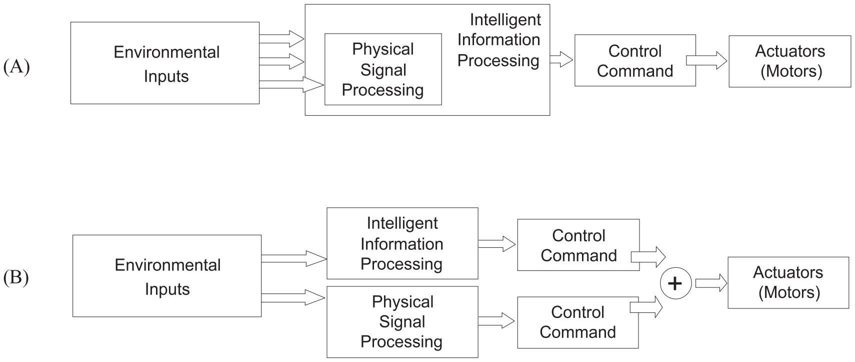


Figure 4.4. Processing flow: The upper figure illustrates the conventional processing that is a single control method even if the informational processing is distributed. While, the lower figure shows the proposed processing method that holds multiple control ways.

Note that this layered architecture enables the robot to hold multiple control ways. At present, the control command for the robot comes from two lines, which means holding two independent operating systems inside the robot. This characteristic is very important from safety point of view.

Fail-safe system is designed to go into a safe mode if and when the system happens to fail as illustrated in the upper figure of **Figure 4.3**. It reduces risk of contamination and provides a highly reliable system with. These systems contains one or some subsystems for monitoring the stabilization of the system. Once an unexpected event happened to the system, the fail-safe subsystem alarms or lockout system for the safety.

This fail-safe system, however, is tied to the fault-tolerant scheme. The problem on the fault detection becomes a subject of discussion. Moreover, the fault detection system is required highly reliable as well as the target system of it own.

On the other hand, the proposed architecture consists of a diverse and redundant system as illustrated in the lower figure of **Figure 4.4**. Here, two subsystems works independently and provides reliability to the whole system.

For instance, in case the user encounters dangerous situations - ill-posed and subternatural behavior by the robot -, he/she can restrain the behavior by pushing the robot body parts. This means a compensation method with an appropriate control command from another channel, which differs from system lockout or shutdown. That is, although command for forwarding motion is processed in mind, the motion does not result since the body is pushed from the front. In that case, because intelligent processing in mind of the robot (realized by computers) and physical signal processing in the whole body (realized by digital/analog circuit) are mechanically differentiated, the control command (from mind) and a real behavior of the robot is not the same but different from each other.

4.2 System Overview

As a first step in the realization of a harmonized human-machine environment, a humanoid robot has been developed. The robot's behavior can reflect continuous inputs from a complicated external environment so that the robot could behave in a natural and intuitive manner. The robotic architecture involves two types of processing: physical and intelligent (logical). The multi-process and independently distributed modules provide adaptive and robust control to the robot.

The developed robot *iSHA* has a number of degrees-of-freedom in its body, especially around the head. Humans can make a variety of body expressions by using muscles. In place of the muscles, actuators (DC or AC) on joints conduct the behavior of the robot. **Figure 4.5** shows the overview of the developed humanoid robot.

iSHA can be divided into two body parts, an upper and a lower body. The upper body resembles a human in shape, while the lower body is a wheelchair. The upper body with a head and two arms has totally 24 DOFs; 8 for the head, 4 for the neck, and 6 for each arm. The lower body has two wheels, which are independently driven, that provide safe and robust locomotion to the robot. The total is thus 26 DOFs in its whole body. In particular, the eye structure has actuators independent of the head movement. This therefore helps the robot to achieve a fast object tracking.

The host computer (ART-Linux, Celeron 700MHz) that works to control the actuators with the hardware-scheduled real-time process is embedded in the backside of the robot. The images obtained by two small CCD cameras are transmitted to another embedded computer (Windows 2000, Pentium III 800MHz) that is engaged in the image processing with the image processing

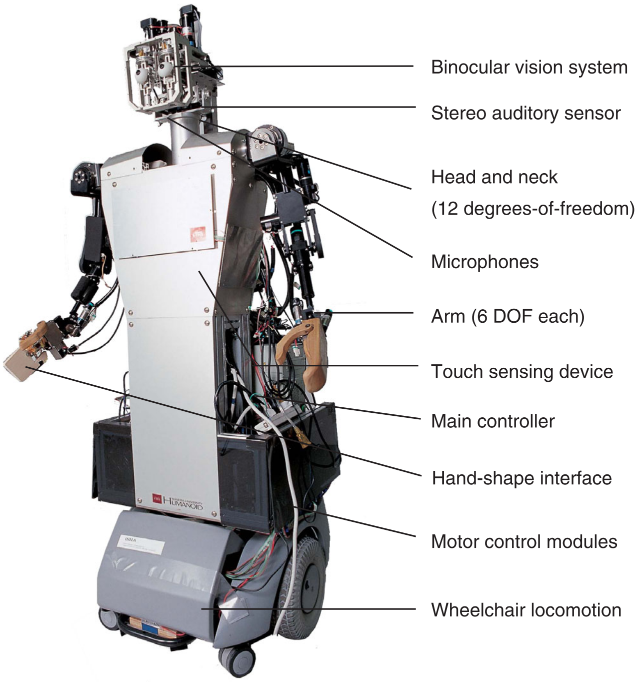


Figure 4.5. Autonomous humanoid robot "*iSHA*": The upper body with a head and two arms has totally 24 DOFs; 8 for the head, 4 for the neck, and 6 for each arm. The lower body has two wheels, which are independently driven.

board (Hitachi IP5000), as well as the processing of the data from the sensory receptors of the microphone, the sound-sensing devices and tactile devices.

4.3 Robotic Architecture

The characteristics of the developed humanoid robot are summarized as follows:

4.3.1 Autonomous and self-subsistent ability

Most individual creatures have an autonomous and self-subsistent ability. These are focused on as the fundamental character of system design. The developed humanoid robot does not need any power supplier from the external environment but can itself move and act with an embedded lead storage battery. Two included computer can then make the robot autonomous.

4.3.2 System plug-in and behavior plug-in

The developed robot has a substantial interface integrating a number of multimodal components. In such a system, it is desirable that any subsys-

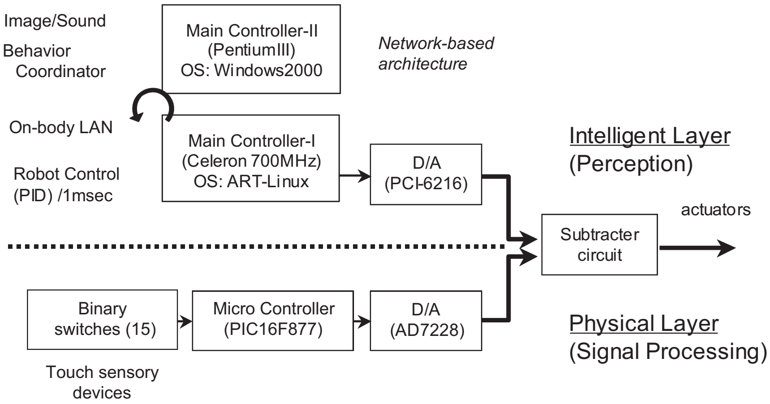

Image/Sound
Behavior Coordinator
Main Controller-II (PentiumIII) OS: Windows2000
Network-based architecture
On-body LAN
Robot Control (PID) /1msec
Main Controller-I (Celeron 700MHz) OS: ART-Linux
D/A (PCI-6216)
Intelligent Layer (Perception)
actuators
Subtracter circuit
Binary switches (15)
Micro Controller (PIC16F877)
D/A (AD7228)
Physical Layer (Signal Processing)
Touch sensory devices

Figure 4.6. The double-layered structure: the *Physical Layer* corresponds to signal processing layer, while the *Intelligent Layer* corresponds to perceptional layer in **Figure 4.2**

tems can be added to the existing robotic system. As for the flexibility and scalability of instruments, a network-based architecture inside the body has been adopted. **Figure 4.7** illustrates the modules such as a speech recognizer/synthesizer, an image processor, a behavior coordinator, a receptor of auditory sensing, and a touch sensing and haptic interface device that have been developed as server applications. Moreover, the robot can be easily connected to the local area network via an embedded wireless connection. The robot, therefore, not only can perform autonomously but can also be handled by a remote control operation over a TCP/IP network connection. The proposed robot thus allows a system plug-in extension with the aid of a general Ethernet connection. Moreover, connecting an existing network and internet, the robot can access to a database and world-wide-web system seamlessly. This feature has enabled the robot be as an application for humanoid network interface.

As for the robotic architecture, the concept of a behavior plug-in has been realized by a behavior-based architecture and a multi-process operating platform, ART-Linux, which is a real-time operating system designed for support development of large-scale real time processing software. In short, the real time processing performance is added to the Linux operating system. Thus, any abilities or types of behavior can be flexibly attached in the developed robotic architecture . The hardware scheduling is guaranteed by the operating system. The details of software architecture are described in the next section.

4.3.3 Human-like modalities

iSHA has several channels of communication with the external environment. These are designed so as to enable humans to give stimuli to the robot in

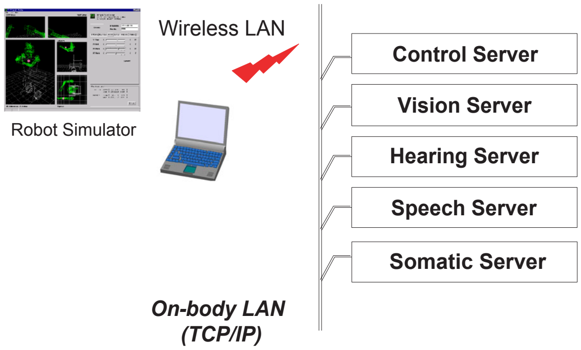


Figure 4.7. Network-based modules: These modules have been designed and developed as server applications. Information through them is exchanged via TCP/IP protocol.

the same way as communication among humans. Each eye is equipped with a small CCD camera that provides binocular vision to the robot. The equipped stereo microphones provide an auditory sense and enable the robot to receive environmental sounds. The robot can thus execute simple sound localization.

Touch-sensing devices have been implemented, which consist of a metal plate and a cushioning material on the front, back and left sides of the body. Additionally, the robot has a hand-shaped force interface to sense the human's intention by hand shaking that is one of the intuitive ways of communication. These allow us physical interaction with the robot through our hands or body.

4.4 Experiments

In this section, some preliminary experimental results are introduced as examples of the two styles interaction: 1) Reaction to touch sensing, 2) Hand shaking, 3) Tracking and reaching an object by the binocular vision system, 4) Dancing movements according to a given tempo, 5) Reaction to auditory sensing and 6) Integrated types of behavior.

The robot allows humans around it to behave freely. The robot's performance is designed for human intuitive understanding so that each type of behavior can be accepted easily by the companions who interact with the robot. Throughout these experiments, the effectiveness of the proposed robotic architecture will be evaluated. The robot performs an action in response to human stimuli in real-time. The types of behavior are chosen and carried out, depending upon the robot's priorities.

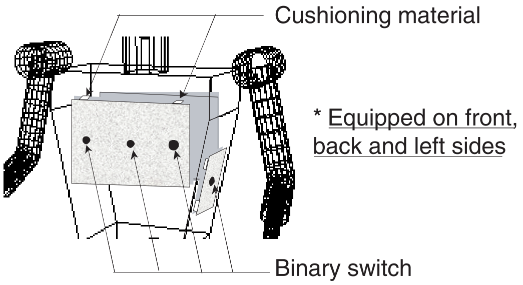


Figure 4.8. Equipped touch-sensing devices: The device on the front or back side has three switches, while the one on the left side has one switch.

In the following experiments, most actuators are operated by position control. By obtaining the angle of each joint with the embedded encoder, the control module provides each joint with the desired angle θ_d. Conventional PID control is applied for each joint. The PID gain parameters are empirically chosen and fixed through experiments. The control module independently processes in the robot operating system by parallel computing.

$$u(t) = -K_p x(t) - K_d \dot{x}(t) - K_i \int_0^t x(\tau) d\tau \tag{4.2}$$

where $x(t)$ denote the position at time t, K_p, K_d, and K_i represent the proportional, derivative and integral controller parameter, respectively. In this work, a discrete PID control (velocity form) is adopted to control for all actuators except for vehicles.

$$\begin{aligned} u(t) &= u(t-1) + \Delta u(t) && (4.3) \\ \Delta u(t) &= K_p(e_t - e_{t-1}) + K_i e_t && (4.4) \\ &\quad + K_d((e_t - e_{t-1}) - (e_{t-1} - e_{t-2})) && \end{aligned}$$

where e_t represents deviation at time t. That is, $\Delta e(t)$ is represented by $(e_t - e_{t-1})$.

4.4.1 Reaction to touch sensing

Reactive movement is one of the basic types of behavior implemented at the physical signal processing layer. Physical interaction with people is the highest priority for the robot.

The developed touch-sensing devices are illustrated in **Figure 4.8**. The device on the front or back side has three switches, while the one on the left side has one switch. By applying an external force to the plate, the robot can obtain a human's intention by physical interaction. In this experiment, the robot moves in a direction so as to cancel the applied external force. The

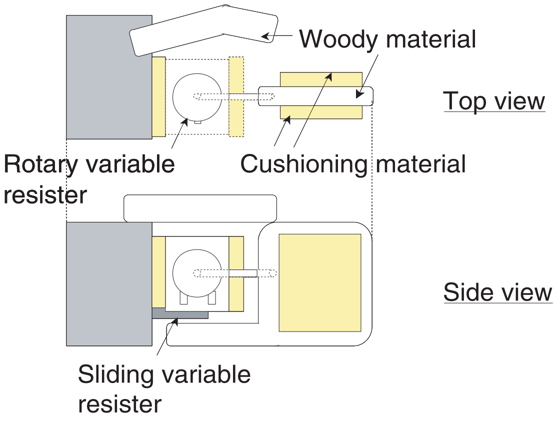


Figure 4.9. Equipped hand-shaped force interface: The interface that is embedded in the right arm enables us to communicate with the robot by shaking hands.

robot moves left, right, backward and forward according to the combination of switches.

For example, a push from the front side causes the robot to move backward. Although the robot is tracking an object and moving his arms to reach the object, physical interaction can still be possible. Therefore, the robot would respond immediately trying to continue tracking the object under the given constraint.

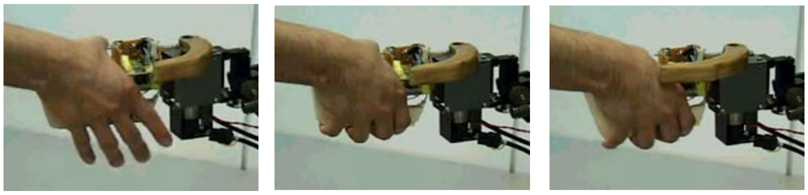
4.4.2 Hand shaking and tracking humans

A hand-shaped force interface have been developed as illustrated in **Figure 4.9**. The interface consists of 2 DOFs with rotary variable resisters and 1 DOF with sliding variable resisters so as to sense the three directions of the human's intention: push/pull, horizontal and vertical motions. The interface that is embedded in the right arm enables us to communicate with the robot by shaking hands. In handshake communication, humans can express his/her mental intention in several ways: for example, holding kindly or strongly. During the handshake, a force emerges according to the difference between the motions of the hands. By applying a force to the other, one can lead the other. The following conditions are set for the intuitive understanding. (see also [Hikiji et al., 2000])

1) When a human grasps the interface, the robot responds by grasping back with the thumb. 2) When a human applies a force to the interface, the robot behaves so as to cancel the applied force by utilizing the right arm (2 DOFs), wrist (2 DOFs) and wheels (2 DOFs).

Figure 4.10 shows examples of the handshake communication.

(a) grasping back (t=0.0, 1.0 and 1.5 [sec])



(b) pulling by the hand (t=0.0, 1.0 and 2.0 [sec])

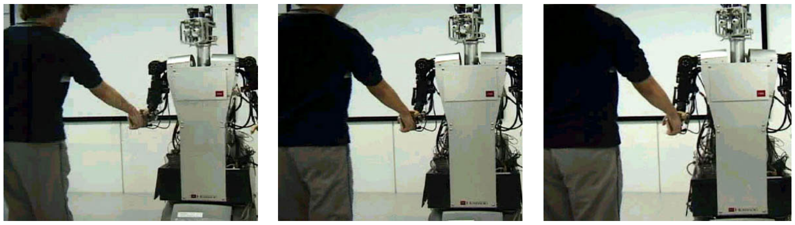


Figure 4.10. Hand shaking and the control: figures show an example of handshaking. It can be seen that the finger moves according to human grasping motion. The user can also push the robot by the hand although it is not natural human behavior.

4.4.3 Tracking and reaching an object

The robot tries to reach a recognized object through coordinated movement. The cooperative movement of each part, such as eyes, neck and wheels, is necessary for a robust and flexible object tracking. However, because the joint structure has redundancy, the specified path can be chosen under an arbitrary criterion. This experiment differs from the conventional binocular vision system [Jain et al., 1995] in utilizing the whole body including the eyes, neck, body and arms. In this experiment, each part is moved in co-operation with the others to track the object segmented by extracting color hue information through the binocular vision system. The tracking and algorithm are independently divided into two terms, head and arm movements.

Head movement:

The head contains 8 DOFs as illustrated in **Figure 4.11**. Each eye has 2 DOFs, horizontal (yaw) and vertical (pitch) axes. Seven of these are used for object tracking.

A small ball that has a specified color for extraction is used for this experiment. The coordinates of the center of the image of the recognized object are calculated. The distance to the recognized object can also be calculated by the azimuth difference. The accuracy is, however, not sufficient to detect the object in 3D space for real use. Therefore, only 2D image

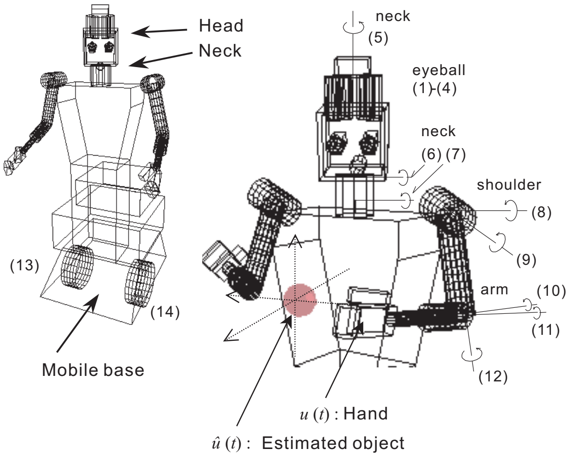


Figure 4.11. Object tracking system: the robot gazes at and tracks an object utilizing the head, the arm and the body with totally 14 DOFs.

coordinates of the object is used. As for the distance, the size of the object as the average of the object in the images obtained from both eyes is used as a substitute for the parallax. Based on these, the position of the object in 3D space can be estimated.

The velocity of response at each joint is different due to the difference in inertia for each part. Each eyeball moves faster than the neck, and the neck moves faster than the body. In general, in order to hold the object in the range of vision, not only eyeball motion but also the cooperative motions of the neck and body are necessary. Before the object goes out of sight, the robot head should follow the object.

The patterns of gazing at the object are thus itemized as follows: When a target is located in the central part of the camera image and a short distance from the robot, eyeball motion is induced due to the fast tracking property. When the target is located peripherally around the camera image and a short distance from the robot, neck motion results. When the target is located at a great distance from the robot, body motion is finally caused. In addition, when the target is located on the left/right side of each camera image, the priority is given to the left/right eye, respectively. The torque applied to each part is tuned so that it reflects the above characteristics.

Arm movement:

So far, many studies on the kinematics of a robot have been reported [Flash et al., 1985][Tevatia et al., 2000][Zatsiorsky, 1998]. As for arm control, a method of the combination of primitive motions is adopted. By using the estimated 3D position from the binocular vision system, the robot changes its arm posture incrementally so that the robot hand reaches the object.

Unlike deriving a locus to the desired position, the robot takes a posture for minimizing the mean-square error to the target incrementally at the local coordinates. This algorithm does not aim to acquire an optimal pathway; however, it contains adaptive and robust features. Even if a joint is broken or disabled, the robot can continue to assume a posture using other joints.

A robot achieves the reaching task with 4+1 DOFs joints as illustrated in **Figure 4.11** (joints (8)-(10), and (12)). The yaw axis of the lower arm (joint (11)) is used only as a conditioning direction for the palm of the hand. The position of the hand at time t can be estimated by the given kinematics. Each joint can be moved with the basis shifting defined as the minim shifting $\Delta\theta$. The prospective posture at time $t+1$ can then be obtained by the product of the minim shifting $\Delta\theta$ and the arbitrary derivative gain k_i. By repeating this operation, the robot arm approaches the object. The arm posture at time t, $u(t)$, is described as:

$$u(t) = f(\theta_1(t), \theta_2(t), \theta_3(t), \theta_4(t)) \tag{4.5}$$

The posture at time t+1 is delivered as:

$$u(t+1) = f \begin{bmatrix} \theta_1(t) + k_1\delta_1 \cdot \Delta_1\theta_1 \\ \theta_2(t) + k_2\delta_2 \cdot \Delta_2\theta_2 \\ \theta_3(t) + k_3\delta_3 \cdot \Delta_3\theta_3 \\ \theta_4(t) + k_4\delta_4 \cdot \Delta_4\theta_4 \end{bmatrix} \tag{4.6}$$

$$where \quad \delta = (-1, 0, 1)$$

where k_1, k_2, $\cdots$, k_4 denote a gain constant to each basis shifting, and δ is a bipolar step function at each joint i. The prospective posture is determined in order to minimize the following evaluation function E, aiming at moving the hand position close to the estimated position $\hat{u}(t)$ of the recognized object.

$$E = (\hat{u}(t) - u(t+1))^2 \tag{4.7}$$

With the use of a motion simulator system, the proposed algorithm is proved to reach the vicinity of the object.

The object position is not estimated precisely and the pathway obtained by the incremental reaching method is not proved to achieve the task in all cases. However, in the cases in which the target is located in front of or near the robot, the reaching movement can be successfully achieved.

Figure 4.12 shows camera images from both eyes, the result of segmentation of a recognized object and the data flow of the object recognition.

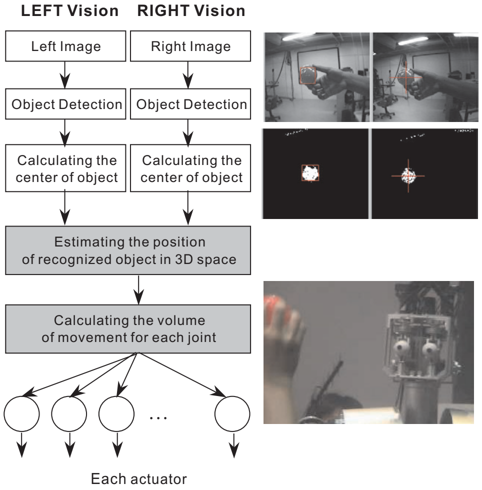

LEFT Vision
RIGHT Vision
Left Image
Right Image
Object Detection
Object Detection
Calculating the center of object
Calculating the center of object
Estimating the position of recognized object in 3D space
Calculating the volume of movement for each joint
…
Each actuator

Figure 4.12. Data flow of the tracking and reaching object: the upper-left figure illustrate images from CCD cameras embedded with eyes of the robot. The black-white images illustrate the segmentation result of a target object.

4.4.4 Dancing movements

The robot can perform a dance movement according to a given tempo. The tempo is defined as the rate of speed, motion or activity. For example, the tempo and beats can be extracted from music and sound. It is considered that the tempo is one of the parameters by which the robot and the external environment are synchronized. The pre-defined gestures are very simple; the robot dances by swinging his arms and head. By giving the size and frequency of motion of the arms and waggling of the head, the robot can dance with his whole body. The movement of both arms and head is synchronized with the given tempo. The position is given by a sinusoidal pattern with a frequency that corresponds to the tempo and an amplitude that corresponds to the robot's intention (currently a fixed value). The above dance movement with three joints for the neck and four joints for each arm is implemented in a behavior coordinator. The movement rule is described as follows.

$$x_i(t) = A_i sin(\theta_i + wt/\Theta_i) \tag{4.8}$$

Ai denotes the amplitude, and w denotes frequency, which corresponds to a given tempo of the movement at joint i. $x_i(t)$ represents the angle. Θ_i represents the predetermined range of movement at each joint i. The robot can perform a coordinated action by hardware scheduling with the body of 13 DOFs.

A tempo tracker have been implemented, which enables us to provide a tempo to the robot by handclaps. The tempo of the robot $T_{robot}(= 1/w)$

is determined according to the obtained tempo with the equipped stereo microphone. By tuning the timing, the robot can thus synchronize the given periodical signals.

4.4.5 Reaction to auditory sensing

The three microphones embedded in the head are used for sound localization. Moreover, the robot can receive some voice commands. One more microphone and stereo speakers are attached for speech recognition and synthesis. In the present system, the robot utters a voice command before executing the ordered task such as stop, start, forward, backward, left, right, tracking a ball and dancing.

4.4.6 Integrated types of behavior

Some integrated types of behavior are illustrated in **Figure 4.13** and **Figure 4.14**. The above-mentioned types of behavior appear in parallel or simultaneously, not in series. For example, the robot can distinguish between objects with different colors. People can ask the robot to gaze at an object with specified color by speech. In addition, when the robot failed to track the object by the head, humans can help the robot to find it by turning it using the hand, e.g.) handclaps can be used for drawing its attention. The important advantage of multimodal interaction is such an integrated type of behavior. The robot should be able to achieve a given task in various ways. The proposed system and architecture allow a number of types of behavior by means of several channels of communication such as vision, auditory and haptics. These channels make the robot behave more sophisticated and flexible.

4.5 Discussion

In this chapter, the specifications and architecture of the autonomous robot *iSHA* is introduced. The experimental results showed that the developed robot provides various channels of communication between humans and the robot.

As for the robotic architecture, a double-layered structure for physical and intellectual interactions is proposed. The physical layer is placed as the lowest one and has the highest priority. The communication ability of the robot with its external environment depends upon the physical constraint. The intellectual processing by computation is dependent upon and limited by that.

In addition, human-like modality is the other characteristic of the developed robot. Following a particular object with the binocular vision system and sound localization by an embedded stereo microphone are implemented.

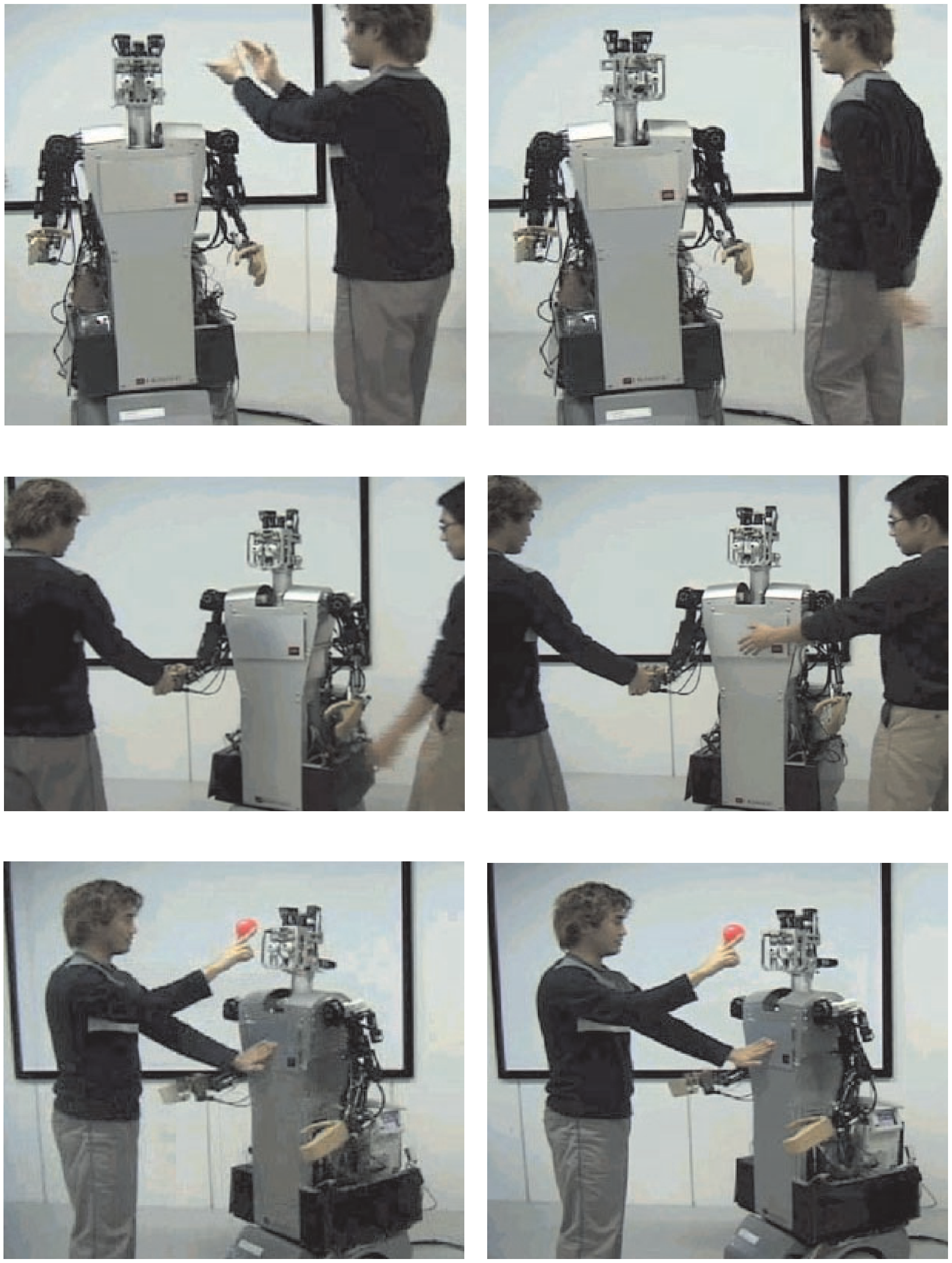


Figure 4.13. Multimodal interaction with *iSHA*: Various types of behavior appear in parallel or simultaneously, not in series. This integrated type of behavior is important advantage of multimodal interaction.

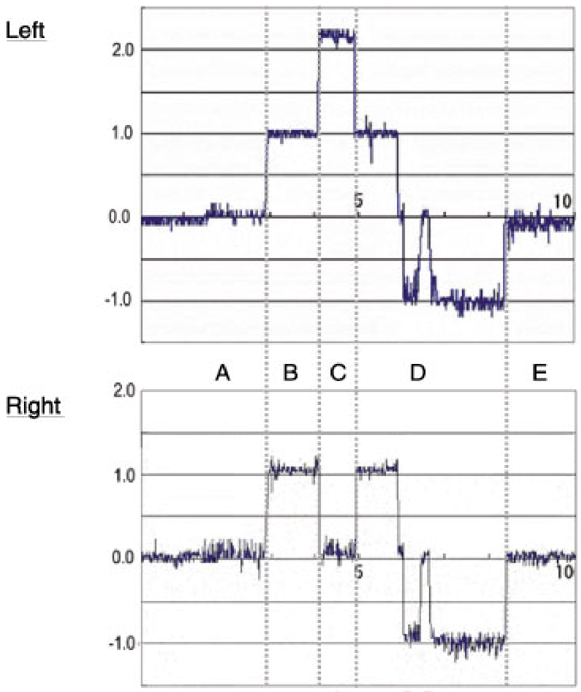


Figure 4.14. An example of the integrated type of behavior: physical interaction and intelligent interaction. Upper and lower figure respectively illustrates a voltage output for left and right vehicles.

Moreover, haptic and tactile interfaces for communication between humans and the robot have been introduced, which allow humans to perform embodied interaction. At present, although binary sensors are implemented for touch sensing, humans can handle the robot with their hands through physical interaction. The physical interaction layer plays a role in gathering the reflex actions of the robot. The developed robot thus provides various ways of sophisticated communication by combining physical and intellectual interactions.

The other focus is the robustness. The robotic architecture allows multi-process tasks, and the control modules are implemented independently. In addition, any instruments can be connected to each other with the network protocol. Therefore, the robot has the robustness of both the system and of behavior. Even if lost connections or machine problems occur in one of these instruments, the robot tries to behave in a possible alternative way. The proposed method of reaching an object is an example. Because the robot assumes a posture by the defined basis shifting of each joint, it can continue to follow the target even in the case where some joints are disabled, or the motion is disturbed by an external force.

In such a complicated robotic system, the scalability of the system is also an important discussed attribute. A system plug-in architecture preserving the advantages of behavior-based architecture is installed. The instruments included in the robotic system are connected with the common TCP/IP network protocol. For example, when a system module installed into the existing robotic system, it can be simply put in the network to exchange data via the network with/without cable. The constructed network system inside the robot can accept the connection from any optional devices such as the speech processing module, the image processing module, and a wireless connection from an external computer. For example, a robot simulator that runs on an external laptop computer can handle the robot over a wireless LAN connection.

As for collaborative work with humans, dancing movements are an example. The robot can dance according to a given tempo. The torque applied to each actuator in phase with a given tempo represents the synergic effect of the robot. By matching the phase between the external signals and the robot's internal cycle, a rhythmic synchronicity would take place. It is considered that a variety of input signals should influence the robot movements. The robot would synchronize its motion with the extracted periodic signal.

In the present work, the developed robot requires numbers of empirically predetermined parameters. Not only fundamental parameters such as the PID gain, but also parameters for intellectual interaction such as the range of reaching its object and the priority of the types of behavior. Tuning the parameters by learning is considered as one of the necessary abilities.

The processing manner is classified into two types of procedures, physical and intelligent processing. In the proposed architecture, each behavior coordinator received the sum of weighted stimuli from input channels. However, when one type of behavior is classified into a procedure, the role is then consequently established. The difficulty is how to associate a behavior type with the type of processing. Self-recognition and self-evaluation are the future research topics. The robot must build an awareness of itself.

In this chapter, the research direction is sketched for building a harmonized human-machine environment for a humanoid robot. The environment is not limited only to humanoid robots but can also be extended to general machines. In the near future, human-machine interaction will enter the next stage, human-humanoid interaction. Such human-cooperative robots will become an "emotion activator", which is regarded as a metaphor meaning that the robot can not only imitate human motions and gestures but also can stimulate humans with its movements and behavior.

Chapter 5

Kansei Interaction

Up to now, the Kansei communication have often been investigated, which is regarded to include some terms such as feeling and sensibility, between humans and machine. In communications between humans, the role of Kansei information is as important as logical information [Hashimoto, 1999]. Notice that such information is often processed not logically but unconsciously and involuntarily. Such characteristics, or rather types of behavior, might be regarded as Kansei. Humans have the ability to understand things by intuition. In other words, aside from logic, humans adopt information processing based on Kansei. Consequently, a robot that interacts with humans should deal with the Kansei information processing. This is one of the requirements for the harmonized human-machine environment.

Synthesis: Kansei-Oriented Communication The next chapter describes a substantial robotic interface for the realization of an interactive musical environment for collaborative work between humans and machines. The robot can be effectively used for musical performances with motion by exploitation of the embodiment. The "moving instrument" can display the refractive motion on stage while producing sound and music through embedded stereo speakers according to the context of the performance.

The three musical platforms utilizing robotic technology and information technology in different circumstances will be introduced. These are effectively designed environments for artists such as musicians, composers and choreographers not only for the music creation but also for the media coordination. The architecture called the MIDI Network enables them to control the robot movement as well as to compose music. Each works as a sort of reflector to create an acoustic and visual space with multi-modality. The proposed approach to equip musical instruments with an autonomous mobile ability is providing new computer music performances in the real world.

"Her Kansei is excellent."
"He is a man of rich Kansei."
"He has no Kansei."
"Her Kansei seems well suited to me."
"The Beatles expressed the Kansei of the times."
"The actress behaves with Kansei."
"This sound stimulates my Kansei."
"Polish up your Kansei to be an artist."
"My Kansei cannot accept his music."

Table 5.1. Typical examples in different contexts of Kansei [Hashimoto, 1997]: People have an ability to understand things by intuition. In other words, people adopt information processing based on Kansei, aside from logic.

Interactive multimodal environments are active spaces capable to observe users and to establish high-level communications with them by means of human gestures, movement, speech and singing. At the same time, such spaces allow users to get feedback in terms of visual media, sound and music.

A number of promising fields of application are emerging for such human-cooperative robots. These include interactive entertainment, interactive home theatre, interactive tools for aerobics and gymnastic, rehabilitation tools, tools for teaching by playing and experiencing in simulated environments, tutoring programs customized to students' different learning styles, tools for enhancing the communication about new products or ideas in conventions and "informative ateliers", cultural and museum applications, and computer-based games.

Analysis: Kansei Quantization On the other hand, another important approach for understanding the Kansei interaction is the measurement of Kansei. The mechanisms and the process of human perception have been widely discussed in many research fields. In the process of perception, one of the possible explanations is that physical features in a high-dimensional space perceived by humans are mapped onto another low-dimensional space in the human mind which contains semantic parameters.

The author first describes the statistical background for the measurement of Kansei, especially the related methods utilizing the similarity or dissimilarity between two objects for the evaluation. The proposed neural network model is then presented with some learning examples. As an example of applications for real world problems, the proposed model will be applied to the analysis of facial expression perception. The author also attempted the measurement of the individual subjectivity and its visualization. Furthermore, the performance on multiclass classification problems will be also described.

Chapter 6

Music-Based Human-Robot Interaction

6.1 Introduction

Sound and music are typical channel of non-verbal communication that humans often use to express their mind. Music is essentially a mode of expression of emotion and affection, and also a Kansei communication protocol among people. Kansei is a Japanese word that means something like "sensitivity", "intuitiveness" and "feeling". People have certain feel about the use of the word "Kansei", although the scale is not measured in a quantitative way, and not visible like the feelings of people [Hashimoto, 1999]. A musical performance is a Kansei communication to express individual ideas and thoughts with the aid of instruments. For increasing the degree of freedoms of a musical expression, a number of researchers and musicians have explored a new type of musical system that exceeds the physical limitation of musical instrument and vocal cords.

These are mainly divided into two categories. One is a new kind of musical instrument to enhance expressiveness by utilizing a computer system for controlling the sound and music effects based on conventional musical instruments. Live interactive electronic music by hyper instruments based on the organ, percussion and cello have been performed in the early stage of these studies [Machover et al., 1989][Chung et al., 1991]. More recently, Cook et al. [Cook et al., 2000] have developed an accordion that embeds a microcomputer for controlling the sound. A Japanese traditional instrument, the Sho, with attached sensors to sense the breathe of the player has been developed by Nagashima et al. [Nagashima, 1999].

Others have attempted to create music according to human gesture and entire body movement [Morita et al., 1991] [Camurri, 1995] [Siegel et al., 1999] [Camurri et al., 2000]. Various types of sensing techniques have been used to detect human motion, and the measured body movements are mapped to music or sound. One example is an instrument magnetic based motion capture system [Paradiso et al., 1999]. Sawada et al. [Sawada et al., 1999] introduced a ball-shaped instrument that can sense human grasping and the movement. In addition, many studies that contain the above two features were also reported [Katayose et al., 1993][Nakatsu, 1997][Paradiso, 1999].

As an example of an environment-oriented musical system, Rokeby has attempted to construct a space in which the movements of one's body create sound and/or music. The Very Nervous System [Rokeby, 1995] is an interactive sound installation utilizing multiple video cameras for music creation. It has been presented as an art installation in galleries, public outdoor spaces, and has been used in a number of performances. The Virtual Cage by Moller [Moller, 1997] is a floor for music creation, which can sense the weight shift of human on it, and can create acoustic feedback by sound and music. At the same time, the pneumatically controlled platform can move and give physical feedback to the user.

On the other hand, some attempts have been reported a mobile robot in order to connect the virtual and real musical worlds. Eto [Eto, 1998] produced a network based robotic art installation. The robots in the site for exhibition communicate with each other and create music. These robots are controlled by users at the site or via the internet. Wasserman et al. [Wasserman et al., 2000] reported a robotic interface for the composition system, RoBoser. It is a small mobile robot and is an autonomous real-world composition system by combining an algorithmic computer-based composition system.

The main purpose of this work is to investigate and explore the paradigms of the Kansei interaction between humans and a robot in the framework of museal exhibitions, theatre, music and art installations. Many studies about the musical interaction between humans and machines have already been proposed. However, although humans often accompany music with body motion, few studies have reported the autonomous mobile robot for a musical performance.

Consequently, focusing on an interaction metaphor that a robotic interface can establish a virtual and real world connection, some sensory systems and mobile robot platforms for the virtual musical environment in the real world are proposed. The robotic interface is one of the possible partners to interact with humans. The first idea was to equip a musical instrument with an autonomous mobile ability for computer music performance in the real world. The robot is effectively used for musical performances with motion because it can move around with the performer while generating sound and music according to the performer's movement and environmental sound and image. Moreover, the robot can display the refractive motion according to the context of the performance to create the human-robot collaborative performance on stage. Applications for interactive art, music, and edutain-

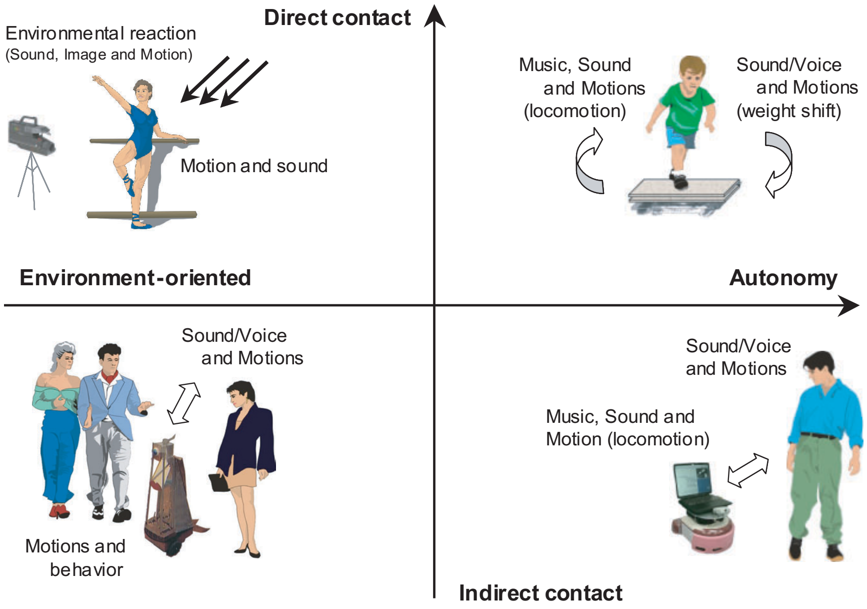


Figure 6.1. Modeling of Human-Machine-Environment: The style of the interaction can be modeled along two axes: the autonomy of the robot and direct/indirect contact. The region is largely divided into four areas according to the characteristics.

ment (education and entertainment) the rehabilitation (e.g., of autism) are promising fields of this study.

This chapter presents an interactive multimodal artistic environment for communication between humans and machines by introducing experimental systems. The developed system is a sort of active aid for a musical performance that allows the users to get feedback for emotional activation in terms of sound, music, image and motion.

The author first describes the modeling of the Human-Machine-Environment in music-based interaction in section 2. Three case studies will then be described. Sections 3 and 4 present two types of semi-autonomous robotic interfaces, Visitor Robot and the iDance platform. Section 5 describes an autonomous robotic interface, MIDItro. Finally, a discussion and conclusions will then be given.

6.2 Modeling of Human-Machine-Environment

In a multimodal musical environment, the style of interaction can be modeled along two axes: the autonomy of the robot and direct/indirect contact as illustrated in **Figure 6.1**. The description of each area is as follows:

1. (left-upper, the fourth quadrant) Reaction of robotic environment by

means of music/sound, image and motion is caused by the emitted sound and human behavior in the environment.

2. (left-lower, the third quadrant) Robot makes its own motion and displays music/sound and image from the surrounding environment, in accordance with environmental sound and image surrounding the robot.
3. (right-upper, the first quadrant) Through a direct contact between human and robot, the robot follows the applied force and also displays its own motion and creates sound and music according to the environmental sound and image, and data upon physical contact.
4. (right-lower, the second quadrant) Through an indirect contact between a human and robot, the robot autonomously displays its motion and creates sound and music according to the environmental sound and image.

In the left-upper field, the fourth quadrant, a model of an environmental robot is considered. An environment is "robotized," that means people interact with the surrounding environment where human motion and the omitted sound can be sensed by the wall, objects, room and the house. The surrounding environment then responds by generating sounds, images and also motions, e.g., the wall moves to people, or object transforms according to the stimuli. The author has contributed to the related work [Takahashi et al., 1999]. This model will not be described in this dissertation, but is indeed a future consideration.

In the framework of human-machine interactions, the assumption was made that humans do not need to wear any special on-body sensors. They are observed by on-board robot sensors and an attached video camera. A camera-based sensor system is developed to increase its input capabilities. This allows humans to carry freely on their tasks.

The following sections present three styles of music-based human-machine interactions with the experimental performance results.

6.3 Case study I: Visitor Robot

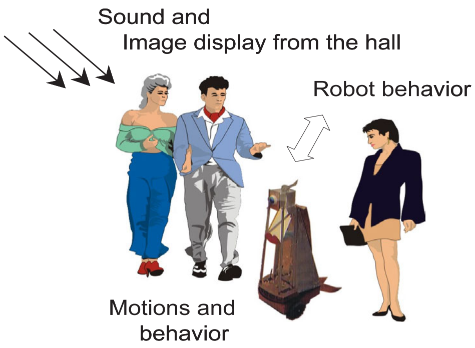


Figure 6.2. Case study I: Visitor Robot, that is a semi-autonomous robot for music-based human-machine interaction. [Suzuki et al., 1998]

Most of studies on the interaction between humans and robots have paid attention to understanding human biomechanics, sensing and control strategies so that the robot can give a performance similar to human beings [Hollerbach, 1996]. However, few have treated the emotional-based interaction of the robot itself.

Based on the above consideration, an intelligent agent for interaction between humans and robots embedding artificial emotions has been constructed. Using the developed system, the gestures of a human, scenes and sounds in the surrounding environment can cause the robot to change its emotional state, and therefore, its behaviors and reactions in the environment. With regard to this issue, some interesting experimental results have been reported [Nakatsu, 1997][Bates et al., 1992]. Many researchers have, however, paid attention to the analysis of movement, extraction of emotional parameters, while trying to realize the interaction through artificial emotion, which has the ability of self-organizing and adaptation being another scenario [Camurri et al., 1997].

The developed system is a robotic environment based on the agent architecture, which is useful software for multimedia applications in real time [Camurri et al., 1998]. The outputs are produced through three components: rational, emotional and reactive. Each component operates under the influence from the other modules. As a whole, they work to map the input parameters of the external world to the output ones.

A robot embeds a computational model of artificial emotion, which is constructed by taking advantage of the self-organization of an improved model of Kohonen's *Self-Organizing Map.* The network is adapted so that it can represent an emotional state; the current emotional state of the robot is determined as a result of competition with other states modeled by the network. Moreover, the state changes dynamically and represents the so-

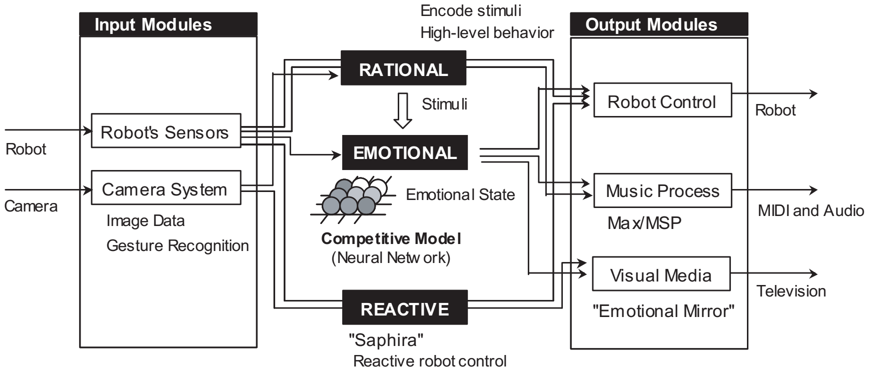


Figure 6.3. An emotional agent architecture *"Robotic Agent"* for multimodal environment: The details of the agent architecture is described in [Camurri et al., 1998]

called "personality" of the robotic agent.

It exhibits its current emotional state by means of integrated visual media, environmental lights, music, and changes in behaviors and style of movement. This allows the human high-level biofeedback effects to generate another motion by feeling.

The gesture, behavior and movement of a human can cause the robot to change its "emotional state", which is exhibited by means of integrated visual media, music, and changes in the robot's style of movement and behavior.

6.3.1 System overview

The small robot on wheels is a Pioneer 1 robotic platform by ActivMedia, Inc, that is equipped with a camera, infrared localization sensors, local audio system, and two wireless communication channels for both audio and video signals. The inputs of the robotic agent are given from the robot's low-level sensor data and a camera placed on the robot. As outputs, the system integrates three kinds of communication channels: movement (the behaviors of robot), visual (environmental light), and acoustic (music and sound).

Figure 6.3 shows the data flow detail in the developed emotional agent "Robotic Agent". The outputs are produced through three components: rational, emotional and reactive. Each module operates under the influence from the other modules. As a whole, they process input parameters from the external world and produce output parameters. It should be noted that only high-level information is processed through the rational and emotional modules. The *Rational* produces controlling data of the behavior of the robot and stimuli to the emotional module. On the other hand, the *Reactive* module produces parameters for dynamic output. The *Emotional* module is the core, which changes the current emotional state of the robotic agent.

Note that the stimuli produced by the *Rational* component are given to the *Emotional* component in order to drive the artificial emotion. The emotional state, which represents the personality of the robotic agent, then influences each communication module at the *Output* component.

The system for the robotic agent works under the Win32s operating systems. The robot communicates by three different radio links (digital I/O control data, video and audio signals) with the supervisor computer on which the model of artificial emotions is realized. The robot also possesses an on-board audio diffusion system, connected by radio, which integrates the audio diffusion system placed in the environment. Three computers connected by an Ethernet network control different various aspects; the first contains the emotional model and control movement, the second deals with "emotional mirrors", and the third generates sounds and music. The robot agent and the other applications are written in C++ (MS Visual C++).

6.3.2 Sensing external environment

Sensor input Input module mainly consists of two components. One is a receiver from the robot's sensor. Its role is gathering data from the robot's sensor every one hundred milliseconds. The following data are computed: the absolute and relative positions of a human and the robot, and distance and area when a human is around the robot. Logarithmic units as the distance have been adopted. The space around the robot is divided into five areas. Five ultrasound sensors are placed in front of the robot. The sensor data gathered by *Saphira* [Konolige et al., 1996] is processed by the Input Module. The special purpose robotic software *Saphira* [Konolige et al., 1996] has been adopted, that was developed by SRI (Stanford Research Institute) in order to handle the low-level details of the robot, such as drive motors and wheels. It is used for control of the Pioneer 1 robotic platform by ActivMedia, Inc. In other words, the component observes the environment around the physical relationship between the human and robot.

Gesture recognition system Another component is a camera-based sensor system, which allows human gestures as inputs of high-level information. This system allows a human to communicate with the robot with the aid of a small light source.

With the aid of a small camera on the robot, the current position of the light is detected every 100 milliseconds, and the current position of the light is sent to the robotic agent about twenty times per second. The detection starts when the user turns on the light in front of the robot, and ends when he turns it off. This duration is extracted as one phase. **Figure 6.4** shows the process of gesture recognition. Each gesture is normalized to the smallest rectangle from the center of the light positions. Therefore, it does not depend upon the area of the human's gesture. The detected area is then

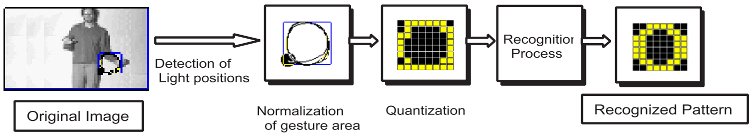


Figure 6.4. Process of gesture recognition process: A camera-based sensor system allows human gestures as inputs of high-level information. The system allows a human to communicate with the robot with the aid of a small light source.

quantified into an 8x8 image data. This image is used as input data for the recognition processes. Moreover, the agent extracts several parameters as input data, not only the pattern of gesture, but also the size of the detected area and duration of one phase.

As the pattern recognition of gesture, a low level processing and a simple back propagation neural network is used. In the present study, it is also possible that the data is sent to the robotic agent as ”negative” and ”positive” information. For experiments, ten gestures seem to be sufficient for communication with the robot. The circle-gesture, for example, can mean a positive stimulus, while a slash-gesture can mean a negative one. The time of recognition is less than one second, which is effective for real-time interaction.

6.3.3 Robot reaction

Robot control The output module consists of three components. One is a component to control the robot.

Two types of control are prepared: behaviors and movements. The former means a high-level behavior so that robot might produce performance similar to the human. In short, it seems that each behavior has a reason such as following a human or escaping, attention or avoidance, turning around a human. On the other hand, the latter corresponds to quite simple movements: forward, backward and turns. These types of behavior and movement patterns are controlled by orders from the Rational component.

Music generation The component of the music process is discussed here. Music is one of the most important ways of communication. In this system, the application to ”modulate” the score skeleton music with MIDI is adopted. The application is described by the Max/MSP patch. The agent outputs particular data to the application so that the generated music can reflect the emotional state. In the component of music generation, the data are arranged so that it can reflect the emotional state and style of movement of the robot. The robotic agent generates parameters used to modulate the score skeleton in each time slot. They include not only the emotional state but also the movement of the robot through its ”Reactive” module. They consist of emotional parameters and physical relations between the human and robot such as distance and area. For example, the music volume is changed by the physical distance obtained at the Input component between the robot and human. In addition, when the robotic agent receives gesture information, parameters are also sent to the music process thus influencing the score skeleton. The main four patterns of music that correspond to each emotional state are prepared. Each music pattern is composed so that the impression would reflect each emotional state.

Visual Media - The "emotional mirror" A further output is the control of visual media. It is available to show the emotional state of the robot, and also the user can see if the robotic agent detects the light carried with his hand. It is expected that performers can understand the state of a robot more clearly with different output channels. The visual component, as another example, is based on the idea of an "emotional mirror". **Figure 6.5** shows an example of the aspect and the implementation of the emotional mirror. The robot sees what is in front of it (people's faces, artworks) warped according to its emotional state. For example, a face could appear "mirrored", distorted in a vortex or re-processed with bright colors, respectively corresponding to positive and negative emotional states. During the performance, the system shows such images on a TV screen in real time.

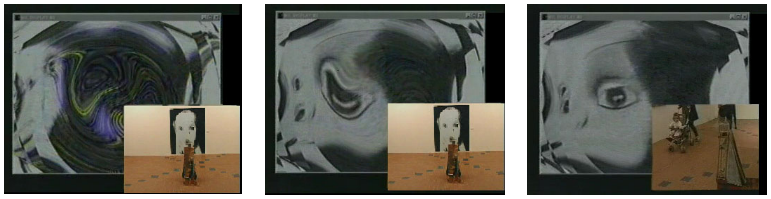


Figure 6.5. Dynamics of Emotional Mirror: What the robot sees appears "mirrored", corresponding to the positive and negative emotional states. During the performance, the system shows such images on a TV screen in real-time.

6.3.4 A model of artificial emotion

One of the motivations to construct artificial emotion is the complexity while making decisions in the robotic agent. The agent receives many inputs from the external environment, such as its position to the human and gestures. In addition, the agent should also refer to the internal state. In other words, artificial emotion is one of the means that supports the agent making decisions dynamically and flexibly. In the real world, the robotic agent divides input into only four vectors for simplicity. The artificial emotion model then consists of four states, which form the personality of the robotic agent. This structure is called "emotional space" as shown in **Figure 6.6**.

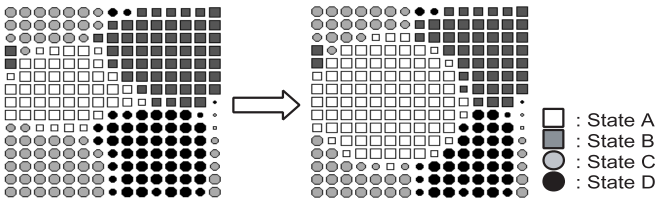


Figure 6.6. The change in the emotional state after a stimulus is given: The model is inspired by the dynamics of human emotion.

The model is divided into four areas that represent the emotional state, and each symbol corresponds to a particular state of emotion. Each state with a typical human emotional condition is called *Happy*, *Angry*, *Melancholy* or *Tranquil* for simplicity. The number of each symbol (cell) represents the rate in emotional space, and each state represents a unique character of the robot. In other words, it represents the personality of the robotic agent. The emotions compete with each other. The area occupied by each state then shows a rate in the emotional space of the robot. When a state changes in emotional space, the other states are also influenced by each other. For instance, when a state changes becomes wide, the others should change to narrow through competition in the network. It should be noted that each state is always competitive in this model. The feature of self-organizing is applied to the artificial emotion of the robot. Considering changes in the emotional state as a result of the competition with different emotional states, an emotional model based on Kohonen's Self-Organization Map (SOM) [Kohonen, 1994] is constructed. The network is improved so that it is suitable for dynamic changes in the emotional state. As described in [Ahalt et al., 1990], the torus model of SOM is used in the present study.

Self-Organization Processing The network is improved so that it is suitable for dynamical changes of the emotional state. As described in [Ahalt et al., 1990], the torus model of SOM is used.

Kohonen's algorithm is established in an unsupervised way. Using the neural network, n-dimensional input space is mapped onto m-dimensional lattice space A. Each unit has a weight vector $\boldsymbol{w}$ of n-dimensions, $w_i = [w_{i1}, w_{i2}, ..., w_{in}]$, and is assigned to each input vector $\boldsymbol{v}$, $\boldsymbol{v}_i = [\boldsymbol{v}_{i_1}, \boldsymbol{v}_{i_2}, ..., \boldsymbol{v}_{i_n}]$. The mapping is formed so that the weight w of active unit is closest to the current input vector $\boldsymbol{v}$ in the nearest neighbor rule as follow.

$$||\boldsymbol{w}_i - \boldsymbol{v}|| \geq ||\boldsymbol{w}_i{}^* - \boldsymbol{v}|| \qquad \forall i \in A \tag{6.1}$$

The network is tuned in a learning step according to the following rule.

$$\Delta \boldsymbol{w}_i{}^{(t)} = \varepsilon \Lambda_{i,i^*}{}^{(t)} (\boldsymbol{v}^{(t)} - \boldsymbol{w}_i^{(t)}) \qquad \forall i \in A \tag{6.2}$$

In the neighborhood function Λ between the nearness of cells i^*, the strength parameter ρ is added so that the network can work dynamically. Λ is determined as:

$$\Lambda_{i,i^*}{}^{(t)} = exp\left(-\rho\frac{||i^{(t)} - i^{*(t)}||_A}{2\sigma^2}\right) \tag{6.3}$$

ε is the constant learning parameter, and $||\cdot||$ denotes the Euclidean distance in A. The neighborhood range is determined by the selection of radius σ. Additionally, the strength parameter ρ is added into the function as continuous establishing the network. The strength parameter reflects a sort of given stimulus to the network. If a human, for example, plays the same gesture wide and fast, the stimulus should be strong, or narrow and slow, it should be smaller. The connection zone $Z(t)$ is also available. As well as the strength parameter ρ, $Z(t)$ is determined every learning process regardless of global time. From this, the rule of renewal weights is given as eq. (6.3).

$$\Delta\boldsymbol{w}_i{}^{(t)} = \begin{cases} \varepsilon\Lambda_{i,i^*}{}^{(t)}(\boldsymbol{v}^{(t)} - \boldsymbol{w}_i^{(t)}) & \forall i \in Z^*(t) \\ \boldsymbol{w}_i^{(t)} & \forall i \notin Z^*(t) \end{cases} \tag{6.4}$$

Network Structure and Dynamics The architecture of the modified self-organization map consists of 15x15 cells in two-dimensions with a torus structure, which means the upper and lower sides of the network, as well as the left and right sides, are connected to each other. Each cell has dimensions of emotional states. Here the input space is four dimensions. When a human gives information to the robot, the agent can understand it by its strength. The larger this parameter, the stronger the influence toward the network (see the details in [Suzuki et al., 1998]).

The kinds of symbols represent the emotional state. Each symbol also represents the activation of a neuron in the unit. Its size represents the amount of activation in each cell. In the center of the occupied area by each state, the size is almost large. However, in parts of the borders, they are always smaller than the others. Therefore, in parts of the border, each state is competitive with each other. This model is inspired by a dynamics of human emotion.

The change in the emotional state is shown in **Figure 6.6**. This figure shows the emotional state after state A has increased. Comparing the left and right figures, it can be seen that the occupied area by state A increased, while, the areas of the other states became smaller than before.

Once the robot is put into multimodal environments, the emotional state begins to be change based on the external world. Through the proposed model of artificial emotion, the state dynamically changes according to the input data.

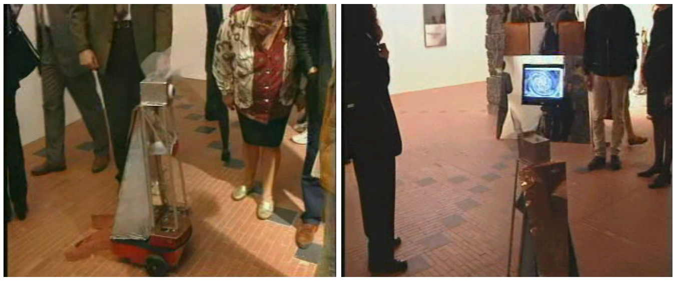


Figure 6.7. Case study I: Visitor Robot that performs at a museal exhibition "*Arti Visive 2 (Visual Arts 2)* ", Palazzo Ducale, Genova in October 1998.

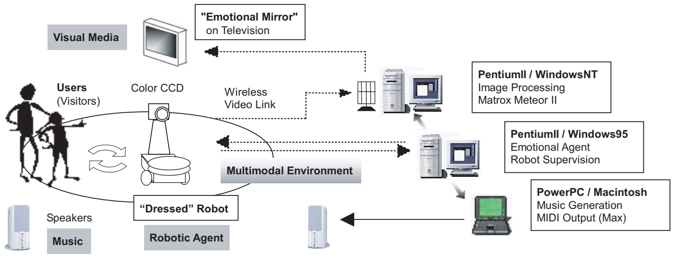


Figure 6.8. The supervisor system at "*Arti Visive 2*" (*Visual Art 2*), which is a museal exhibition held in *Palazzo Ducale*, Genova, Italy, in October 1998.

6.3.5 Performance demonstration

The developed system has been demonstrated in the interactive art installation at "Arti Visive 2", a museal exhibition held in Palazzo Ducale, Genova, in October 1998. The robot freely tours in the exhibition as one of the many people who frequent it, a sort of medium between humans and machines living together in the exhibition area. Sensors allow him not only to avoid collisions with people surrounding him, but also to observe the artworks and the visitors in order to interact with them. In the latter presentation, the robot also has been "dressed" with a scenography for the art installation. On the top of the dressed robot, a small camera has been installed.

Visitors can communicate with the robot in several ways. For example, they can approach it, act in front of its "eyes", follow it, ignore it, or become an obstacle in its path. The robot interprets some stimuli as positive, other as negative causing the evolution of its emotional state. As described below, the emotional state is exhibited by means of the robot's movement, music, sound, and visual media.

6.4 Case study II: the iDance Platform

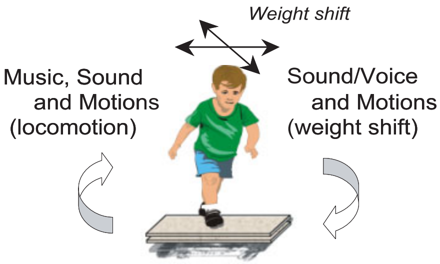

Weight shift
Music, Sound and Motions (locomotion)
Sound/Voice and Motions (weight shift)

Figure 6.9. Case study II: the iDance platform [Suzuki et al., 1999] that is a semi-autonomous robot for music-based human-machine interaction.

A semi-autonomous human cooperative robot will be described in this section. Through direct contact between a human and the robot, the robot follows the applied force and also displays its own motion and creates sound and music according to the environmental sound and image, and data upon physical contact.

The system software is based on an agent architecture for real-time multimodal interactions. The user can easily associate the agents on Max/MSP GUI. The module of sound analysis mainly extracts the pitch and velocity of the input sound including the human voice and instrumental sounds every 100ms. The module of image analysis extracts the color composition and temporal structure of the input image including the environmental scene and human gestures at the same rate of sound analysis. The sound, image and robot motion are compiled from the composition patch referring to the outputs of the behavior coordinator. The input data from the four strain-gage sensors, and sound and image analysis module are mapped into stimuli to activate the internal process components of the agent. These components produce the robot behavior including the displayed image and MIDI sounds, which can be influenced by the performer's movement, environmental sounds and images. It should be noted that communication data are exchanged through the MIDI channel among the main controller and others.

6.4.1 System overview

The overall of the developed system is shown in **Figure 6.10**. The system consists of four components: mobile robot, motion interface, main controller and output devices. In this study, an omni-directional mobile platform [Hirose et al., 1993] is used for a mobile base. In addition, a motion interface called "plate" is installed in order to receive external force information. The interface "plate" enables simple locomotion by a weight shift

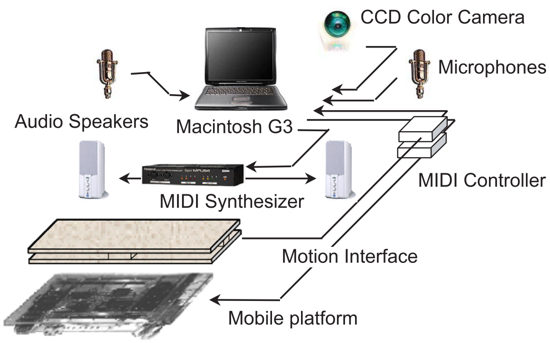


Figure 6.10. System overview of the iDance platform: To make the interaction with human, the system integrates three kinds of communication channels: acoustic (music and sound), movement (the behaviors of robot), and visual.

and force application [Yokono et al., 1998]. Also, a CCD camera and two microphones are installed in order to get environmental visual and auditory information, All these instruments and others including a Macintosh G3 computer and audio speakers are installed to make the mobile robot semi-autonomous. A number of useful modules for motion devices have added to the Max/MSP architecture. The modules communicate with the robot and motion interface through the MIDI controller to exchange serial and MIDI data. Therefore, this provides a effective musical platform where users can easily associate with each other including not only music generation but also movement of the mobile robot.

6.4.2 Sensing external environment

To have interactions with a human, the system integrates three kinds of communication channels: acoustic (music and sound), movement (the behaviors of robot), and visual. First, the input component consisting of these three modules will be described.

Motion Interface Module The first module is an action receiver that has the role of gathering data from the motion interface at the MIDI rate (31.25kbps). The interface can obtain the center of gravity data of objects on the mobile robot, which is measured by four strain-gage sensors bonded under the plate. If a user provides a force to the object on the plate, the four strain-gage sensors bonded under the plate measure the center of gravity. Since the strain-gage element changes its resistance value according to the applied load, the applied load can be measured as a change in voltage, and the center of gravity calculated using these voltage values by a single chip computer. The module on the Max/MSP patch receives a data list about the center of gravity, and calculates the desired direction.

Sound Input Module The second one is a part of the sound input. This module can obtain auditory information with two microphones installed on both sides of the mobile platform. From this, the system allows users to interact with the robot by using his voice and handclap. The following three sub-modules were developed for the Max/MSP patch.

Volume and Pitch Tracking: The volume and pitch data of the obtained sound from the installed microphone are calculated. The component called the *Sound Analyzer* works to extract the sound features and auditory information of the environmental sound. The input source of the object is the sound wave from the microphone that comes equipped with the standard Macintosh MIC-in. It outputs the following sound features:

Sound features
  (i) Velocity
  (ii) Fundamental frequency
Auditory information
  (iii) Environmental state

The sound from each microphone is measured using eq. (6.5), where N and A denote the number of samples per frame and the maximum amplitude, respectively.

$$V[db] = 10\ log_{10}\left(\frac{1}{N}\sum_{n=0}^{N-1} x^2\left(t+n\Delta t\right)\bigg/\frac{A^2}{2}\right) \tag{6.5}$$

The cepstrum method is adopted for the identification of the fundamental frequency. The method uses the harmonic structure obtained by Fourier transform at the high range of the cepstrums. In order to decrease sampling errors, the fundamental frequency is regarded as a value of the nth peak of the frequency divided by n as shown in eq. (6.6).

$$f_1 \approx \sum_{n=1}^{p} f_n \bigg/ \sum_{n=1}^{p} n \tag{6.6}$$

In addition, the system can also recognize the state of the environment from the auditory information. an experimental thresholding of the sound velocity and spectrum density were determined in order to distinguish each mode of auditory information such as noisy, silent and singing.

A Max/MSP object for extracting the specified pitch information of the environmental sound including a human voice is constructed using this method. As well as handclaps, a singing voice can be directed toward the robot as auditory information. For example, when the system obtains human's voice and is able to capture the pitch information, the backing scale

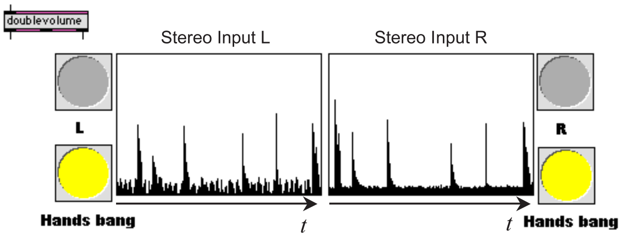


Figure 6.11. Tempo tracking object on Max/MSP: While users clap hands, this sub-module calculates the tempo by extracting the peak of the volume data.

will be changed to allow the users to control the high/low chord with his voice.

In addition, simple sound localization has been also realized. By using the difference in amplitudes from each microphone, the system can roughly estimate the location of the sound sources. It is not easy to detect exact sound sources with only two microphones. However, because the robot can turn toward the measured target, it helps to capture the exact sound sources more precisely.

Tempo Tracking: While users clap hands, this sub-module calculates the tempo by extracting the peak of the volume data. The system can synchronize the generated music with the estimated tempo in real time. The player usually has a flicker error to produce a tempo. The module renews the next tempo in order to take the flicker into account by the experimental threshold so that audience can easily listen to the music. **Figure 6.11** shows an example of tempo tracking with two microphones. Since the sound is captured every 10 ms, the time difference between each microphone is not found in this experiment. In **Figure 6.11**, the left figure represents the sound input from a microphone equipped with the left side, while the right figure represents the sound input from the right one. The x-axis of each graph represents time t, while the y-axis represents the volume of the input sound source.

Camera-based Sensor System The third one is a camera-based sensor system to obtain environmental information and human gestures. Moving image data from the CCD camera are calculated in order to get color information such as RGB data, hue, saturation and lightness every 100ms. From this, the spatial and temporal features are also obtained such as the density of the edge, pattern data, and blinking information. The system allows a human to communicate with the robot with the aid of a small symbolic source such as an LED light or color flags. From this, users can provide a sign to the robot with the location of detected symbol.

This component is called the *Image Analyzer* can extract temporal and spatial features [Gong et al., 1995] from moving images. The input source of the object is a moving image from the video camera, while the output consists of the following image features:

Color information

(i) RGB (Red, Green, Blue) components

(ii) Hue, saturation, and lightness components

Spatial and temporal features

(i) Edge density

(ii) Scene changing value (binary data)

In the present study, the sized of the captured frames are 320×240 pixels. Each frame is divided into $M \times N$ areas. The features of the moving image are calculated in both the whole frame and the image of each area to obtain the global features and the local ones.

The values of RGB components obtained by the image data in each frame is calculated as follows:

$$\begin{cases} L_k(i,j,t) = \sum_{x_i}^{n} \sum_{y_i}^{m} l_k(x_i, y_j, t) \\ L_{k(all)}(t) = \sum_{i=1}^{M} \sum_{i=1}^{N} L_k(i,j,t) \end{cases} \tag{6.7}$$
$$(k = R, G, B, H, S, L)$$

where L denotes the summation of each RGB brightness, and l_k represents each one of the pixels. In the same manner, the calculations of hue, saturation and lightness can be performed. m and n represent the width and height of each divided area, respectively.

By using this color information, the average values of the edge density are also extracted as the spatial features in eq. (6.8) by a two-dimensional filter as: .

$$\begin{cases} E_k(i,j,t) = \sum_{x_i}^{m-1} \sum_{y_i}^{n-1} \begin{pmatrix} -1 & 0 & -1 \\ 0 & 4 & 0 \\ -1 & 0 & -1 \end{pmatrix} \cdot l_k(x_i, y_j, t) \\ E_{k(all)}(t) = \sum_{i=1}^{M} \sum_{i=1}^{N} E_k(i,j,t) \end{cases} \tag{6.8}$$
$$(k = R, G, B)$$

where E_k denotes the summation of each edge density.

On the other hand, for the acquirement of temporal features, the image analyzer stores basic frame data as a background image at the beginning of

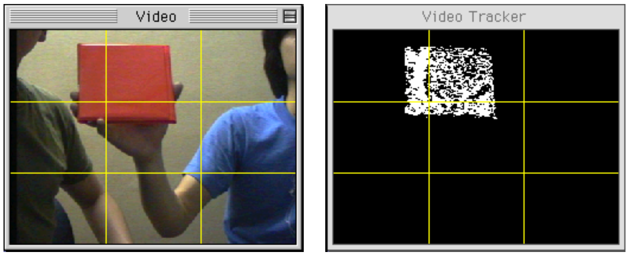


Figure 6.12. The Image Analyzer object on Max/MSP: Moving image data from CCD camera are calculated to get color information such as RGB data, hue, saturation and lightness every 100ms.

the detection of the moving image. By comparing to the background image, scene changing can be detected by calculating the temporal difference for every newest frame (eq. (6.9)).

$$\begin{cases} D_k(i,j,t) = \displaystyle\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\Big| l_k(t)(x_i,y_j) - l_k(0)(x_i,y_j) \Big| \\ D_{all}(t) = \displaystyle\sum_{k=R,G,B}\ \sum_{i=1}^{M}\sum_{i=1}^{N} D_k(i,j,t) \end{cases} \tag{6.9}$$

where D denotes the difference in brightness and $l_k(0)$ and $l_k(t)$ represent the brightness of the background and present image, respectively. The threshold θ represents the horizontal line in the figure. When D_{all} exceeds the threshold, a scene change has occurred.

6.4.3 Robot reaction

The output module consists of two parts: sound and music generation, and control movement of robot. Each output module operates under an influence from the input modules. They calculate the output parameters from the input parameters of the external environmental information through two kinds of process component: long-term and short term reactive ones. The details of the music creation are described as follows and the data flow is shown in **Figure 6.13**.

Omni-Directional Mobile Robot Control The first output module is the part that controls the robot. At present, two types of controls, active and passive reaction, are prepared.

The former one works as a sort of robot behavior to chase humans. For example, with the aid of symbolic flags, a robot can detect a human and simply follow the symbolic flag. This chasing reaction also is caused by the

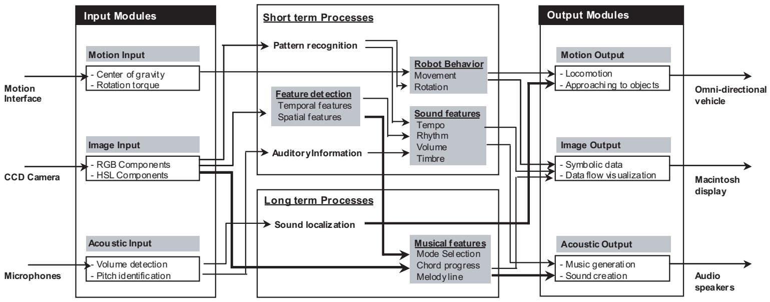


Figure 6.13. Data flow of the iDance platform: Based on the prototype of generated music, sound and musical features can be modified according to the input from sound, image and motion modules.

input from the sound detection module. By using sound localization data, the robot can turn and change the direction by itself. The latter one is a kind of tool that a human can use to control the robot. When he pushes the robot by his hand, the robot moves itself along the direction that the force is given. In other words, this module allows a human to show his intention with his action. The communication data format is the same as the MIDI configuration, and sent to the external MIDI Controller (Motion MIDI). The special hardware translates the MIDI format into the control of the mobile robot.

Music and Sound Creation The second one is a part of the music and sound generation. Some basic modes of music generation are prepared. Based on the prototype of generated music, sound and musical features can be modified according to the input from the sound, image and motion modules.

By using the key information such as scene changing and applied force, the system changes the current mode of the music generation. A few examples of music generation mode are described as follows.

Rule-based Music: We humans are familiar with the created music based on musical theory. Also in the field of computer music, the structure of chord progress and melody harmonizing has been often applied from many kinds of musical theory. In this mode, the theoretical music generation is simply adopted in order to make the simple chord progress. Using five typical patterns based on the C chord, a basic melody is also composed within the possible notes for each code.

Stochastic Music: Another kind of music generation, namely the stochastic music mode is installed. Up to now, several research studies about

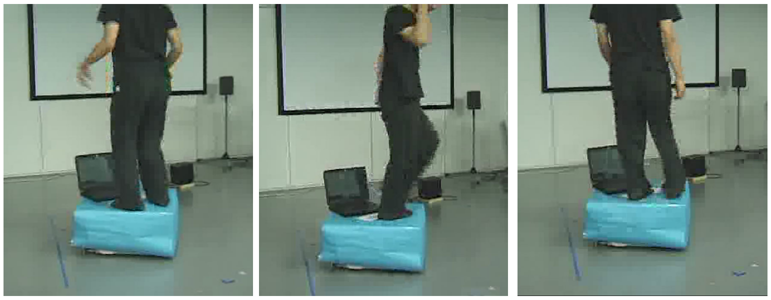


Figure 6.14. Performance of the iDance platform: Figures show an example of dance performance with the iDance platform.

stochastic music generation have been reported. For example, a historical work is [Hiller et al., 1959][Grubb et al., 1997]. In this study, every beginning phrase of starting this mode, the note set is determined by the input data from the sound/image analysis. The chord progress and melody can be created with a random value within the note sets. The prepared four kinds of note sets are: major / minor pentatonic, and Japanese major / minor scale. Pre-recorded Music: 72 sets of drum patterns are prepared: These patterns have 6 different tempos and appear so that the rhythm could correspond to the change in the image and sound input. The backing chord progress is also generated with the tempo based on musical theory. Some other MIDI files of melody are also used.

The musical features such as timber, pitch, volume, tempo and style of music generation are modified by the users in real time. The mobile robot becomes active when it is put into the environment where the robot and humans perform. All of the output can then be continuously changed, and also be modified according to both acoustic and visual features of the environment.

6.4.4 Performance demonstration

The performer freely produces action on the platform, which continues to create sound and music according to a given stimuli. The experiments prove us that the system is an effective and interesting interaction between users and robot by using multimodal communication channels.

6.5 Case study III: MIDItro

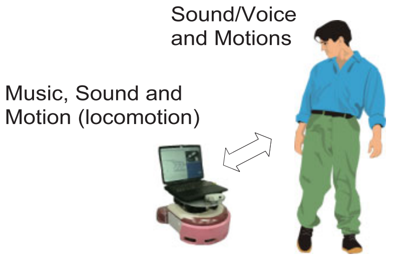


Figure 6.15. Case study III: MIDItro, an autonomous robot for music-based human-machine interaction. [Suzuki et al., 2000]

Finally, an autonomous robot will be described in this section. Through an indirect contact between a human and the robot, the robot autonomously displays its motion and creates sound and music according to environmental sounds and images, and physical contact data.

The system hardware includes two microphones, two speakers, a CCD camera, and Macintosh G3 computer, all of which are installed on the mobile robot. The software developed on Max/MSP environment consists of the sound analyzer, image analyzer and behavior coordinator with the sensor and motor drivers. The sound analyzer mainly extracts the pitch and velocity of the input sound including the human voice and instrumental sound at the frame rate every 30ms. The image analyzer extracts the color composition and temporal structure of the input image including the environmental scene and human gestures 10 times per second. The sound, image and robot motion are compiled according to the composition patch referring the outputs of the behavior coordinator.

The system can exhibit a style ”human-robot dance collaboration” where the robot moves in concert with the human performer sensing the visual and audio information.

6.5.1 System overview

An overview of the developed system installed in a two-wheeled mobile robot is shown in the left figure of **Figure 6.16**. The system consists of four components: a mobile base, main controller, and input and output devices. In this study, an omni-directional mobile robot is used. A color CCD camera and two microphones were installed in order to obtain environmental visual and auditory information. All these instruments and others including the Macintosh G3 computer and audio speakers are installed to make the mobile robot autonomous. A hardware connector is constructed so as to adapt the mobile base on Max architecture. The modules communicate with the robot

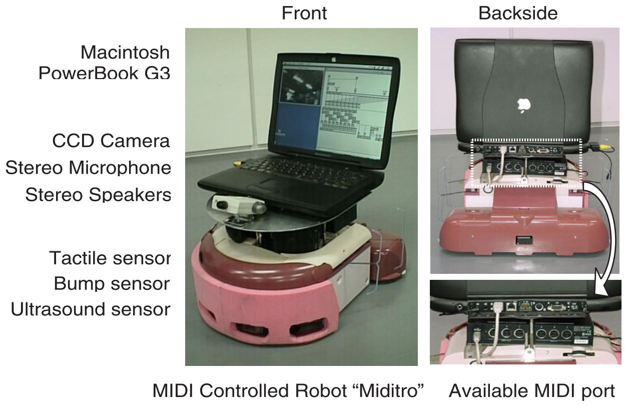


Figure 6.16. Case study III: MIDItro, system overview. All the equipment is installed on an omni-directional mobile base.

and main controller through MIDI controller in order to communicate with the robot.

The robot consists of two components; an upper installed part, and a lower mobile part. The CCD cameras and two microphones for stereo input are attached in front of the upper part on the robot. This robot has bumper sensors on the front and the back, tactile sensors on the sides, ultrasonic sensors on the front, and an encoder, a gyro sensor in the lower mobile part. Therefore, the location of the robot can be roughly calculated using these sensing data. The cushion on the bumper reduces the shock from any collision. A single chip computer is embedded in the robot in order to handle the low-level control, drive motors and wheels. For the whole system, the size of the robot is approximately 40×45×20 (width×length×height(cm)).

6.5.2 Sensing external environment

To modify the musical parameters such as melody, backing, tempo and pitch, three types of information as described below are available to obtain the sensing parameters.

Sensor Information The robot has three kinds of different sensors. The bumper sensor installed on front and back can sense obstacle contact in seven different directions. The dour-wire tactile sensors installed on both sides of the robot can sense an applied force. The ultrasonic sensors can measure the distance to obstacles in front of the robot in four directions in the range of 50mm to 500mm. The sensor configuration is illustrated in **Figure 6.18**. These three sensors are used not only to directly link musical parameters but also to avoid obstacles without transmitting MIDI data. For safety, in case when a bumper sensor sensed obstacles, the robot once stops to retrace its

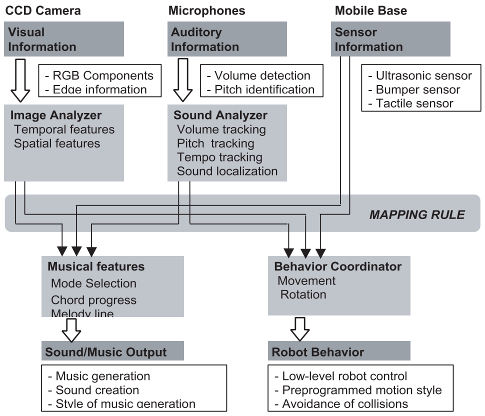


Figure 6.17. Processing flow of MIDItro: The mapping rule between input and output modules is determined by composers and choreographers. The output module consists of two parts for the sound and music generation and the control of the robot movement. Each output module operates under the influence of environmental data from the input modules.

steps. The other two sensors can be useful for avoiding collisions in advance. These sensing data are converted by the MIDI controller and transmitted to the main controller when the events occur.

Visual Information This module is a camera-based sensor system to obtain environmental visual information and human gestures. By using a modified Image Analyzer as described in section 5, moving image data from the CCD camera are calculated to get the spatial and temporal features such as RGB data, hue, saturation and lightness, the density of the edge, pattern data, and blinking information every 100ms.

Auditory Information This module can obtain auditory information with two microphones installed on both sides of the mobile platform. From this, the system allows users to interact with the robot using his voice and handclap. The following three sub-modules were developed on the Max patch.

The first is a sub-module of simple sound localization. By comparing the volume of sounds from each microphone, the system can roughly estimate the location of the sound sources. It is not easy to detect exact sound sources with only two microphones. However, the robot can turn toward the measured target to more precisely capture the exact sound sources.

The second is a sub-module of pitch tracking by the Sound Analyzer. The last sub-module has the role of tracking tempo. More details on the detection of the visual and auditory information can be found in section 6.4.2.

Robot movement	MIDI Code	Primary Factor
Basic movement		Bumper Sensor
Forward	B-01	Contact at backside
Backward	B-02	Contact at front
Sliding		Tactile Sensor
left direction	B-03	Contact at right part
right direction	B-04	Contact at left part
Rotate		Sound Localization
clockwise	B-05	Source located at right
counterclockwise	B-06	Source located at left
Zigzag motion	B-07	Randomly occurred
Circle motion	B-08	Scene Change

Table 6.1. An example of predetermined set of primary factors for robot movement: Sensory data of the robot is converted into the MIDI format by a special component, the MIDI Converter.

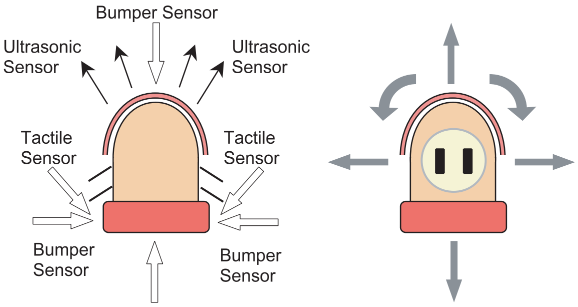


Figure 6.18. Sensors equipped with MIDItro. There are three kinds of sensing devices such as bumper sensor (binary signal), ultrasonic sensor for the distance measurement, and tactile sensor for sensing data upon physical contact.

6.5.3 Robot reaction

Mobile Base Control The robot can move omni-directional by rotating the lower mobile part against the upper installed part about the center axis of the robot, and by independently controlling two driving wheels. In this study, the robot is limited to six types of movements such as forward/backward, left/right sliding, and rotate-clockwise/counter-clockwise. The maximum speed is set to approximately 30 cm/s. For the experiment, the primary factors that cause each style of robot motion are predetermined, as described in **Table 6.1**. The MIDI code shows the defined addresses for transmission to the computer.

Also, some experiments with regard to direct robot control by the MIDI keyboard have been done. In this proposed platform, any additional musical instruments that can transmit MIDI data are possible instruments to be associated with.

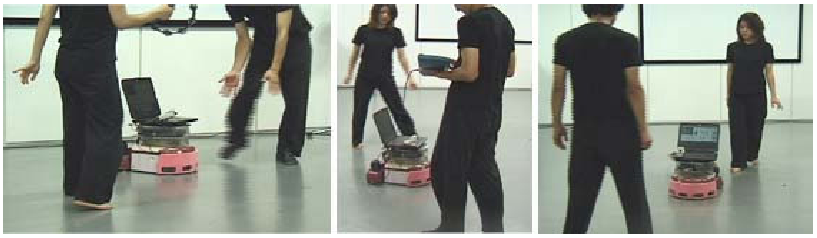


Figure 6.19. Performance of MIDItro: The left figure illustrates an example of interaction based on the surrounding sounds. In the middle one, the robot is controlled by MIDI keyboard with the aid of MIDI Network. The right one also illustrates a vision-based interaction.

Mapping from input parameters to music By mapping from these parameters to music, the music is created so as to reflect the input parameters from the external environmental information through two kinds of process components; long-term deliberate and short-term reactive processes. In part of the music and sound generation, some basic modes of music composition are prepared; stochastic, the algorithmic and pre-recorded modes. Based on these prototypes of generated music, music and sound features such as timber, pitch, volume, tempo and style of music generation can be modified according to the input from the sound, image and motion modules. The harmonic structure refers to the compositional rules. By using the key information such as scene changing and applied force, the system changes the current mode of music generation. In addition, 72 sets of drum patterns have been prepared. These patterns have six different tempos and emerge so that the rhythm could correspond with the change in the image and sound input. Some other MIDI files of melody are also used. The musical features are modified by users in real time. The mobile robot becomes active when it is put into the environment where the robot and humans perform. All of output can then be continuously changed.

6.5.4 Performance demonstration

Through a number of experiments with dancers, the synchronization between humans and robot can be found. Dancers do not significantly care about the compositional structure of music, but the variety of the composed music. These experiments proved to us that the developed multimodal communication channels allow them to make an effective and interesting interaction with the robot.

6.6 Discussion

Four robotic interfaces in the virtual musical environment are successfully employed in art installations and demonstrations. These systems can exhibit a style "human-robot dance collaboration" where the robot moves in concert with the human performer by sensing the visual and audio information. Each system works as a sort of reflector to create an acoustic and visual space around the moving instrument. Moreover, the robot can display the refractive motion according to the context of performance to create the human-robot collaborative performance on stage.

In the effective interaction system for the felicitous performance, the appropriate responses must be required in real time. The substantial presence of a robotic interface is one of the possible solutions to make an effective reaction according to expressive motion of dancers. By providing the user's intention to the robot with his action, the robot not only reflects the intention with music generation but also with motion in real space. The small motions of the human may be amplified by the robot to make the performance more dynamic and exciting.

Moreover, the proposed musical environment is constructed under the Max/MSP programming environment. Therefore, users can easily associate the unrestricted relationship between different inputs and outputs because all the components communicate with each other through the MIDI network.

6.6.1 MIDI network

In the performance systems that allow users to get feedback for the emotional activation in terms of sound, music, image and motion, the flexibility of the instrumentation must be considered. The conventional interactive systems have paid less attention to designing the environment for users. The proposed platform provides a useful design environment for artists such as musicians, composers and choreographers not only to create music but also to coordinate the media with the aid of the MIDI network.

The developed systems as described in the case studies have realized the basic concept of MIDI Network, which provides a seamless communication among the devices of the proposed moving instruments. MIDI is a sufficient communication protocol because computer-based music creation is often done by MIDI sound, and easily available. All the exchanged data among the robot, audio synthesizer, main controller and other instruments is combined with the MIDI network in the developed systems. The diagram of the experimental system is shown in **Figure 6.20** (the iDance platform) and **Figure 6.21** (MIDItro). The robot can receive data as a control command, and transmit data from the sensors. The microchip converter operates to exchange between these data and the MIDI data. Therefore, the main controller can control the mobile robot with MIDI data just like mu-

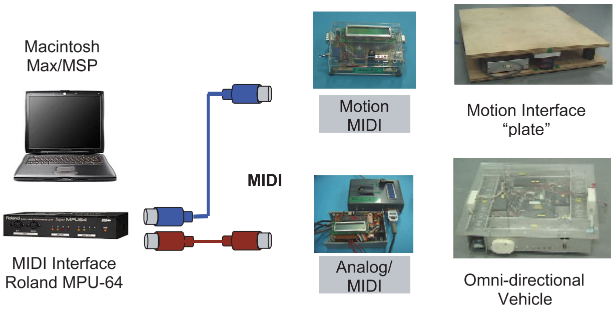


Figure 6.20. MIDI network in the iDance platform: Motion MIDI and Analog/MIDI are small logic components whose processor is a microchip PIC controller.

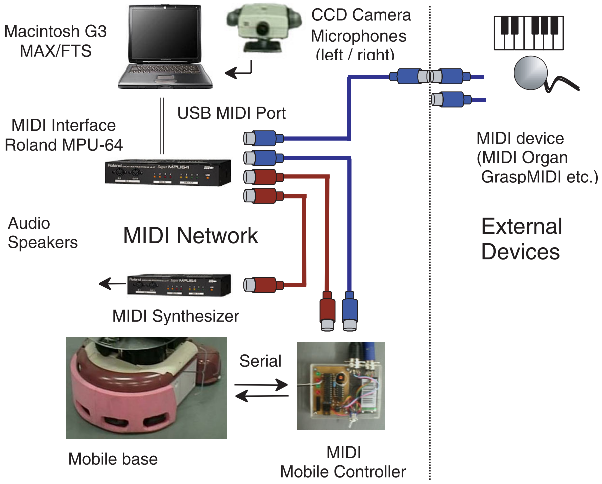


Figure 6.21. MIDI network in MIDItro: MIDI Mobile controller is a small logic component whose processor is a microchip PIC controller.

sical instruments with a MIDI software environment. By taking advantage of the MIDI format, other MIDI devices can be adopted with the system. For example, the MIDI organ enables users not only to play music but also to control the mobile robot.

6.7 Conclusions

In the near future, a human cooperative robot as a partner that makes cooperative work with people will appear. These robots are required to have a multimodal interface such as visual, audio and other sensory abilities, to enable them to share information space with humans. It is very important for such robots to have abilities not only to actively work in the human environment, but also to have a flexible and safe interface without the requirement of specific training or tuning.

In this study, reactive responses in human-machine communications have mainly been addressed. Recently, it is considered to add agents capable of reflecting the users' preferences. Further consideration is to provide more sets of musical rules to make the system more impressive.

The sketched multimodal musical environment is an interactive space where the robot would behave in response to the given stimuli and its internal state in the real environment, and where humans who play with the robot can continuously interact with the robot.

Therefore, the system works as an emotion activator stimulating human creativity. This means that it not only behaves like a human based on the emotional understanding from human movement, but also to "activate" humans by integrated outputs. In the dance and music performances that are required for its reaction in real time, a substantial interface plays an important role. Music shrouds people with a hypothetical space and is thought of as the creation of a virtual environment. The conventional graphical user interface is not sufficient for interactions in such an environment. It is considered that a substantial interface such as robot and virtual reality is absolutely imperative. The author considers that the embodied interaction between a human and the robot will open the next stage of human-machine collaborative musical performance.

Chapter 7

Kansei Quantization

7.1 Statistical Background

7.1.1 Quantitative theory IV

The quantitative theory [Hayashi, 1952] is a method to convert obtained data into a suitable quantity according to the purpose of use. Especially, Hayashi's Quantitative Theory IV aims at reconstructing the group by using a distance (similarity) between each object. Therefore, the theory is utilized to make clear about the structure of a given group, and to make a relative arrangement using the similarity e_{ij} between each other. In brief, when a pair has a large similarity, they are placed on near coordinates. On the other side, when a pair that has a small similarity, they are placed on far coordinates. The method is the same to minimize the following Q with the rearranged coordinates x_i using the similarity e_{ij} between data i and j.

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} e_{ij}(x_i - x_j) \tag{7.1}$$

The method is also called e_{ij} quantification. The various set of similarity are available to rearrange the coordinates. From the mathematical point of view, it is required to solve the following peculiar equation as shown in eq. (7.2). Human's subjective evaluation can be visualized by a possibly low dimensional - two or three - dimensional space, namely perceptual space.

$$-\left(\sum_{j=1,j\neq i}^{n} a_{ij}\right) \cdot x_i + \left(\sum_{j=1,j\neq i}^{n} a_{ij}\right) \cdot x_j = \lambda x_i \tag{7.2}$$

$$\begin{aligned} where \quad a_{ij} &= e_{ij} + a_{ji} \\ \lambda &= \frac{Q}{N\left[\left\{\sum_k^N {x_k}^2 \middle/ R\right\} - \left\{\sum_k^N x_k \middle/ R\right\}^2\right]} \end{aligned}$$

7.1.2 Multidimensional scaling

The aim of multidimensional scaling (MDS) [Kruskal et al., 1978] is to provide an arrangement of a set of objects in a geometric configuration from the pattern of proximities (i.e., similarities or distances) between each two objects. The small proximities and distances among points on the obtained map mean the closer (smaller) the distance between the input objects. While, the large proximities and distances among points on the obtained map mean the further apart.

The process of MDS is to find a set of vectors in p-dimensional space such that the matrix of Euclidean distances among them corresponds as closely as possible to some function of the input matrix according to a criterion function called stress. The equation and formulation will be described later.

The simplified algorithm is as follows:

1. Assign objects to arbitrary coordinates in p-dimensional space.
2. Compute a matrix of Euclidean distances $\hat{D}$ among all pairs of objects.
3. Compare the $d(p)$ with the input distance matrix D by evaluating the stress function. The smaller the value, the greater the correspondence among them.
4. Adjust coordinates of each point in the direction for the best minimum stress.
5. Repeat steps 2 through 4 until stress get the lowest value.

Figure 7.1 shows an example of MDS with ten U.S. cities from square matrix of dissimilarity that corresponds to flying mileages between each two cities. It can be seen that the arrangement of ten cities is roughly displayed in the figure. Note that the arrangement is upside down in the vertical line due to the property of MDS.

7.1.3 Sammon's nonlinear mapping

A nonlinear mapping algorithm for data structure analysis has been proposed in [Sammon, 1969]. The purpose of this algorithm is to describe the structure in a lower dimensional space such that the inherent structure of the data is approximately preserved under the mapping.

The approximate structure preservation is maintained by fitting N points in the lower-dimensional space. The interpoint distances approximate the corresponding their interpoint distances in the original space.

Let $X_i, i = 1, ..., N$ be N vectors in the original space (m-dimensional), and let $Y_i, i = 1, ..., N$ be N vectors in the lower-dimensional space (n-dimensional). Then, let the distance between the vectors X_i and X_j be defined by ${d_{ij}}^*$, and also let the distance between the corresponding vectors

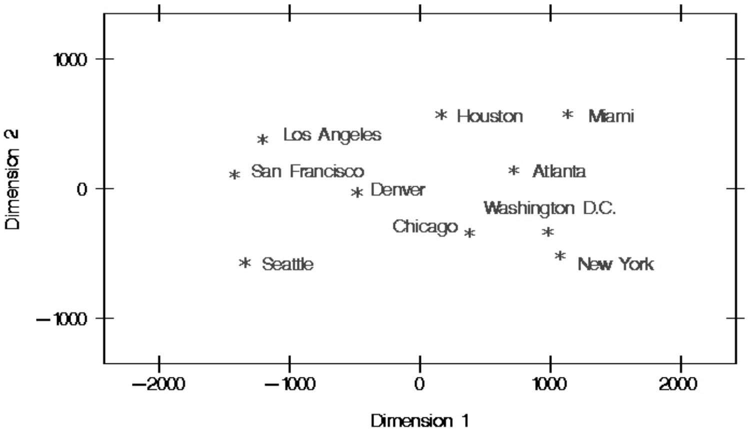


Figure 7.1. Analysis of flying mileages between ten U.S. cities. (©1999, SAS Institute): Created from symmetric & square matrix of dissimilarity that corresponds to flying mileages between each two cities.

Y_i and Y_j be defined by d_{ij}. The error E is defined how well the present configuration of N points in the n-space represents the N points in the original space, m-dimensional space.

$$E = \frac{1}{\sum_{i<j} [d_{ij}{}^*]} \sum_{i<j}^{N} \frac{[d_{ij}{}^* - d_{ij}]^2}{d_{ij}{}^*} \tag{7.3}$$

At the initial phase, after the number of dimension n is chosen, N vectors, Y_i, are randomly assigned in the n-dimensional space. The simplified algorithm is as follows:

1. Choose the number of dimension, n, for an initial configuration.
2. Assign objects to arbitrary coordinates, Y_i in n-dimensional space.
3. Compute an error E among all pairs of objects.
4. Adjust coordinates of each point in the direction, or equivalently change the number of dimension for decreasing the error. In the original work, a steepest descent procedure to search for a minimum of the error (see [Sammon, 1969])
5. Repeat steps 1 through 4 until the error E get the minimum value.

7.2 Review on Multilayered Perceptron

This section gives a review how multilayer networks can learn the nonlinear discriminant function, and can provide the optimum solution to an arbitrary classification problem. **Figure 7.2** shows a simple three-layer neural network, consisting of an input layer, hidden layer[1], and output layer, links

---

[1] Rumelhart called this internal representation layer. The output of this layer can not be directly obtained by the external environment - the input or output of the network.

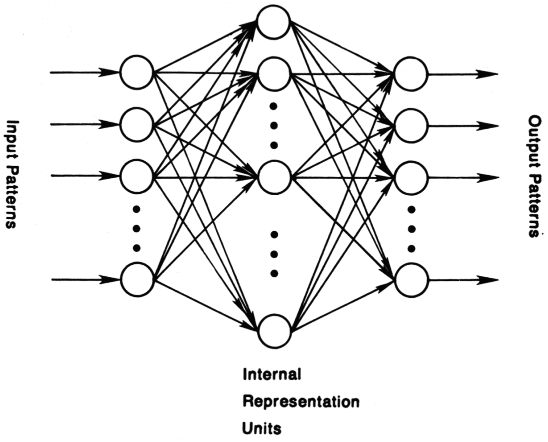


Figure 7.2. A Multilayer Perceptron (MLP): The middle layer is named as *internal representation layer* in [Rumelhart, 1986]. They remarked that the output are generated by the internal representation rather than by the original pattern. This figure is quoted from [Rumelhart, 1986]. In the original figure, the flows are drawn in a vertical line.

between layers, which are interconnected by modifiable weights.

The function of units is often called neurons or cells because these are based on properties of biological neurons. Each neuron computes the weighted sum of its input x_i to emit its output u_j with weights w_{ji} and its nonlinear function of its activation, f.

$$u_j = \sum_{i=1}^{m} w_{ji} x_i + w_{j0} = \sum_{i=0}^{m} w_{ji} x_i \equiv \boldsymbol{w}_j^t \boldsymbol{x} \tag{7.4}$$

The function of its activation - activation function - is chosen according to the desired training set, such as logistic function (sigmoidal function), and hyperbolic tangent function and other arbitrary functions, however, the differentiability is the only requirement of activation function.

$$f(x) = \frac{1}{1+\exp(-ax)} \qquad a > 0 \tag{7.5}$$

$$f(x) = a\tanh(bx) \qquad a > 0, \quad b > 0 \tag{7.6}$$

In multilayer network, when an input signal $\boldsymbol{x}$ are given to the network, the signal of each output unit represents the discriminant function $g_k(\boldsymbol{x})$.

$$g_k(\boldsymbol{x}) \equiv y_k = f\left(\sum_{j=1}^{n_H} w_{kj} f\left(\sum_{i=1}^{m} w_{ji} x_i + w_{j0}\right) + w_{k0}\right) \tag{7.7}$$

where the number of input units, hidden units, and output units denotes m, n_H, n, respectively. The interconnected weight between the input and hidden layer represents w_{ji}, and the weight between the output and hidden layer represents w_{kj}.

Consider the training error on a pattern to be the sum over output units of the squared difference between the desired output t_k given as a teacher signal and the actual output y_k. The cost function $E(t)$ is described as

$$E(\boldsymbol{w}) \equiv \frac{1}{2}\sum_{k=1}^{n}(t_k - y_k)^2 = \frac{1}{2}||\boldsymbol{t} - \boldsymbol{y}||^2 \tag{7.8}$$

One of the most popular methods for training multilayer neural networks is the backpropagation algorithm due to intuitive graphical representation and the simplicity of design of models. It is also called that a natural extension of the LMS algorithm and generalized delta rule.

Because the models are powerful and applicable for real world problems, the backpropagation learning rule is widely used in many works [Bishop, 1995]. At the same time, the theoretical properties have also been studied so far by many researchers [White, 1990] [Hornik et al., 1989] [Hornik, 1991] [Hornik, 1993].

The backpropagation learning rule is based on gradient descent. The weights are initialized with random values, and then modified so as to reduce the training error step by step learning. The updating weights are described as

$$\Delta\boldsymbol{w} = \eta\frac{\partial E}{\partial \boldsymbol{w}} \tag{7.9}$$

where η is the learning rate, that is often a function of time, $\eta(t)$. This iterative algorithm requires taking a weight vectors at iteration t and updating it as

$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + \Delta\boldsymbol{w}(t) \tag{7.10}$$

In the case of three-layered network, to start with the hidden-to-output weights, w_{kj}. eq. (7.10) can be re-written in component form

$$\Delta w_{kj} = -\eta\frac{\partial E}{\partial w_{kj}} \tag{7.11}$$

w_{kj} is not explicitly dependent on E. Therefore, eq. (7.11) can be described using the chain rule for differentiation:

$$\frac{\partial E}{\partial w_{kj}} = \left(\frac{\partial E}{\partial u_k}\right)\frac{\partial u_k}{\partial w_{kj}} = \left(\frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial u_k}\right)\frac{\partial u_k}{\partial w_{kj}} \tag{7.12}$$

where δ_k is defined as the sensitivity of unit k

$$\delta_k = -\frac{\partial E}{\partial u_k} = -\frac{\partial E}{\partial y_k}\frac{\partial y_k}{\partial u_k} = (t_k - y_k)f'(u_k) \tag{7.13}$$

The last partial derivative of eq. (7.12) is from eq. (7.4):

$$\frac{\partial u_k}{\partial w_{kj}} = h_j \tag{7.14}$$

Bringing together these equations, the learning rule for the hidden-to-output can be described as:

$$\Delta w_{kj} = \eta \delta_k h_j = \eta (t_k - y_k) f'(u_k) h_j \tag{7.15}$$

The learning rule for the input-to-hidden units is also described with the delta rule. Calculate carefully the summation using the chain rule.

$$\begin{aligned} \frac{\partial E}{\partial w_{ji}} &= \frac{\partial E}{\partial h_k} \frac{\partial h_k}{\partial v_k} \frac{\partial v_k}{\partial w_{ji}} \\ &= \left(\frac{\partial}{\partial h_j} \left[\frac{1}{2} \sum_{k=1}^{n} (t_k - y_k)^2 \right] \right) f'(v_j) x_i \\ &= \left(- \sum_{k=1}^{n} (t_k - y_k) f'(u_k) w_{kj} \right) f'(v_j) x_i \end{aligned} \tag{7.16}$$

The sensitivity for a hidden unit is defined to be:

$$\delta_j \equiv f'(v_j) \sum_{j=1}^{n} w_{kj} \delta_k \tag{7.17}$$

Finally, the learning rule for the input-to-hidden units is summarized as:

$$\Delta w_{ji} = \eta \delta_j x_i = \eta \left(\sum_{j=1}^{n} w_{kj} \delta_k \right) f'(v_j) x_i \tag{7.18}$$

7.3 Distance Mapping Learning

The numbers of Euclidean distance between any two points of N points is $N(N-1)/2$. In n dimensions, the distance d_{ij} is defined as:

$$d_{ij}{}^2 = \sum_{k=1}^{n} \left(y_k^{(i)} - y_k^{(j)} \right)^2 \tag{7.19}$$

The $y_k^{(i)}$ represents the value of object i at k the dimension. However, some objects are not satisfied with the formula of distance at all times. Such objects often has redundancies, for example, subjective data. Therefore, even by taking less notice of a strictly mathematical definition, to obtain the data arrangement in possible low dimensions directly and approximately using the relationship between each data must be important.

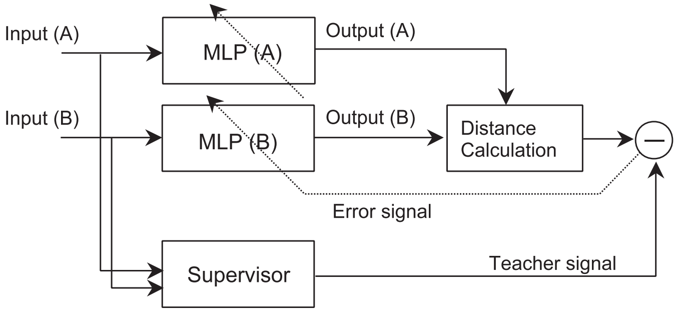


Figure 7.3. Illustration of *Distance Mapping Learning*: The MLP(A) and MLP(B) are the identical three-layered perceptrons whose weight coefficients are the same. Thus those give the same mapping.

In this work, by using a new type of neural network, a method to construct a non-linear mapping while keeping the distance between each data is proposed, namely, *Distance Mapping Learning* (DML), which can rearrange data in case that the only distance is given.

The purpose of the proposed method is to construct a non-linear mapping. Therefore, it is possible to obtain an evaluation about the relationship of distance toward unknown data, which means the data that is not used at the training process.

7.4 Neural Network Learning

The framework of the above-mentioned arrangement problem is formulated as follows. In the m-dimensional physical feature space X, a pattern i is represented as vector $\boldsymbol{x}^i = (x_1^i, x_2^i, \cdots, x_m^i)$. The target is to produce n-dimensional vector outputs $\boldsymbol{y}^i = (y_1^i, y_2^i, \cdots, y_n^i)$, $\boldsymbol{y}^j \in \mathrm{R}^n$ that preserve a desired distance $s_{ij}(\geq 0)$ with regard to given inputs $\boldsymbol{x}^i$ and $\boldsymbol{x}^j \in \mathrm{R}^m$.

The similarity s_{ij} in human perception approximates Euclidean distance d_{ij} between nonlinearly mapped patterns $\boldsymbol{y}^i, \boldsymbol{y}^j$ $(=\Phi(\boldsymbol{x}^i), \Phi(\boldsymbol{x}^j))$ from physical feature space X. To solve the above problem, the following fitting value W is minimized under a provisionally determined dimension order n.

$$W = \sum_i \sum_j (s_{ij} - d_{ij})^2 \tag{7.20}$$

$$d_{ij} = \begin{cases} \|\boldsymbol{y}^i - \boldsymbol{y}^j\| & (a) \\ \|\Phi(\boldsymbol{x}^i) - \Phi(\boldsymbol{x}^j)\| & (b) \end{cases} \tag{7.21}$$

where d_{ij} denotes the distance between mapped vectors. The similarity value s_{ij} can be obtained by human evaluations by means of a questionnaire, i.e., a pair comparison method. The geometric arrangement in Euclidean space is chosen because it is helpful for intuitive understanding of

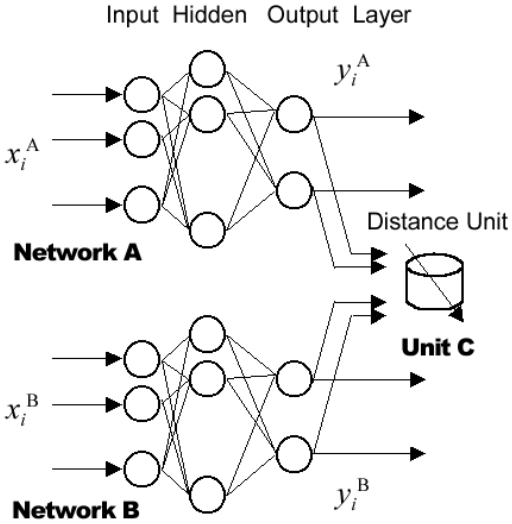


Figure 7.4. Network structure: This model is available for non-linear mapping based on the distance data. The Network A and Network B are the identical three-layered perceptrons. This is named as *Distance Mapping Learning* network.

the data structure. In the conventional multidimensional scaling method [Kruskal et al., 1978], the target is to minimize eq. (7.20) under the constraint of eq. (7.21)(b), not containing mapping function Φ with input vector $\boldsymbol{x}$. Therefore, the arranged vector $\boldsymbol{y}$ does not have any relationship with $\boldsymbol{x}$. However, the proposed method is to minimize eq. (7.20) under the constraint of eq. (7.21)(b) so that nonlinear mapping function Φ could be derived from neural network learning. The diagram of the network is illustrated in **Figure 7.3**.

7.4.1 Network structure

The structure of *Distance Mapping Learning* is shown in **Figure 7.4**. Network A and network B are identical three-layered perceptrons. Two objects A and B are required for the inputs to the networks. The outputs are the semantic parameters of objects A and B. The nonlinear mapping between the physical space an semantic space is therefore done by networks A and B. The outputs of A and B are connected in parallel to unit C which calculates the distance d_{AB} in the semantic space. When x_i^A and x_i^B are input, the network outputs y_i^A and y_i^B from each output layer, respectively, while input vectors $x_i^A, x_i^B (i = 1, 2, \cdots, m)$ are given. Each output vectors of networks A, B can be of arbitrary dimensions.

In the learning process, rather than providing networks A and B with absolute coordinates as a teacher signal, the semantic distance between the inputs to network A and network B, s_{AB}, is given.

7.4.2 Formulation

Error backpropagation algorithm Multi-layer perceptron (MLP) based on an error backpropagation algorithm for minimizing the mean square error is most popular [Rumelhart, 1986]. The least-squares learning and regression are discussed in [Geman et al., 1992]. This states that among all the functions of $\boldsymbol{x}$, regression is the best predictor of $\boldsymbol{y}$ given $\boldsymbol{x}$, in the mean-squared-error sense.

A training set $(x^1, y^1), (x^2, y^2) ... (x^p, y^p) ...$ is a collection of observed $(\boldsymbol{x}, \boldsymbol{y})$ pairs. In other words, the pair $(\boldsymbol{x}, \boldsymbol{y})$ obeys some unknown joint probability distribution.

To construct a non-linear function $\Phi(\boldsymbol{x})$ based on the training set is equivalent so that Φ satisfies the desired output $\boldsymbol{y}$. Φ is generally determined so as to minimize a given cost function. The learning process of a multi-layer perceptron is described as:

$$\varepsilon^2(\Phi) = \int \|y^p - \Phi(x^p)\|^2 p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x} \tag{7.22}$$

where $p(\boldsymbol{x}, \boldsymbol{y})$ and ε denote the probability density function and mean-squared error of $\Phi(\boldsymbol{x})$ that represents a nonlinear functional of the neural network, in which the goal of learning process is to minimize ε^2.

Recalling Bayes formula and let $\boldsymbol{\theta} = \Phi(\boldsymbol{x})$, the derivative $\partial\varepsilon^2(\Phi)/\partial\boldsymbol{\theta}$ is calculated:

$$\begin{aligned}
\frac{\partial\varepsilon^2(\Phi)}{\partial\boldsymbol{\theta}} &= \frac{1}{\partial\boldsymbol{\theta}}\Big[\iint \|y^p - \theta^p\|^2 p(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} d\boldsymbol{x}\Big] \\
&= \int -2(\boldsymbol{y} - \boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{x}) \Big[\int p(\boldsymbol{x}) d\boldsymbol{x}\Big] d\boldsymbol{y} \\
&= \int -2(\boldsymbol{y} - \boldsymbol{\theta}) p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} \\
&= -2\left[\int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} - \int \boldsymbol{\theta} p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y}\right] \to 0
\end{aligned} \tag{7.23}$$

eq. (7.22) that is a simple quadratic formula regarding $\hat{\boldsymbol{y}}$ is delivered by probability distribution $\boldsymbol{y}$ conditioned by $\boldsymbol{x}$:

$$\hat{\boldsymbol{y}} = \Phi_{opt}(\boldsymbol{x}) = \int \boldsymbol{y} p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} \tag{7.24}$$

Distance mapping learning On the other hand, the proposed neural network differs from the conventional MLP in which the desired distance between input data is given as a teacher signal. The mean-squared error ε is described as follows with probability density function $p(\boldsymbol{x}^A, \boldsymbol{x}^B, \boldsymbol{s})$ where

$\boldsymbol{s}$ denotes the desired distance between $(\boldsymbol{x}^A, \boldsymbol{x}^B)$, which may not be a deterministic variable even though $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$ are determined.

$$\varepsilon^2(\Phi) = \int \left\| s_{ij} - d_{ij} \right\|^2 p(\boldsymbol{x}^A, \boldsymbol{x}^B, \boldsymbol{s}) d\boldsymbol{x}^A d\boldsymbol{x}^B d\boldsymbol{s} \tag{7.25}$$

As well as the above operation, the derivative $\partial \varepsilon^2(\Phi)/\partial \boldsymbol{\theta}$ is calculated:

$$\begin{aligned}
\frac{\varepsilon^2(\Phi)}{\partial \boldsymbol{\theta}^A} &= \frac{1}{\partial \boldsymbol{\theta}^A}\Big[\iiint \left\| s_{ij} - d_{ij} \right\|^2 p(\boldsymbol{x}^A, \boldsymbol{x}^B, \boldsymbol{s}) d\boldsymbol{x}^A d\boldsymbol{x}^B d\boldsymbol{s}\Big] \\
&= \iiint \left(-2\frac{\boldsymbol{s}-\boldsymbol{d}}{\boldsymbol{d}}\right)\left(2\boldsymbol{\theta}^A\right) p(\boldsymbol{s}|\boldsymbol{x}^A, \boldsymbol{x}^B) p(\boldsymbol{x}^A) p(\boldsymbol{x}^B) d\boldsymbol{x}^A d\boldsymbol{x}^B d\boldsymbol{s} \\
&= \int \left(-2\frac{\boldsymbol{s}-\boldsymbol{d}}{\boldsymbol{d}}\right)\left(2\boldsymbol{\theta}^A\right) p(\boldsymbol{s}|\boldsymbol{x}^A, \boldsymbol{x}^B)\Big[\int p(\boldsymbol{x}^A) d\boldsymbol{x}^A\Big]\Big[\int p(\boldsymbol{x}^B) d\boldsymbol{x}^B\Big] d\boldsymbol{s} \\
&= -4\left[\int \left(\frac{\boldsymbol{s}}{\boldsymbol{d}} - 1\right) \boldsymbol{\theta}^A p(\boldsymbol{s}|\boldsymbol{x}^A, \boldsymbol{x}^B) d\boldsymbol{s}\right] \to 0
\end{aligned} \tag{7.26}$$

From this, a simple relation is delivered regardless of $\hat{\boldsymbol{y}}$ and probability distribution $\boldsymbol{s}$ conditioned by $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$:

$$\boldsymbol{d} \to \boldsymbol{s} \tag{7.27}$$

Note that the proposed method is independent of linear transformation and rotation in output space. Because the network that is trained by the relative distances, these transforms can be applied without missing the property.

$$\begin{aligned}
d_{ij}(\Phi(\boldsymbol{x}^i), \Phi(\boldsymbol{x}^j)) &= \|\Phi(\boldsymbol{x}^i) - \Phi(\boldsymbol{x}^j)\| && (7.28) \\
&= \|(A\Phi(\boldsymbol{x}^i) + B) - (A\Phi(\boldsymbol{x}^j) + B)\| && (7.29)
\end{aligned}$$

(A: rotation matrix, B : translation matrix)

7.4.3 Learning rule

The connection weights in networks A and B are tuned to make the distance d_{AB} close to teacher signal s_{AB}. The sigmoid output function is used for each neuron, and a modified back propagation method is applied for the learning. Networks A and B start from the same initial connection weights and are trained in the same manner to give the same mapping.

The learning rule is described below. The cost function of DML E_d is defined as follows:

$$E_d = \sum^{n} (s_{AB} - d_{AB})^2/2 \tag{7.30}$$

$$d_{AB} = \sqrt{\sum_{k=1}^{n} (y_k^A - y_k^B)^2} \tag{7.31}$$

Updating the connection weights in each iteration is performed as follows:

$$\Delta w_{ij}(t+1) = \alpha \frac{\partial E_d}{\partial w_{ij}} + \eta \Delta w_{ij}(t) \tag{7.32}$$

where α and η represent the learning parameter and the momentum, respectively.

In this case, since w_{ij} is not explicitly dependent on E, the equation should be re-written using the chain rule for differentiation. Moreover, y_k is either dependent on E explicitly, but dependent on d_{AB} in this network. The above equation is therefore described as:

$$\begin{aligned}
\frac{\partial E_d}{\partial w_{ij}} &= \sum_{k=1}^{n} \left[\left(\frac{\partial E_d}{\partial d_{AB}} \right) \left(\frac{\partial d_{AB}}{\partial y_k^A} \right) \left(\frac{\partial y_k^A}{\partial w_{ij}} \right) \right. \\
&\quad \left. + \left(\frac{\partial E_d}{\partial d_{AB}} \right) \left(\frac{\partial d_{AB}}{\partial y_k^B} \right) \left(\frac{\partial y_k^B}{\partial w_{ij}} \right) \right] \\
&= \left(-(s_{AB} - d_{AB}) \sum_{k=1}^{n} (y_k^A - y_k^B) \frac{\partial y_k^A}{\partial w_{ij}} \right) \\
&\quad + \left(-(s_{AB} - d_{AB}) \sum_{k=1}^{n} (y_k^A - y_k^B) \cdot \left(- \frac{\partial y_k^B}{\partial w_{ij}} \right) \right) \\
&= -(s_{AB} - d_{AB}) \sum_{k=1}^{n} (y_k^A - y_k^B) \left(\frac{\partial y_k^A}{\partial w_{ij}} - \frac{\partial y_k^B}{\partial w_{ij}} \right)
\end{aligned} \tag{7.33}$$

The partial derivative $\partial y_k^A / \partial w_{ij}$ and $\partial y_k^B / \partial w_{ij}$ are calculated as the conventional backpropagation algorithm as described in 7.2:

$$\frac{\partial y_k^A}{\partial w_{ij}} = f'(u_k^A) h_j^A, \qquad \frac{\partial y_k^B}{\partial w_{ij}} = f'(u_k^B) h_j^B \tag{7.34}$$

For describing the learning rule simply, the sensitivity δ_k^A, δ_k^B is defined to be:

$$\delta_k^A = (s_{AB} - d_{AB}) \sum_{k=1}^{n} (y_k^A - y_k^B) f'(u_k^A) \tag{7.35}$$

$$\delta_k^B = (s_{AB} - d_{AB}) \sum_{k=1}^{n} (y_k^A - y_k^B) f'(u_k^B) \tag{7.36}$$

The learning rule for hidden-to-output units is finally described:

$$\Delta w_{kj} = \alpha (\delta_k^A h_j^A - \delta_k^B h_j^B) \tag{7.37}$$

The learning parameter α and the momentum η are empirically chosen for the stable convergence.

The network does not need absolute coordinates as the desired output, only the desired distance between input data. It should be noted that the structure and initial condition of network A and network B is identical, while the number of cells in the output layers of network A network B can be set arbitrarily, depending upon the structure of the required semantic space. In the present work, experiments are carried out with two-dimensional output space. Euclidean distance is used as a teacher signal. In this framework, this approach is basically similar to Sammon's nonlinear mapping [Sammon, 1969] and realizes the algorithm by a modified multilayer perceptron.

7.4.4 Visualization of the mapping aspect

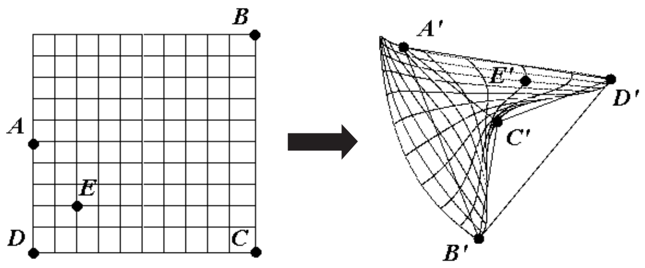


Figure 7.5. An aspect of the obtained mapping by distance mapping learning: In the left figure, the significance of X and Y axis is not determined.

Figure 7.5 shows an example of non-linear mapping with four points that consisting of a triangle and the center of gravity. In the left figure, the significance of X and Y axis is not determined. It is possible to determine the significance according to the purpose of use. From now, the space is named as *initial coordinates space.* While, the right figure shows a non-linear mapped space by the proposed method. The space is named as a *mapped space* or *description space.*

In the initial coordinates space (left figure), objects B, C and D are arranged at each corner square, and A is arranged at a down half of side from the other square. These four points are used for the training of the network. E is not used at the training phase. The teacher signal is respectively given to AB, BD and DA so that A, B and D could form a regular triangle. The relation of distance with regard to C are then given in order that C is arranged on the circumcenter of ΔABD.

In the arrangement of initial coordinates space, the position C is put at the outer side of ΔABD. Therefore, the space around C is expected to

Training set	Test set
Flute*	Clarinet
Classic Guitar*	Tenor Saxophone
Violin*	Fagotto
Oboe*	Harmonica
Trumpet*	Euphonium
Trombone*	Tuba
Cello*	Horn
Piano*	Electric Guitar
Cembalo*	Harp
	Contrabass
	Viola
	Organ
	Accordion

Table 7.1. Training and test data set with musical instruments, totally 22 sounds: The 9 sound of musical instruments marked * are used for learning in the experiment.

be strained inside. In the initial coordinates space, distances of all sides at ΔABD are different from each other, and have the relation $BD > AB > AD$. Since the distance $B'D' = A'B' = A'D$ is given as teach signal at the training process, the different aspects of constriction are shown in the left figure along each side. Moreover, when an unknown data E is given, it is rearranged onto E' of the right figure. E was on the line BD in the initial coordinate space. However, mapping of C makes an effect on E that mapped onto the inside of ΔABD. In this way, the proposed neural network enables the evaluation toward such an unknown data.

An experiment with a real data set was then conducted. **Figure 7.6** illustrates a result with the sound of musical instruments [Suzuki et al., 1997].

The first step is to extract the feature parameters from the sounds. 9 sounds are used as the example data to train the neural network. The sounds are digitized into 16bit digital value at the sampling rate of 44.1kHz. The duration of one sound is about 1.5 seconds to make 65,535 digital data. The data is divided into 125 frames. The length of frame is 2048(46ms). Every frame overlaps by 512 words(11.5ms). FFT (Fast Fourier Transformation) is performed for every frame data to obtain the spectrograph. The sound spectrograph is divided into 8×16 parts; 8 parts along the time axis, 16 parts along the frequency axis. Total power x_i in part i $\{i=1,2, \ldots , 8\times16\}$ is used as the primary sound features. Therefore a sound data is characterized by a physical feature vectors of 128 dimensions.

$$\begin{aligned} X(n)\{x_1, x_2, ..., x_{128}\} \\ x_i = \sum_{area(\alpha,\beta)} p(t,f) \\ \alpha = 1, 2, ..., 8 \end{aligned} \tag{7.38}$$

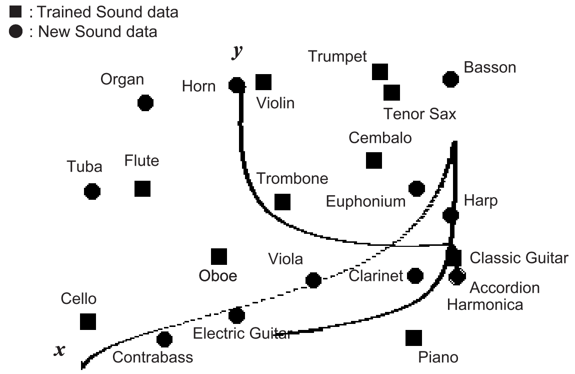


Figure 7.6. An arrangement of musical instruments: This is also an aspect of the obtained mapping by distance mapping learning.

$$\beta = 1, 2, ..., 16$$

To reduce the feature vector dimension the Principle Component Analysis (PCA) is introduced. The QR-law using a covariance matrix is adopted to calculate numerically. Sound data can be represented by using the principle component scores. The valid dimension p of principle components is determined so that the cumulative distribution rate is more than 95%. Therefore sound data can be represented by the vector $\{P_i\ ;\ i{=}1,2,...,\ p\}$ which consists of principle component scores as shown in eq. (7.39) where e_{ij} denotes the eigen vector element of jth principle component. The distribution rate of the first principle component was 61.5%, and that of the second component was 25.9%. And four principle components could satisfy the cumulative distribution rate 95% of the source data.

$$S(n)\{P_1, P_2, ..., P_m\}$$

$$P_i = \sum_{j=1}^{8\times 16} e_{ij} x_j \tag{7.39}$$

The second step is to construct the rule of non-linear mapping based on comparisons between the sound data. The evaluations of a subject toward 10 musical instruments is used as teach signals. The subjects have no special knowledge on music.

The learning time using these data took about 300 seconds CPU time on Pentium III computer. From a result of the first experiment, an emotional sound space can be obtained according to the data of subjects. After the learning, the rest 12 data which were not used in learning, are mapped into the emotional sound space to examine the appropriateness of the obtained mapping function. The mapping results were almost the same as the arrangement using a multidimensional scaling method (MDS).

The third step is to consider the case of new data. This means that it is not used on learning. Above-mentioned 22 instruments are illustrated in **Figure 7.6**, namely 'Emotional Sound Spaces'. Since the network has ability to map continuously, such new sound data can be arranged on the obtained emotional space. It should be noted that the mapped axes are represented in the figure where two distorted lines correspond to the accentuated axes in the original feature space whose axes correspond to principle component. In these figures after mapping, each axis mean renew axis to evaluate sound from emotional information. Noted that the proposed method will not only give the similar measure as in MDS method but also map a new sound in the obtained emotional space, which is a great advantage comparing with the conventional statistical methods.

The experimental result shows that human evaluation towards sounds is not always determined according to the instruments type such as wind wood, brass, and strings, but is inclined to focus on the pitch. It can be seen that the pitch of the instrument is higher in the top-left area, while is lower in the bottom-right area.

7.5 Model of Facial Expression Perception

Many research studies have been reported so far concerning the relationship between a facial expression and its impression, especially in the field of psychology, cognitive science and engineering, for example, [Ekman, 1992]. Facial expression has several physical features such as the shape and location of each facial part, the skin color and the wrinkles. Non-verbal communication takes place by perceiving and generating his/her emotional facial expression. Humans can guess another person's mental state and emotion according to the movement and posture of facial expression.

To judge and categorize emotional facial expressions have long been a research subject, mainly in experimental psychology. It is well-known that Schlosberg [Schlosberg, 1952, Scholsberg, 1954] proposed a three-dimensional theory of facial expressions, which states that facial expression is located along three scales: pleasant - unpleasant, attention - rejection, and sleep - tension. The low-dimensional semantic space is composed of the high-dimensional physical features of the face. In such a research, the method of measuring the human impression of facial expression is usually done by means of questionnaires. The facial expressions are arranged onto a low-dimensional semantic space using statistical methods. The analyzed result, however, does not contain the relationship between the physical features and the impression parameters. There have been few research studies conducted to relate the physical features of facial expression into its impression parameters. In such a case, a neural network is appropriate for acquisition of nonlinear mapping between qualitatively different data. The nonlinearity in the perceptional processing of humans, which is regarded as a sensitive or intuitive process, can be effectively performed using a multilayer perceptron classifier [Zhang et al., 1998][Padgett et al., 1997], compared with conventional statistical approaches which are based on linear functions [Russell et al., 1985][Katsikitis, 1997].

A modeling of the human perception of facial expressions is conducted for the evaluating the performance of the distance mapping learning network. Note that an element of semantic space corresponds to the output space of the proposed neural network, where the distance is measured by the similarity (similar - dissimilar) via facial expression perception. In other words, this is a new method of treating symbolic values such as similarity.

Some similarity-based learning methods have been reported, for example, a similarity-based distance is used, for instance, for image database organization [Squire, 1998], the classification method [Duch, 1997]. However, the proposed method differs from these approaches in the training of the network. In particular, the network is trained by a pair of inputs and their distance, a similarity value, between given inputs.

The obtained human judgment of facial image is employed to create a reduced feature space that is derived from the input feature space. The

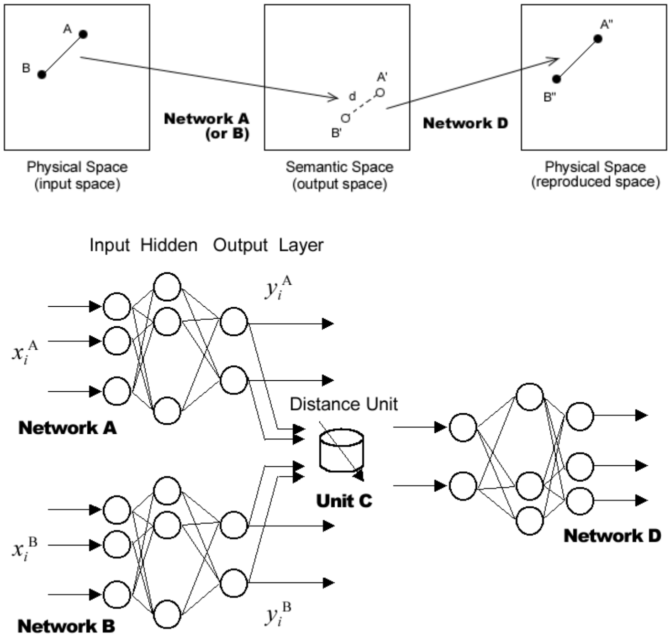


Figure 7.7. The concept of interrelationship and the structure of the learning network: Both of the upper and lower figures correspond to each other. Parameters in Physical Space is input to Network A or B. Outputs of Network A or B are then semantic parameters in Semantic Space. By a reproduction network, Network D, an inverse mapping is constructed with multi-layer perceptrons.

human impressions of faces can thus be visualized by the mapped physical facial parameters onto the constructed two-dimensional semantic space. Additionally, by obtaining the inverse nonlinear mapping, facial expressions can be reproduced from the semantic parameters.

7.5.1 Interrelating modeling

The characteristics of the proposed interrelation model of facial expression perception are summarized as follows:

1. The interrelating between the perceived physical feature of a face and its impression is regarded as a nonlinear mapping.
2. The mapping rule is determined by the desired distances between any two given inputs.

It is proper that human impression of a facial expression is caused by the perceived physical feature of the face. A human's impression is usually obtained by means of questionnaires. Humans, however, cannot arrange a facial image onto an absolute coordinate in the semantic space. Therefore, the common way is to represent data with an adjective pairs questionnaire. In addition, the simplest way is a pair comparison method, in which each subject scores a similarity value while two data are shown. In both cases, the tested sample data are arranged in low-dimensional space by the obtained human judgment. The axes of this space conventionally refer to the principal components which are calculated by a particular statistical method.

Unlike those methods, this approach aims to solve the arrangement problem by a nonlinear mapping between the facial physical feature space and a low-dimensional semantic space utilizing a distance-based neural network. Data such as human judgment are often noisy and redundant. In such cases, it can be important to obtain the arrangement in a low-dimensional space.

The proposed system consists of two stages: (1) construct the semantic space and (2) reproduce facial expression from its semantic parameter. The neural network in the first stage which is called Distance Mapping Learning (DML) consists of three parts, networks A, B and unit C. Network D in the second stage is called the reproduction network.

The reproduction network, network D, is an identical three-layered perceptron, which is connected to network A or network B. The sigmoid output function is used for each neuron, and the back propagation method is applied for the learning. The input of the reproduction network is used as the output of network A (or B). The teacher signal of the network corresponds the input data of network A (or B) so as to obtain a nonlinear inverse mapping from the semantic parameters to the physical feature parameters in order to reproduce the facial expression.

The proposed network differs from a sandglass type neural network in updating the weight. The reproduction network is trained once after DML is converged. Thus the whole input space of DML that represents any possible facial expressions can be mapped onto the semantic space. The reproduction network is then trained in order to construct an inverse mapping to reproduce physical feature parameters.

7.5.2 Description of facial expression

As for the description of facial expression, FACS (Facial Action Coding Unit) [Ekman et al., 1978] became popular, and the related studies have been reported in the engineering field as well as psychology. This method enables us to describe any facial expression with 44 kinds of Action Units (AU) that are related to the movement of facial muscles.

In the present work, we, however, adopted a line drawing facial model, that is an abstraction of facial expression, proposed in [Yamada, 1993a, Yamada, 1993b] as a set of typical physical features of facial expression. It consists of nine physical facial parameters, which correspond to facial movements such as raise/lower of inner/outer eyebrows, eyes and upper/lower

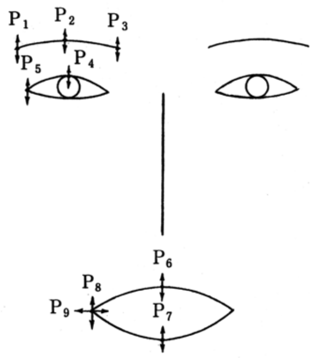
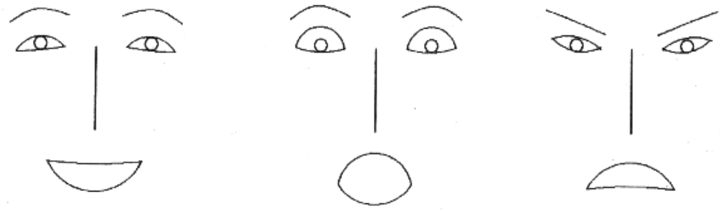
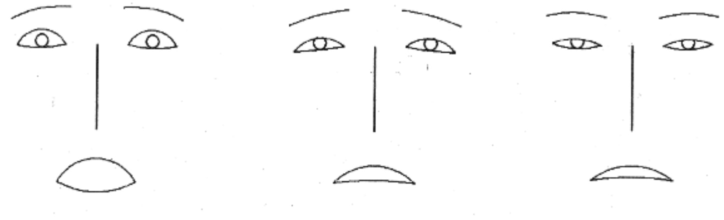


Figure 7.8. A line drawing model of facial expression proposed by Yamada [Yamada, 1993a]. The nine parameters indicated by P_i are used to generate an image of facial expression. The movement of each parameter which is allocated to the eyebrows, eyes and mouth changes the intensity of facial expression.



A Happiness C Surprise E Anger



S Fear J Sadness H Disgust

Figure 7.9. Six fundamental emotional faces: Each alphabet letter corresponds to one in the following experimental results. The correspondences of the rest of the alphabet is referred to in [Yamada, 1993b].

Configuration	DML	RN
Parameters		
Learning parameter α	0.4	0.25
Momentum η	0.3	0.3
Number of Cells		
Input Layer	9	2
Hidden Layer	12	9
Output Layer	2	9
Learning		
Error limit	1×10^{-5}	1×10^{-5}
Instance	21	21

Table 7.2. Parameter configuration in the experiment for the analysis of facial expression. DML, RN represent *Distance Mapping Learning* and *Reproduction Network*, respectively. The results are illustrated in **Figure 7.10**, **Figure 7.11**.

lips as illustrated in **Figure 7.8**. Each face has line symmetry with respect to the vertical central line on which the nose is located and fixed. Each feature point is consistently connected to the others, introducing the Spline interpolation.

Using this line drawing model, six faces of typical emotional categories such as *happiness, anger, disgust, fear, surprise* and *sadness* are acquired on the average of images drawn by 36 subjects in the previous work (see [Yamada, 1993b]. Subjects moved the feature points of **Figure 7.8** in order to create the desired emotional face. **Figure 7.9** shows the six fundamental emotional faces, where each letter indicated on each face corresponds to one in the later experiments.

Twenty-one physical facial expressions are then provided that include fifteen in-between images of each emotional face in addition to the above six fundamental faces. The mean image between the two categories is made by the average of each parameter.

The line drawing model focuses not on multiple shapes of the face but on the essential expressions. The advantage is the ease of quantitative operation. Unlike FACS that contains redundancy due to the facial muscle system, this model can operate directly with the geometrical features of facial expression.

The input layer of the network has nine neurons corresponding to nine feature parameters of a facial expression, $x_i^{(k)}(i = 1, 2, \cdots, 9; k = 1, 2, \cdots, 21)$. Twenty-one prepared facial expressions, which consist of six fundamental images; *happiness, anger, disgust, sadness, surprise* and *fear*, and fifteen intermediate images of these, are used as input patterns. A semantic distance which is obtained in advance by calculating adjective pairs scores is used as the teacher signal in this experiment.

The numbers of neurons in the hidden and output layer of networks A and B are set to 12 and 2, respectively. The learning parameter α is 0.4, and

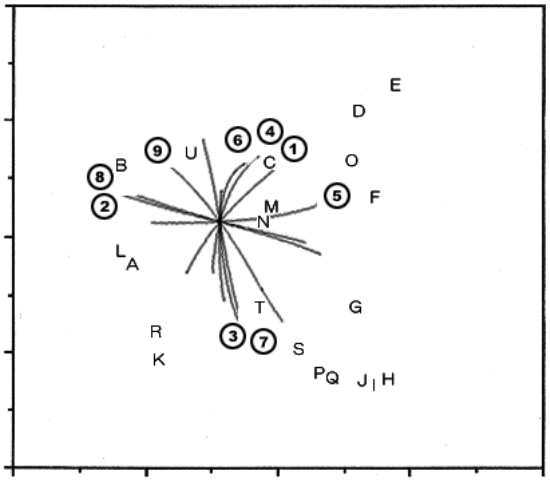


Figure 7.10. Twenty-one faces arranged on the semantic space: The figure was modified with the translation and rotation of the axes so that the horizontal and vertical axes could correspond to pleasantness and activity, respectively.

the moment η is 0.3. When the error is smaller than 1×10^{-5}, the training of the network is terminated. The average iteration was in the order of ten thousands.

After the training of DML is completed, the reproduction network enters the training phase. The number of neurons in each layer of network D is set empirically to 2, 9, 9, respectively. The learning parameter α is 0.25, and the moment η is 0.3. When the error is smaller than 1×10^{-5}, the training of the network is terminated.

Figure 7.10 shows the obtained semantic space. Each coordinate of the letter represents the output vector of the DML network, $\boldsymbol{y}^A$ (or $\boldsymbol{y}^B$). The figure was modified with the translation and the rotation of axes so that the axes could correspond to those extracted by previous research [Yamada, 1993b]. In this figure, the horizontal and vertical axes roughly correspond to pleasantness (pleasant-unpleasant), and activity (low-high), respectively. Each letter corresponds to each facial expression image in **Figure 7.8**. The correspondence of the rest of the alphabet letter is referred to in [Yamada, 1993b]. The numeric character represents each physical axis which corresponds to a feature parameter axis of the facial expression model. The intersecting point is the origin in the physical feature space, which corresponds to the neutral (base) face shown in **Figure 7.8**. The nonlinearly mapped axes of the physical feature can be shown in the semantic space. With the aid of the generalization ability of the neural network, any given artificial facial expressions are mapped onto the semantic space.

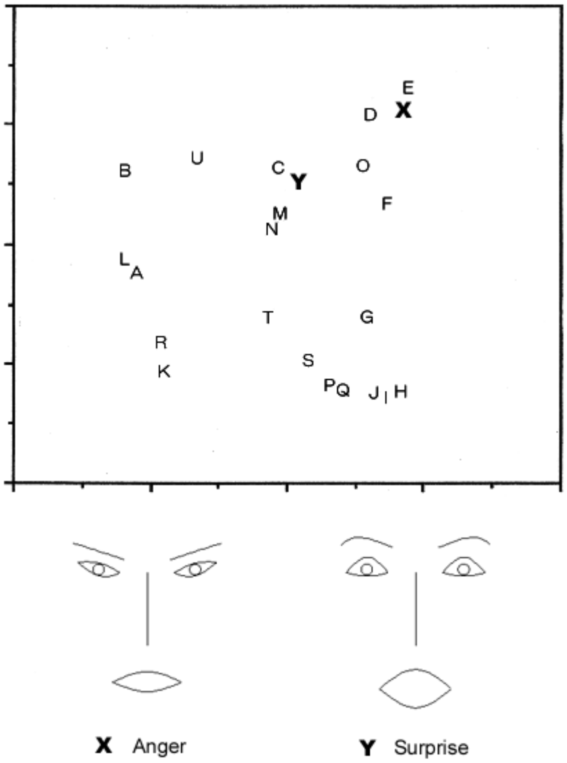


Figure 7.11. Estimation of reproduced faces: estimated faces are illustrated in the bottom of figure. Faces X and Y are estimated from a coordinate near E (*Anger*) and C (*Surprise*), respectively.

7.5.3 Reproduction network

Inverse mapping can be constructed by using a multi-layer perceptron with the back-propagation rule, presenting data in the semantic space as input and data in the physical feature space as output. The upper part of **Figure 7.7** shows the description of the neural network and the concept of the proposed model that contains the bi-directional correspondences between the physical and semantic spaces.

An example of the estimated facial image by the reproduction network is illustrated in **Figure 7.11**. The letter indicated with X, Y around E (Anger), C (Surprise) are coordinates in the semantic space. The reproduced faces from these locations that are shown in the bottom part of **Figure 7.11** can be recognized as faces in each emotional category. Even if the network with the training set of the line drawing model is used, the mapping result shows good performance with a real image set due to the generalization ability of the neural network.

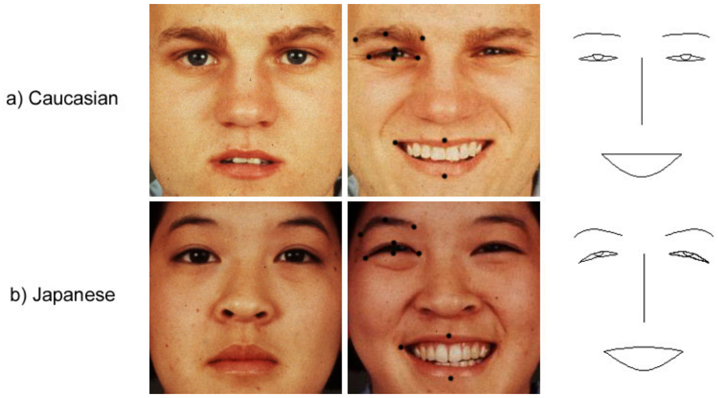


Figure 7.12. Examples of real image sets of Japanese and Caucasian facial expressions: images are extracted from JACFEE and JACNEUF facial image database [Biehl et al., 1997].

7.5.4 Application to a real image set

In this experiment, the proposed model with a real image set has been evaluated. The input of the network is facial stimuli of a real image, consisting of an expression of neutrality and happiness by one male Caucasian and one female Japanese, from the JACFEE and JACNEUF facial image database [Biehl et al., 1997], as illustrated in **Figure 7.13**.

The upper and lower lines show a male Caucasian and a female Japanese, respectively. The leftmost figure in each line shows a neutral expression, while the central figure presents an expression of happiness. Each right image is each facial image along the line drawing model according to the center figure. The nine feature points of each figure are obtained by a manual process. The differences between neutrality and happiness are thus used as the feature parameters of the facial expression of happiness.

As shown in **Figure 7.13**, both the Caucasian and Japanese facial images indicated double circles which are mapped onto the semantic space. It can be seen that these faces are arranged in close to the coordinate of *Happiness*. The same mapping rule of the neural network as shown in **Figure 7.10** was used for evaluation in this experiment.

7.6 Performance in Multi-Class Classification Problems

In this section, a method of multiclass classification by utilizing a distance mapping learning network is proposed. The network can obtain the nonlinear mapping between the input objects and the outputs by providing a pair of objects and the desired distance between them. It thus realizes

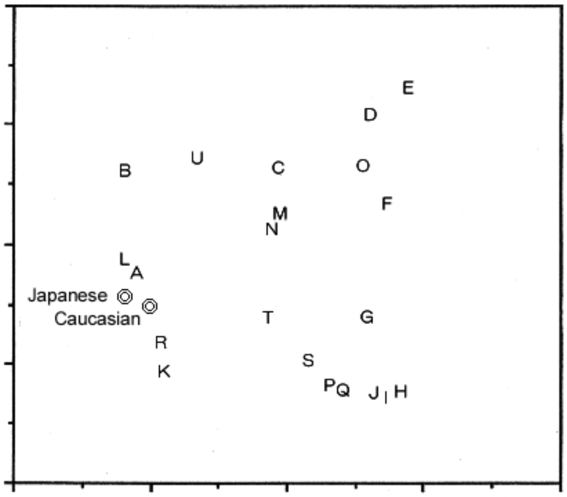


Figure 7.13. Application to a real image set: Both Caucasian and Japanese facial images are indicated by double circles which are mapped onto the semantic space. The learning set to construct the semantic space is the same as in **Figure 7.10** in section *7.5.2*.

multiclass classification based on pairwise classifications iteratively. The validity of the model with two classification problems will be shown, e.g., Iris classification and facial expression classification.

To date, a number of techniques of multiclass classification have been proposed. In the conventional classification problems, one tries to distinguish between two (or more) classes of objects, and most methods try to estimate the probability density of the target set. On the other hand, a powerful and effective method of two-class classification is proposed such as Support Vector Machine [Vapnik, 2000]. One approach to classify multiple classes is to combine two-class classifiers. Another approach has been reported about multiclass classification by combining one-class classification [Tax et al., 2001] that tries to distinguish between a set of objects and all other objects. In these cases, although one can choose the classifiers so as to adapt each two-class classification of target set, the number of class is given or provisionally has to be set. However, it is often impossible or difficult to know the number of class in the real-world problem.

In this section, a method of multiclass classification by utilizing *Distance Mapping Learning* network will be described. The network can obtain the non-linear mapping between the input objects and the outputs by providing the desired distance between the objects, not the desired output. The desired distance represents the similarity between the input objects. The network thus realizes multiclass classification based on pairwise classifications.

By comparing to other existing methods, the proposed method differs

Configuration	Setting
Parameters	
Learning parameter α	0.3
Momentum η	0.2
Scaling parameter ρ	0.6
Number of Cells	
Input Layer	4
Hidden Layer	6
Output Layer	2
Learning	
Iteration	10,000
Instance	75
Test Data	75

Table 7.3. Parameter configuration in the experiment for Iris classification. The result is illustrated in **Figure 7.14**.

from these approaches in the training of the network. It should be noted that the proposed model can deal with classification for an unknown number of class. Only by presenting a pair of objects and the desired distance, the network can map the objects onto the output space of arbitrary dimensions, which is regarded as data description space. In this section, an application to multiclass classification with some experimental results will be shown.

7.6.1 Iris classification problem

In this experiment, the proposed model is applied to Iris Classification problem [Fisher, 1936]. An Iris flower that has 4 attributes (length and width of the flower's sepal and petal) is classified into one of three classes (Iris-Setosa, Iris-Versicolor, Iris-Virginica). One class is linearly separable from the other two classes; the latter are not linearly separable from each other.

In the learning phase, a pair from 75 instances (contains evenly three classes) and the teacher signal s_{AB} as described below are given for training the network.

$$s_{AB} = \begin{cases} 0 & \\ \rho & (\rho > 0) \end{cases} \tag{7.40}$$

where ρ represents the scaling parameter that is the scale of the distance between two classes, A and B. If a training pair is chosen from the same class, the teacher signal is set to 0.0. While, if they are chosen from different classes, the value is set to ρ. In this experiment, ρ is set to 0.6. The number of cells in the input, hidden and output layers are 4, 6 and 2, respectively. The learning rate is set to 0.3, and the momentum is set to 0.2. After training the network, the model holds a non-linear mapping from 4 attributes of Iris flower to 2 dimensional output space.

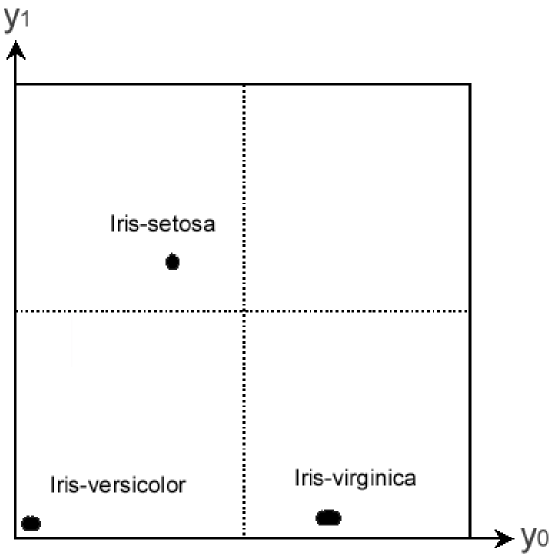


Figure 7.14. Performance of Iris Classification: x-axis and y-axis represents y_0 and y_1, respectively, that corresponds to the cells in the output layer. The dotted center line was added afterwards.

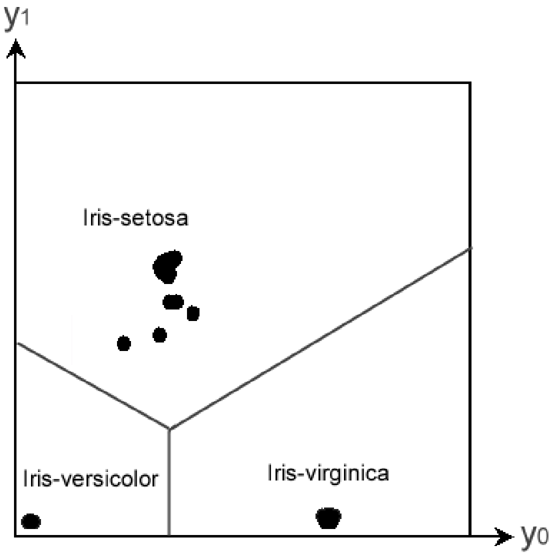


Figure 7.15. Generalization ability of the model in Iris Classification: Displaying the Voronoi diagram (lines) with regard to the spot of each class gives an indication of the class.

Figure 7.15 shows the mapped objects in the two dimensional output space after the 10,000 iterations. The objects are clearly classified into three classes, and the center point of each class (called *spot*) forms a regular triangle.

In order to verify the generalization ability, experiments with test data that are not used for the training will be carried out. 75 test data are mapped onto the output space as illustrated in **Figure 7.17**. Displaying the Voronoi diagram with regard to the spot of each class gives an indication of the class. The Voronoi diagram has the property that for each spot, every point in the region around that spot is closer to that spot than to any of the other spots. Although the network does not give the identification of the class, the aspect of classification can be displayed with the aid of the diagram.

7.6.2 Facial expression classification

The proposed model is applied to facial expression classification.

A line drawing model [Yamada, 1993b] of the facial expression and five facial images of typical emotional categories such as *happiness, anger, fear, surprise* and *sadness* are used for the classification problem. The learning object is the same figures as **Figure 7.9**.

In the learning phase, a pair from 50 instances (contains evenly five classes) is given for the training. The scaling parameter ρ is set to 0.6. The number of cells in the input, hidden and output layers are 9, 12 and 2, respectively. The learning rate is set to 0.3, and the momentum is set to 0.2. **Figure 7.16** shows the mapped objects in the 2 dimensional output space after the 30,000 iterations. It can be seen that the objects are classified into five classes, and the spot of each class forms a deformed pentagon. In case that the number of spot becomes more than four, every spot that have a constant distance to other spots cannot be arranged on 2 dimensional space. Therefore, the training of the network is terminated when the number of iteration steps becomes 30,000 in this experiment.

Figure 7.17 illustrates 50 test data that are mapped onto the output space. The Voronoi diagram is also displayed as well as the previous experiment. The result proves that the trained network holds a non-linear mapping from 9 attributes of facial expression model to 2 dimensional output description space.

7.7 Modeling of Individual Perception

An attempt to create an individual model for each subject is described in this section. The difference in semantic space among subjects has been investigated. As an evaluation of the impression of facial images, the data is

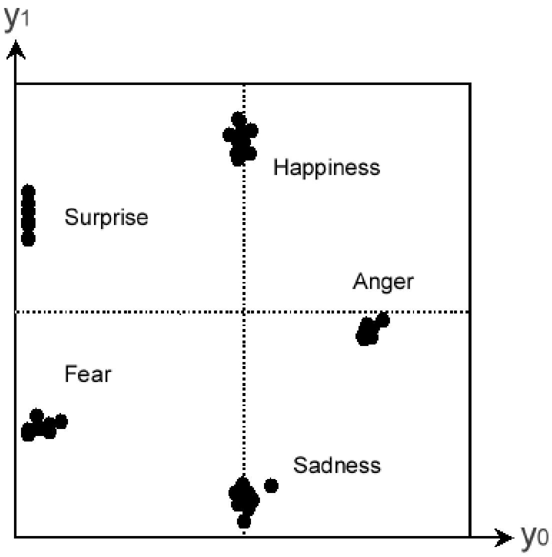


Figure 7.16. Performance of facial expression classification: Displaying the Voronoi diagram (lines) with regard to the spot of each class gives an indication of the class.

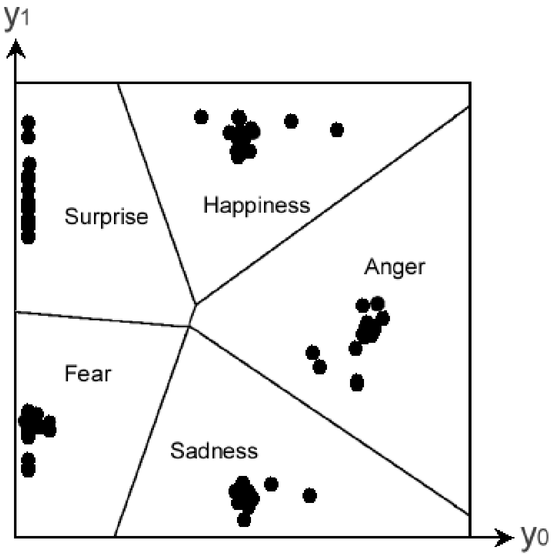


Figure 7.17. Generalization ability of the model in facial expression classification: x-axis and y-axis represents y_0 and y_1, respectively, that corresponds to the cells in the output layer. The dotted center line was added afterwards.

Configuration	Setting
Parameters	
Learning parameter α	0.3
Momentum η	0.2
Scaling parameter ρ	0.6
Number of Cells	
Input Layer	9
Hidden Layer	12
Output Layer	2
Learning	
Iteration	30,000
Instance	50
Test Data	50

Table 7.4. Parameter configuration in the experiment for Facial Expression Perception. The result is illustrated in **Figure 7.16**.

obtained by means of a pair comparison method. On showing a pair of facial images from six fundamental emotional facial images as shown **Figure 7.9**, each subject marks the score for each pair in the seven-scaled semantic differential method (dissimilarity, 1: most similar, 7: dissimilar). The obtained data are normalized to 0-1 values as the teacher signal. The total number of test sets is 15 ($_6C_2$). 30 subjects (23 male / 7 female) are mostly graduate/undergraduate students (21-32 age) in our laboratory.

The numbers of neurons in the hidden and output layer of networks A and B are set to 12 and 2, respectively. The learning parameter is 0.4, and the moment is 0.3. When the error is smaller than 1×10^{-5}, the training of the network is terminated.

Figure 7.18 shows the difference in individual semantic space. In Figure (a) three arrangements, of which one is the mean score of the subjects, and the other two are examples that have representative forms as shown in Figures (b) and (c) are illustrated. Each letter corresponds to one in **Figure 7.9**. A similar trend can be seen with respect to the arrangement of facial images except for the coordinate H (*Disgust*). This result supported the belief of that the facial expression of *Disgust* is one of the ambiguous faces from the psychological point of view. A certain measure of relationship between the physical features of the face and its impression can be seen in addition to subjective characteristics.

The mapping aspect demonstrates individual differences. For example, comparing between Figures (b) and (c), although the difference in the arrangement is just face H, the mapping aspect is quite different. In this manner, visualizing the mapping aspect allows us to have a holistic interpretation in addition to the evaluation of each facial expression.

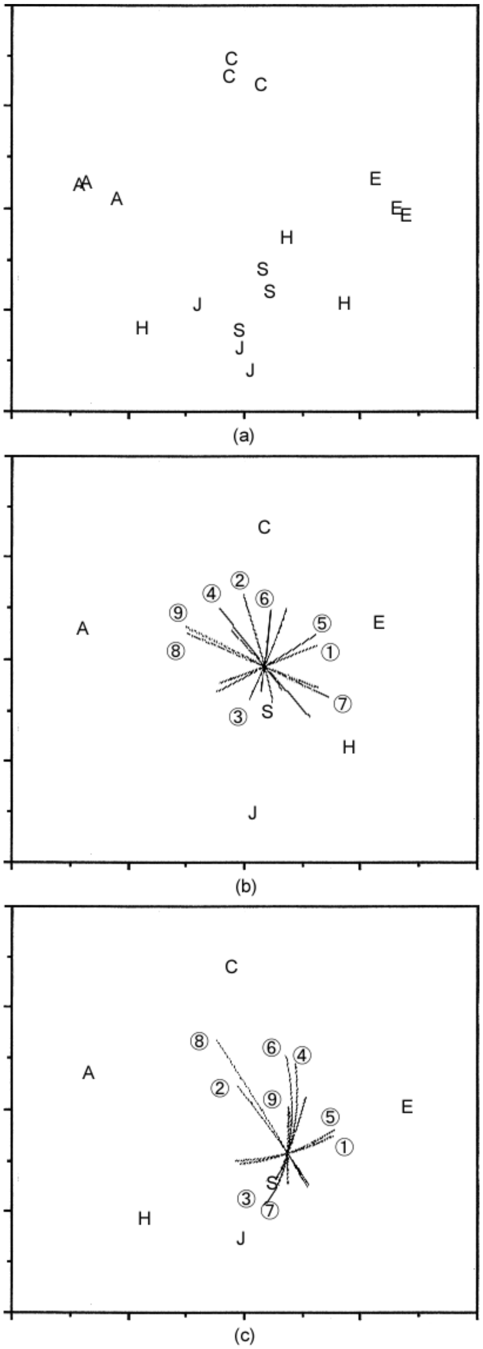


Figure 7.18. Comparing obtained semantic spaces of individual subjects: (a) combined three arrangements; the mean result of all subjects and two distinguishing subjects whose results are illustrated in (b)(c) with the mapped axes.

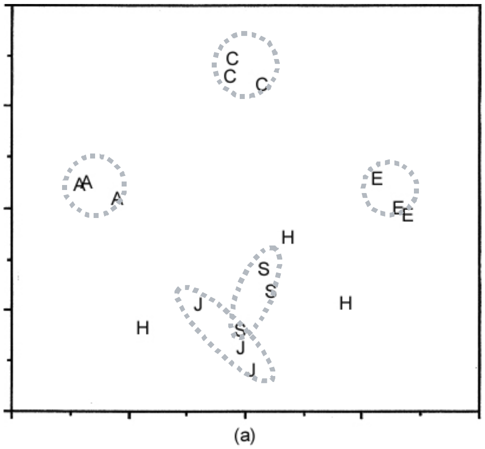


Figure 7.19. The mapping aspect demonstrates individual differences: the three obtained map are overlapped in this figure. A similar trend can be seen with respect to the arrangement of facial images.

7.8 Discussion

With regard to the structure of the proposed neural network, the proposed distance-based learning has some distinguishing properties compared with conventional multi-layer perceptrons. The convergence is relatively stable, but the converged arrangement often depends upon the initial conditions of weights due to the characteristics of the network. Although the Euclidean distance is given as a teacher signal in the present work, the algorithm allows any other distance metric such as norm distance on Mahalanobis' generalized distance. The influence of non-Euclidean distance functions is also considered. In the framework of the proposed network model, nonlinear mapping obtained by network A or B is constrained by the distance metric in unit C. The analysis of the obtained manifold is one further consideration. The issue is that unit C is replaced with a multilayer perceptron so that the metric can be trained for given training sets.

The relationship between the number of training sets and errors is illustrated in **Figure 7.20**. In accordance with the number of training sets, the mean squared error increases since the output space is limited to a two-dimensional space. The figure thus shows the reliability of the output dimensions. As the result of the same experiments with different learning sets, a similar trend can be seen in any case.

Also, a method of multiclass classification by utilizing Distance Mapping Learning network is described in this section. The notable point is the method to obtain a nonlinear mapping between input data and the description parameter by using a neural network model in which the desired distances between the input pair are given as a teacher signal set. The novel

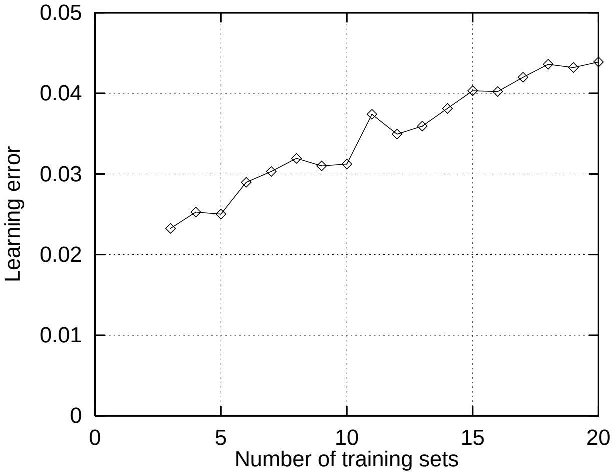


Figure 7.20. Relationship between learning error and number of training sets: The learning set to construct the semantic space is the same as that in **Figure 7.9** in section *7.5.2*.

aspect of the proposed method is that the number of class is not needed for the classification.

Further consideration is to apply the proposed method to the categorical perception of facial expressions [Etcoff et al., 1992]. The analytical evaluation of the network is one of the future issues. Although the Euclidean distance is given as a teacher signal in the present work, the algorithm allows any other distance metric such as Hamming distance, city-block distance and Minkowsky distance. The influence of non-Euclidean distance functions will be also considered.

7.9 Conclusion

The interrelation between the physical features of a facial expression and its impression is realized by obtaining both physical-to-semantic and semantic-to-physical mappings. A method to obtain a nonlinear mapping between physical and emotional quantity has been introduced using a neural network model in which the semantic distances between input data are given as a teacher signal set.

Also, an application to display an artificial facial expression and a semantic space has been developed. The system allows users to modify the facial expression by a mouse operation and to show its correspondence in the semantic space.

The proposed method has been proved to deal with a real image set with the same approach. The categorical perception of facial expressions is one of further considerations; for example, this is discussed in [Etcoff et al., 1992].

Of future consideration is the proposed system for judging human facial expressions, automatically extracting the particular feature parameters of the face. In addition, the acquired bi-directional correspondences between physical and semantic spaces can be applied to the semantic recognition and generation of facial expression.

Chapter 8

Conclusion

8.1 Summary of Results

The author states the arguments of these studies made in this dissertation with some keywords.

Development of a Humanoid Robot iSHA

- These studies are much stronger with actual experimental results.
- By differentiating the processing between physical and intelligent interactions in the mechanical level, the experiment showed the emergence of a sophisticated and integrated type of behavior.
- The control command for the robot comes from two lines, which means providing two independent operating systems inside the robot. This characteristic is very important from a safety point of view.
- A time trace showing activation levels of different behaviors, annotated with verbal descriptions of human stimuli.

Music-Based Human-Robot Interaction

- Three styles of interactions have been investigated, based on the modeling of the human-machine-environment. These studies are also much stronger with actual experimental results.
- Since users can provide intention to a robot with his actions, a new style of possible music generation can be provided. The embodied interaction between humans and the robot has opened the next stage of human-machine collaborative musical performance.
- The MIDI Network allows users to easily associate the relationship between the input and output modules, not only for music generation but also behavior coordination.
- Sub-systems inside the robot perform in parallel with multimodality in the real and continuous worlds.

Kansei Quantization

- A new method of Kansei quantization is realized by the model of distance mapping learning. The proposed method extended the multidimensional scaling method scheme.
- By applying an individual data set on impression to the developed system, visualization of the nonlinear mapping can be used for the measurement of differences among individuality and characteristics common to all.
- New data that are not used in the training in the network can also be evaluated by the generalization ability of the network.

8.2 Contributions

The contributions made in this dissertation are summarized as follows.

Intuitive Robot Operationality: The developed robots described in this dissertation has been presented to the public at a number of demonstrations. In various scenes, people successfully interacted with the robot without any pre-knowledge of operating them. It is considered that the intuitive control of the robot have contributed to this operationality criterion. For instance, shaking hands and reactions to sounds are sufficient for users to easily and intuitively interact with the robot.

Diversity and Redundancy: With regard to the interaction with the robot *iSHA*, people can express his/her intentions to the robot in different ways. For example, there are various ways to draw its attention: 1) handclaps (making sounds), 2) Showing an object, 3) By speech, simply say "turn left/right, please." 4) Grasping its hand and pulling it closer, and 5) Pushing the body and making it turn around.

People can choose these ways according to the purpose of use and scene setting. This is an important requirement for realization of natural interaction. When one wanted to lead the robot to a specified location, the easiest way is to take it by the hand, or to force it to there by pushing its body. However, there may be a situation where one has something in their hands, and it is difficult to accomplish the given task with physical contact. In that case, one may need to operate the robot by speech or sounds. In this way, diversity and redundancy are constitutive and important features in multimodal interaction, which enables the machine to achieve a given task by different ways through multimodality, despite an easier way to achieve the given task.

Safety Factors: In case of errors in the installed control mechanisms in the operating system, or unexpected events occur, it is effective and true to stop its motion with physical contact by hand. The layered architecture

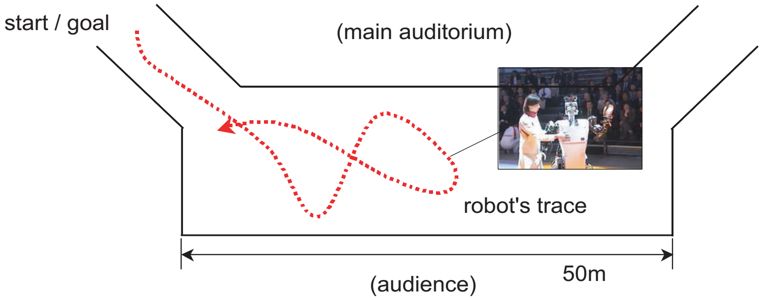


Figure 8.1. Robodex2002 demonstration. (Pacifico Yokohama, from Mar. 27th - Mar. 31st, 2002. *iSHA* appeared in *Robot Parade* during the exhibition.

enables people who interact with the robot to limit activity in the robot while other types of behavior are operative . Also, when the robot failed to track the object using its head, humans can help the robot to find it by turning it around with physical contact. Considering that the robot has multiple operating systems and control paths, the robustness of the system is enhanced.

Robustness and Dependability: The humanoid robot *iSHA* was successfully demonstrated at a special exhibition, *Robodex2002*, with a human partner. In fact, it had with only 1 failure, the embedded battery died during the demonstration. **Figure 8.1** illustrates the demonstration stage. The dashed arrow is a trace of the robot, where a human partner brought the robot around by pushing it to turn around, and grasping and pulling its hand to lead it to a specific location. She was a person untrained in robot control before. However, after being lectured for a few minutes, she successfully performed the control, rather ”lead” the robot. Other robots on the stage were not led by any human partners, but operated by remote-control or behaved in preprogrammed way. Sufficient evidence of the robustness and dependability of the robot are seen in the experimental results.

Synthetic Approach to Kansei Interaction: The effective mobile robot platforms in multimodal artistic environment for music and dance performance have been introduced. The proposed approach to equip musical instruments with an autonomous mobile ability will provide a new computer music performance in the real world. The developed system enables one to reflect environmental visual and auditory information around the human and the robot for the creative and dynamic performance. Since the users can provide their intention to robot by action, a new style of possible music generation can be provided. The author considers that the system has the capability of creation in the virtual world to extend robot control in the real world.

New Method of Kansei Measurement: This approach utilizing non-linear mapping delivered good results in terms of fitting rate compared with the conventional statistical analysis. In addition, new data which are not used in the training in the network can also be evaluated by the generalization ability of the network. It is considered that the proposed method can extend the scheme of the multidimensional scaling method. Also by applying an individual data set on impression to the system, visualization of the nonlinear mapping is used at an early stage to measure differences among individuality and characteristics common to all.

8.3 Further Consideration

The author concludes by addressing the further consideration with the following four arguments.

Machine Vision Many studies on a robot vision system have been emphasized, which aims at gaining a visual feedback for robot control such as object recognition in a 2D or 3D environment, and creating an environmental map surrounding the robot, for example [Ayache et al., 1987] [Jarvis et al., 1998]. The author became interested in the interaction metaphors such as a multimodal interaction system including emotional affects in interactive dance/music systems. A method to extract *emotional* information from human gestures in real time is one of further considerations with regard to machine vision research.

Some of the key concepts found in the exploration of human motion are taken from Rudolf Laban's (1879-1958) theoretical studies [Laban et al., 1947] [Laban, 1963]. One of the most important approaches concerns his theory of movement called Theory of Effort.

Laban introduced the notation method known as "Labanotation". By using this notation, it is possible to record a variety of human motions. It should be noted that the Labanotation is different from other similar studies. (ex. Benesh Notation based on classical ballet). Labanotation is not limited to a singular, specific style of dance but concerns every kind of human motion.

With regard to Laban's work and his theory, from a scientific point of view, several attractive studies have been done, for example, a platform of human-robot interaction has been developed in order to apply Laban's theory to the movement of the mobile robot [Nakata et al., 1998].

The focus differs from other existing systems that pay attention to gesture recognition from a gesture vocabulary. The focusing point is the difference in intentions between two performances of the same segment.

The motion analysis is involved in mapping physical parameters onto emotional information, namely Kansei analysis of a dance performance. The author is trying to realize a sort of "Kansei extractor" as the last goal by extracted emotional information in a dance performance with the aid of a camera-based computational system and based on Laban's theory of movement. (see [Camurri et al., 2000]) People can understand the emotional mode, such as happy, angry and sad, and in which mode the actor performs. It seems that human movement carries emotional information although it is difficult to express the emotional effects with particular physical parameters. From this point of view, the extraction of emotional information in a dance performance with the aid of a computer system based on Laban's theory of movement is one needing further consideration.

Machine Listening Until today, many researchers have emphasized the acoustic perception by machines. The term "machine listening" refers to the ability to self-generate musical preference according to the given audio signals. As a primary attempt of constructing a machine listening system, the author has contributed to the development of an artificial music listening system that can understand an audio sound sequence as music [Suzuki et al., 2002].

The music composition and performance by human is done as a result of the repetition of creation and evaluation. Therefore, in the autonomous music creation and performance by a machine, the ability of machine listening is essentially required for the self-evaluation of music.

In case of music creation by humans, his/her musical preference may be highly crucial in creating his/her own music. However, the unbiased evaluation of the created music is quite difficult as the musical preference is different from one to another. In the present study, assumed that the musical preference is built up through the listener's musical experiences. If a person is familiar with western style music, his/her own musical preference is based on the rhythm, tone and many other musical features of western style.

The proposed machine listening system can learn from music without teacher using multiple neural networks. The system has the ability to establish its own music theory from the experiences of listening to the raw sound without any musical knowledge given in advance. Note that the proposed system can accumulate these features by learning using neural networks. The details of these features and the process of extraction are described in the rest of this section.

In general, the analysis of the audio stream is assumed to be a spectral-temporal pattern analysis by a dynamic clustering. First, the system discriminates streams from the acoustical sequence of sound. The temporal structure of the stream such as tempo and rhythm is extracted in parallel. The relationship between the tempo and separated stream of sound is then integrated.

Machine Initiative With regard to the judgment, initiative is the ability to use one's judgment to make decisions and do things without needing to be told what to do. In addition, initiative is also regarded as the first action or movement, often intended to solve a problem. If one has/seizes the initiative, one has/takes power and is able to control events. If one loses the initiative, one no longer has the position of control or power that one had before. In any case, the term of initiative is used in the scheme of communication or interaction.

In communication between people and robots, people always take the initiative until today. People usually control and govern the robot. Future robots are required to make decisions in the collaborative works with people. At that time, initiative plays a very important role because the initiative transfer is one of typical human-human communication. Therefore, the development of a machine that takes, rather tries to take, initiative is one of further considerations.

For example, in the musical performance, it is natural that players exchange the initiative of performance with each other, without spoiling the musical harmony. The author has contributed to a novel human-machine interface system which allows the smooth initiative exchange between human and machine during a performance [Taki et al., 2000]. Another advantage of the developed system from an artistic point of view is that the initiative transfer between the human and the system can realize a performance with an unexpected and surprising style to stimulate the human creativity by providing novel ideas for the music performance and music composition.

Machine Awareness In this dissertation, the hierarchical structure of robot control provides a principle condition that the behavior of the robot does not correspond to the intelligent process by the robot itself, because physical signal processing is mechanically separated from that and the behavior emerged by the integration between these two processes. This aspect is quite important and essential for the emergence of behavior.

The robot's behavior and its range is given by some explicit rules, and the robot behaves within this frame. This is the underlying problem for robot control. In recent years, researchers in the intelligent robot field have emphasized the behavior acquirement by a visual feedback system or learning. These are largely aimed at this acquirement for a given task. On the other side, a behavior acquirement by imitating human motions has been an active topic in the robotics field. However, it is considered that there are passive methods of behavior acquirement.

It is still difficult for the robot to perform in the real world, where unexpected events, disturbances and interruptions occur. Moreover, since the robot does not have the ability to determine the range of its own action, and unexpected trouble and failure will cause crucial problem.

Based on the above consideration, the robot should have a somatic sense to measure its own body conditions, especially pain. If the robot can measure its own pain, it can determine the range of action by itself. The pain is divided into two areas: one deals with the joint parts consisting of the actuator and gear. The other deals with the non-joint parts caused by physical contact.

This research aims to formalize the pain occurring in the robot body, and to associate it among the control command, and the data of physical contact and physical constraints. In the hierarchical architecture, a module to sense the pain can be embedded in the physical layer, and the learning module that holds the condition, e.g., over current, inverse electromotive force, the angle of the joint, the strain of the joint, and physical contact data around the joint, can be embedded into the intelligent layer. These modules thus work to avoid the situation in which the pain arises in the joint part.

The proposed mechanism allows the robot not only to have a self-calibration ability but also to obtain knowledge about a prior risk avoidance. Note that this is an active method of behavior acquirement and determination of a range of action through the interaction between the robot and the environment.

Pain is one of the principles of action in the law of nature. Acquiring the range of its own action by itself allows the ability of self-preservation. This research provides important knowledge about the mechanism of self-preservation, which is the fundamental ability of autonomous systems.

Appendix A

iSHA Configuration

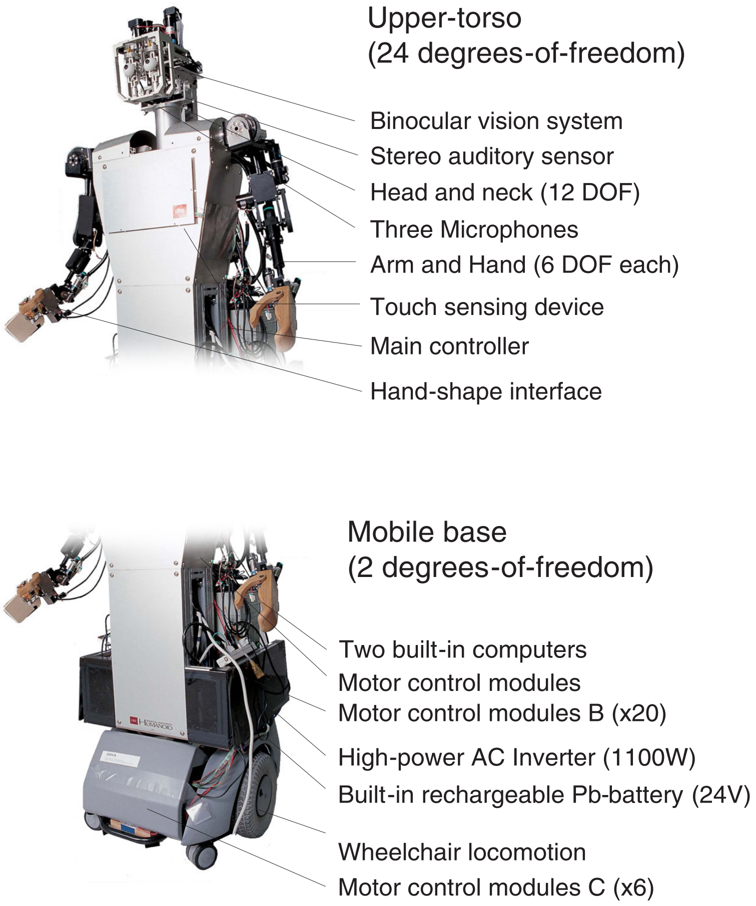


Figure A.1. Autonomous humanoid robot "iSHA". The upper torso resembling a human in shape, with a head and two arms has 24 degrees-of-freedom. The lower base with two wheels equipped under the body that provide safe and robust locomotion.

Body parts	Number of actuators	Motor vender
Eye ball	DC servo × 4 (each 2)	Maxon motor
Head	DC servo × 4	Harmonic Drive
Shoulder	AC servo × 4 (each 2)	Panasonic MINAS
	DC servo × 2 (each 1)	Harmonic Drive
Arm	DC servo × 4 (each 3)	Harmonic Drive
Hand/wrist	DC servo × 4 (each 2)	Harmonic Drive
Finger	DC servo × 2 (each 1)	Escap
Foot (wheels)	DC servo × 2	Hitachi Car Eng.

Table A.1. Mechanical configuration of Humanoid robot *iSHA*: Total number of the actuators is 26; 24 DOFs for the upper body and 2 DOFs for the lower base.

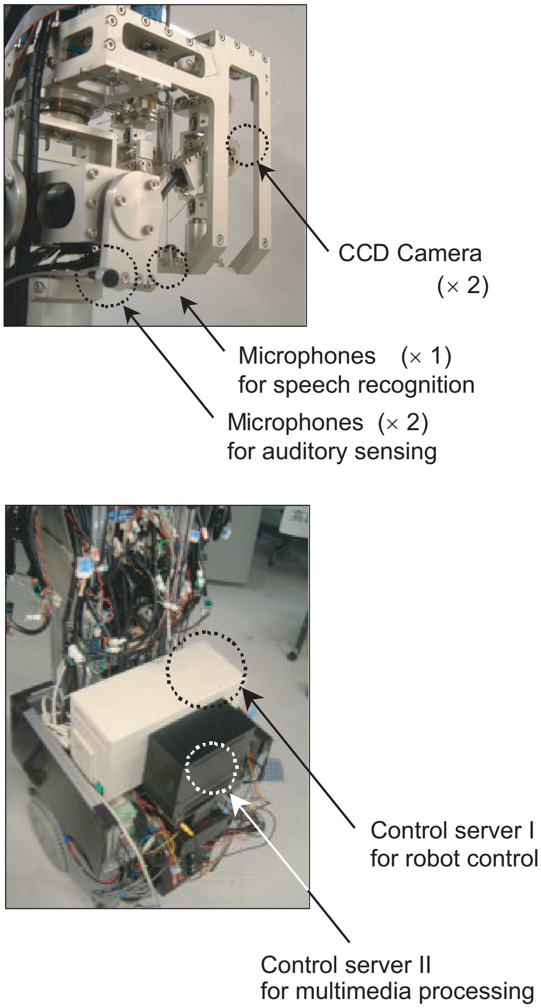

CCD Camera (× 2)
Microphones (× 1) for speech recognition
Microphones (× 2) for auditory sensing
Control server I for robot control
Control server II for multimedia processing

Figure A.2. The head system and control system of "iSHA": Each eye equipped with a small CCD camera, small three microphones embedded in the head.

Body parts	Installed items
Control Server I	ICP Electronics (PAC-107 Type)
	(AMD K6-II 400MHz, 128MB RAM)
Operating System	ART-Linux ver. 2.1.9 (for robot control)
Data Acquisition	Adlink PCI-6216 ×2
	(16-ch Analog Output Cards)
	32-ch ISA Counter Board (custom-built)
Control Server II	ICP Electronics (PAC-53H Type)
	(Pentium III 800MHz, 256MB RAM)
Operating System	Windows2000 (for multimedia processing)
Data Acquisition	Hitachi IP-5000
	Image Processing Board
CCD Camera	Tokyo Electronic Industry CS6100 ×2
Microphones	Small capacitor microphone ×3
Speaker System	Sanei House CEMI (ceramics speaker system)
Battery	Lead storage battery 24V24Ah ×2
High Power Inverter	Exeltech XP1100
	(DC24V input 1100W true sin wave output)
Uninterruptible Power Supply (UPS)	Sanwa Supply UPS-500DE
Motor Driver	Panasonic MINAS AC servo driver ×4
	TITECH Ver. 1 (PC-0121-1) ×16
	TITECH Ver. 4 (PC-0144-1) ×6
Wireless LAN	Allied Telesys
	CentreCOM WR11Mbps (IEEE802.11b)
Video Transmitter	RF Systems lab.

Table A.2. Hardware configuration of Humanoid robot *iSHA*: Most listed instruments are commercially available. Some of them are custom-built instruments.

Appendix B

Laban's Theory of Effort

Some of the key concepts in the exploration of human motion intention are taken from Rudolf Laban's work [Laban et al., 1947, Laban, 1963]. In his theory of effort, he pointed out the dynamic nature of movement and the relationship among movement, space, and time. Laban's approach is an attempt to describe, in a formalized way, the characteristics of human movement without focusing on a particular kind of movement or dance expression. Effort theory principles can be applied to dance and to everyday work practices.

The basic concept of Laban's theory is effort that is a property of movement. From an engineering point of view, it is considered that a vector of parameters that identifies the quality of a movement performance. The most important note is the description of the quality of movement. Theory of effort is not concerned with the degrees of joint rotation or moment directly, but it considers movement as a communication media and tries to extract parameters related to the its expressive power. During a movement performance the vector describing the motion quality varies in effort space. Laban studies the possible paths followed by this vector and the intentions they can express. Therefore, variations of effort during the movement performance should be studied.

Effort Space Laban indicates 4 components that generate what is called "effort space": space, weight, time and flow. Each component is measured on a bipolar scale, in this way every component of effort space can have binary values to represent opposite quality.

Laban's basic theory considers the first 3 factors to develop a description system for human movement.

In this way 8 possible combinations of the space can be identified; time and weight factors, corresponding to states that the movement can assume in its developing.

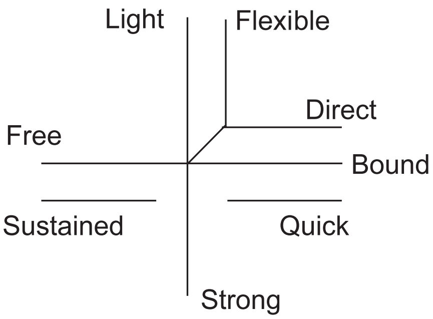


Figure B.1. Effort Space with symbol expression, which is composed of space, weight, time and flow.

Axes	Indulging Effort	Fighting Effort
Space	Direct	Flexible
Time	Sustained	Quick
Weight	Light	Strong
Flow	Free	Bound

Table B.1. Efforts table by Laban: a verbal description of the correspondence between effort and movement.

Space: Regarding space, Laban says "... whenever the body moves or stands, it is surrounded by space. Around the body is the sphere of movement, or *Kinesphere*, the circumference of which can be reached by normally extended limbs without changing one's stance, that is, the place of support..." [Laban, 1963, p. 85]. The *Kinesphere* is also referred to as personal space, while the whole space surrounding the *Kinesphere* (i.e., the environment in which the act of movement is taking place) is referred to as general space. When the body moves in space the *Kinesphere* follows it, so the study of movement can be divided in two main branches: the movement of the *Kinesphere* in general space and the movement of the limbs inside the *Kinesphere*. The approach follows this method. A movement, in both kind of spaces, will follow a definite direction or a sequence of different directions. If the movement follows those directions smoothly the space component in effort space will be flexible, while if it follows them straightly it will be marked as direct.

Time: Laban considers two aspects of time: an action can be sudden or sustained, which allows the binary description of the time component of effort space. Moreover, in a sequence of movements, each of them has a duration in time, the ratio of the duration of following movements gives the time-rhythm, as in a music score.

Weight: Weight is a measure of how much strength and weight is present in a movement, so in pushing away an heavy object it will be necessary to use body weight in order to succeed.

Flow: Flow is a measure of how bound or free appears a movement or a sequence of movements.

The author has contributed to a study of attempting to extract Kansei information directly from basic physical properties of movement (Investigating personal space), and to find a symbolic representation of the qualities of movement suitable for Kansei analysis (Investigating general space) based on the Laban's basic theory from the engineering point of view [Camurri et al., 1999]. The developed system is able to locate the stretches in space movement, which is the first important step in classifying movement using Laban's approach. Future work will develop a method to detect the space component of effort while performing a stretch. Very important expressive information can be obtained by the way in which paths are followed in space. A direct movement has got different content from a flexible, round movement performed to reach the same target point.

On the other hand, the work will also go toward the direction of improving the process of direct extraction of Kansei and developing more effective symbolic descriptions of movement performances following Laban's approach. However, one important point showed by the work is that information is carried not only by the state of a set of observed variables, but by their change in time, so the rhythm of variations during time is and will be a central part of the study.

The study to better classify the emotional impact of a live performance of dancers. This means that the study of the Kansei of movement will try to provide a high level description of the dancer performance, modulated by a factor that is a function of the spatial position in which the movement is performed.

Acknowledgements

The present work is a collection of studies which have been carried out under the guidance of Professor Shuji Hashimoto at the Graduate School of Science and Engineering, Waseda University. I have no words with which to thank Professor Shuji Hashimoto for leading me toward this insight. The range of his thinking, expertise, and boundless interests have always surprised me. He indeed has provided a wealth of inspiration, encouragement and suggestions. Needless to say, this thesis work would not have been realized without his guidance and continuous support.

I acknowledge Professor Antonio Camurri of the University of Genoa for his splendid support, great help and warm treatment during my stay in Italy. During the various stages of my study, he provided me with many important opportunities such as exhibitions and a public seminar.

I would like to acknowledge the members of the committee, Professor Tomiji Hisamura, Professor Hirochika Nakajima, Professor Atsuo Takanishi and Professor Atsushi Takeuchi for their kindness in providing me with helpful comments and their time for reviewing this dissertation.

This thesis work was supported in part by a grant from the research fellowships of the Japan Society for the Promotion of Science (JSPS) for Young Scientists (JSPS-H12-DC1-05931), and also partly supported by the JSPS Research for the Future Program in the Area of Kansei (Intuitive & Affective) Human Interfaces, and by a grant-in-aid from the Japanese Ministry of Education, Science, Sports and Culture (A-11305021). I also thank a lot of many organizations for their partial support of our robot demonstrations.

My great appreciation goes to Professor Atsuo Takanishi and Professor Hideaki Takanobu for the mechanical design of the humanoid robot *iSHA*. Also, for two performers, Mr. Tatsuya Hashimoto and Ms. Eri Nomura that showed their remarkable talents during the demonstration of ”*iDance*” and ”*MIDItro*”. The sensuous music was composed by Mr. Riccardo Dapelo for both of exhibitions ”*Il giardino della musica*” and ”*Arti Visive 2*”. The scenography intervention (”dress”) on the robot in the ”*Arti Visive 2*” exhibition was developed by Ms. Emanuela Pischedda. This thesis work could

not have done without the great help by these wonderful scientists and artists.

With my whole heart, I would like to express my heartfelt thanks to Dr. Pitoyo Hartono and Dr. Yoshimitsu Aoki who spent countless cheerful talks and tiny works with me. These times with them are indeed unforgettable. Also, my sincere thanks to the best friend, Dott. Riccardo Trocca, for not only his collaboration in the dance analysis project, but also for the many joyful times we shared. He left me with a lot of food for thought.

I would like to thank the secretary Ms. Hisako Ohta for helping with the accounting procedures. Because she kept me away from the extremely complex process of accounting, I could then concentrate on my studies.

I had the good fortune to be involved in our laboratory, SHALAB, for many years during my bachelor and master courses, and doctoral study. Especially, the fruitful discussions, critics and suggestions with the members of the Neural Network Group have encouraged me to achieve this thesis work. I would like to thank the members of the Robotics Group who shared the everlasting cycle of the assembly, programming and debugging in the development of robots that appear in this dissertation.

Also, I am deeply grateful to all members of the Laboratorio di Informatica Musicale, InfoMus, for their kind support and increasing my skill of language by wonderful talks, especially for Dott. Massimiliano Di Stefano and Dott. Alessandro Stroscio for their continuous help in developing the motion detection system.

I thoroughly enjoyed the opportunity of studying there with all others at these two laboratories, who all made great contributions to this thesis. For all the colleagues and friends in Japan, Italy and any other country, I extend my thanks for their support and help over the years. Although the names are not shown, everybody who has known me has made a contribution to this thesis.

I would like to thank my parents who supported me with the freedom of choice for a long time. Talking with my brother gives me new findings and encouraged my motivations. What has been learned from them is the attitude for life. I am really fortunate to be raised in a happy family.

Finally, there is no way to put into words as to how I express my sincere gratitude and appreciation for my dearest wife, Yukiko. I could not have done, and will not do anything without you.

References

[Ahalt et al., 1990] Ahalt, S. C., Krishnamurthy, A. K., Chen, P., and Melton, D. E., "Competitive learning algorithm for vector quantization," *Neural Networks*, Vol.3, pp. 277-290, 1990.

[Anzai, 1993] Anzai, Y., "Human-computer-interaction in multiagent environments," *Proc. of HCI International 1993*, pp. 2–7, New York, Elsevier, 1993.

[Arkin, 1998] Arkin, R., *Behavior-Based Robotics*, MIT Press, Cambridge, MA, United States, 1998.

[Ayache et al., 1987] Ayache, N., Faugeras, O. D., "Building, registrating and fusing noisy visual maps," *Proc. of 1st International Conference Computer Vision*, pp. 73-82, 1987.

[Bates et al., 1992] Bates, J., Loyall, A.B., Reilly,W.S., "An architecture for action, emotion and social behavior," *Proc. of the Fourth European Workshop on Modeling Autonomous Agents in a Multi-Agents World*, S. Martino al Cimino, Italy, 1992.

[Biehl et al., 1997] Biehl, M., Matsumoto, D., Ekman, P. et al., "Matsumoto and Ekman's Japanese and Caucasian Facial Expressions of Emotion (JACFEE): Reliability data and cross-national differences," *Journal of Nonverbal Behavior*, 21(1), 3–21, 1997.

[Birk, 1996] Birk, A., "Learning to survive," *Proc. of the 5th European Workshop on Learning Robots*, Klingspor, V. (ed.), 1996.

[Bischoff et al., 2002] Bischoff, R., Graefe., V., "Dependable Multimodal Communication and Interaction with Robotic Assistants," *Proc. of IEEE International Workshop on Robot and Human Interactive Communication*, Berlin, pp. 300-305, 2002.

[Bishop, 1995] Bishop, C. M., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[Breazeal et al., 2000] Breazeal, C. and Scassellati, B., "Infant-like Social Interactions between a Robot and a Human Caregiver," *Adaptive Behavior*, 8 (1), pp. 49–74, 2000.

[Brooks, 1986] Brooks, R., "A Robust Layered Control System for a Mobile Robot," *IEEE Journal of Robotics and Automation*, vol. RA-2, no. 1, pp. 14–23, 1986.

[Brooks, 1991] Brooks, R., "Intelligence without reason," *Proc. of the International Joint Conference on Artificial Intelligence.*, San Mateo, Morgan-Kaufman, pp. 569–595, 1991.

[Brooks, 1997] Brooks, R., "Evolutionary Robotics; Where From and Where To," *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, pp. 1–19, 1997.

[Brooks et al., 1999] Brooks, R., Breazeal, C., Marjanovic, M., Scassellati, B. and Williamson, M., "The Cog project: Building a humanoid robot," *Computation for Metaphors, Analogy and Agents, C. L. Nehaniv, editor, Lecture Notes in Artificial Intelligence*, vol. 1562, Springer-Verlag, 1999.

[Charwat, 1992] Charwat., H. J., *Lexikon der Mensch-Maschine-Kommunikation (Encyclopedia of man-machine communication)*, Oldenbourg, 1992.

[Camurri, 1995] Camurri, A., "Interactive Dance/Music Systems," *Proc. of International Computer Music Conference*, 1995.

[Camurri et al., 1997] Camurri, A., Ferrentino, P. and Dapelo, R., "A computational model of artificial emotions," *Proc. of Intl. Workshop on KANSEI - Technology of emotion*, pp.16-23, 1997.

[Camurri et al., 1998] Camurri, A. and Coglio, A., "An Architecture for Emotional Agents," *IEEE Multimedia*, October - December 1998, pp. 24-33, 1998.

[Camurri et al., 1999] Camurri, A., Hashimoto, S., Ricchetti, M., Suzuki, K., Trocca, R., Volpe, G., "Kansei Analysis of Dance Performance," *Proc. of IEEE International Conference on System, Man and Cybernetics*, Tokyo, Japan, pp. IV- 327-332, 1999.

[Camurri et al., 2000] Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A., Suzuki, K., Trocca, R. and Volpe, G., "EyesWeb - Toward Gesture and Affect Recognition in Dance/Music Interactive Systems," *Computer Music Journal*, MIT Press, vol. 24, No. 1, pp. 57–69, 2000.

[Cheng et al., 2000a] Cheng, G. and Kuniyoshi, Y., "Complex Continuous Meaningful Humanoid Interaction: A Multi Sensory-Cue Based Approach," *Proc. of IEEE International Conference on Robotics and Automation*, vol. 3, pp. 2235–2242, United States, 2000.

[Cheng et al., 2000b] Cheng, G., Nagakubo, A. and Kuniyoshi, Y., "Continuous Humanoid Interaction: An Integrated Perspective - Gaining Adaptivity, Redundancy, Flexibility - In One," *Proc. of First IEEE-RAS International Conference on Humanoid Robots*, USA, 2000.

[Chung et al., 1991] Chung, J., Gershenfeld, N. and Norris, M.A., "A Development Environment for String Hyperinstrument," *Proc. of the International Computer Music Conference*, pp. 150-152, 1991.

[Cook et al., 2000] Cook, P. and Leider, C., "SqueezeVox: A New Controller for Vocal Synthesis Models," *Proc. of International Computer Music Conference*, Berlin, 2000.

[Duch, 1997] Duch, W., "Neural minimal distance methods," *Proc. of Third Conference on Neural Networks and Their Applications*, Kule, pp. 183–188, 1997.

[Ekman et al., 1978] Ekman, P. and Friesen, W., "Facial action coding system : a technique for the measurement of facial movement," *Consulting Psychologists Press*, 1978.

[Ekman, 1992] Ekman, P., "Facial expression of emotion," *American Psychologist*, 48, pp. 384–392, 1992.

[Etcoff et al., 1992] Etcoff, N. and Magee, J., "Categorical perception of facial expressions," *Cognition*, 44, pp. 227–240, 1992.

[Eto, 1998] Eto, K., *SoundCreatures*, http://www.canon.co.jp/cast/artlab/scweb/, 1998.

[Fisher, 1936] Fisher, R.A, "The use of multiple measurements in taxonomic problems," *Annual Eugenics*, 7, Part II, pp. 179–188, 1936.

[Fogel, 1999] Fogel, D. B., *Evolutionary Computation : Toward a New Philosophy of Machine Intelligence*, IEEE Press, 1999.

[Flash et al., 1985] Flash, T., and Hogan, N., "The Coordination of Arm Movement; an Experimentally Confirmed Mathematical Model," *Journal of Neuroscience*, 5, pp. 1688–1703, 1985.

[Geiser, 1990] Geiser, G., *Mensch-Maschine Kommunikation*, Oldenbourg, 1990.

[Geman et al., 1992] Geman, S., Bienenstock, E. and Doursat, R. "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, 4, pp. 1–58, 1992.

[Giorsi et al., 1990] Girosi, F. and Poggio, T., "Networks and the Best Approximation Property," *Biological Cybernetics*, Vol. 63, pp. 169–176, 1990.

[Gong et al., 1995] Gong, Y., Hook-Chuan, C. and Xiaoyi, G., "Image Indexing and Retrieval Based on Color Histograms," *Multimedia Modeling*, World Scientific, pp.115-126, 1995.

[Grubb et al., 1997] Grubb, L., and Dannenberg, R. B., "A Stochastic Method of Tracking a Vocal Performer," *Proc. of 1997 International Computer Music Conference*, pp. 301-308, 1997.

[Harvey et al., 1996] Harvey, I., Husbands, P., Cliff, D., Thompson, A., Jakobi, N., "Evolutionary Robotics: the Sussex Approach," *IEEE Journal of Robotics and Autonomous Systems*, 20, pp. 205–224, 1996.

[Hashimoto, 1997] Hashimoto, S., "KANSEI as the Third Target of Information Processing and Related Topics in Japan," *Proc. of AIMI International Workshop on KANSEI - The Technology of Emotion*, Genoa Italy, pp. 101-104, 1997.

[Hashimoto, 1999] Hashimoto, S., "Humanoid Robot for Kansei Communication," *Proc. of the Second International Symposium on Humanoid Robot (HURO99)*, pp. 156–160, Tokyo, 1999.

[Hashimoto et al., 2002] Hashimoto, S. et al., "Humanoid Robots in Waseda University - Hadaly-2 and WABIAN," *Autonomous Robots*, Vol.12, pp. 25–38, 2002.

[Hayashi, 1952] Hayashi, C., "On the prediction of phenomena from mathematicostatistical point of view," *Annuals of the Institute of Statistical Mathematics*, 3, pp. 69–98, 1952.

[Hikiji et al., 2000] Hikiji, R. and Hashimoto, S., "Hand-Shaped Force Interface for Human-Cooperative Mobile Robot," *Proc. of First International Workshop on Haptic Human-Computer Interaction*, pp. 113–118, 2000.

[Hiller et al., 1959] Hiller, L. and Isaacson, L., *Experimental Music*, McGraw-Hill Book Company, Inc., 1959.

[Hollerbach, 1996] Hollerbach, J. M., "Anthropomorphic Robot and Human Interactions," *Proc. of 1st. Intl. Symposium on Humanoid Robots*, 1996.

[Hirai et al., 1998] Hirai, K., Hirose, M., Haikawa, Y. and Takenaka, T., "The Development of Honda Humanoid Robot," *Proc. of the 1998 IEEE International Conference on Robotics and Automation*, pp. 1321–1326, 1998.

[Hiraki et al., 1996] Hiraki, K., and Anzai, Y., "Sharing knowledge with robots," *International Journal of Human-Computer Interaction*, 8 (3), pp. 325–342, 1996.

[Hirose et al., 1993] Hirose, S., Amano, S., "The VUTON: High Payload High Efficiency Holonomic Omni-Directional Vehicle," *Proc. of Intl. Symposium on Robotics Research*, pp. 253–260, 1993.

[Hornik et al., 1989] Hornik, K., Stinchcombe, M. and White, M., "Multilayer Feedforward Networks are Universal Approximators," *Neural Networks*, Vol. 2, No. 5, pp. 359–356, 1989.

[Hornik, 1991] Hornik, K., "Approximation Capabilities of Multilayer Feedforward Networks," *Neural Networks*, Vol. 4, No. 2, pp. 251–257, 1991.

[Hornik, 1993] Hornik, K., "Some New Results on Neural Network Approximation," *Neural Networks*, Vol. 6, No. 8, pp. 1069–1072, 1993.

[Imai et al., 1999] Imai, M., Ono, T., and Etani, T., "Attractive Interface for Human Robot Interaction," *Proceedings of 8th IEEE International Workshop on Robot and Human Communication (ROMAN'99)*, pp. 124–129, 1999.

[Iwata et al., 2001] Iwata, H., Hoshino, H., Morita, T. and Sugano, S., "Force Detectable Surface Covers for Humanoid Robots," *Proc. of IEEE/ASME Int. Conf. on Advanced Intelligent Mechatronics (AIM'01)*, pp.1205-1210, 2001.

[Jain et al., 1995] Jain, R., Kasturi, R. and Schunck, B., *Machine Vision*, McGraw-Hill, 1995.

[Jarvis et al., 1998] Jarvis, R. A. and Byrne, J. C., "An automated guided vehicle with map building and path finding capabilities," *Proc. of 4th International Symposium on Robotics Research*, pp. 497-504, 1998.

[Katayose et al., 1993] Katayose, H., Kanamori, T., Kamei, K., Nagashima, Y., Sato, K., Inokuchi, S. and Shimura, S., "Virtual Performer," *Proc. of the International Computer Music Conference*, pp. 138-145, 1993.

[Katsikitis, 1997] Katsikitis, M., "The classification of facial expressions of emotion: A multidimensional scaling approach," *Perception*, 26, pp. 613–626, 1997.

[Klingspor, 1997] Klingspor, V. and Demiris, J. and Kaiser, M., "Human-Robot-Communication and Machine Learning," *Applied Artificial Intelligence*, Vol. 11 (7/8), 1997.

[Kohonen, 1994] Kohonen, T., *Self-Organizing Maps*, Berlin, Springer Series in Information Science 30, Springer-Verlag, 1994.

[Kohonen, 1990] Kohonen, T., "The self-organizing map," *Proc. of IEEE International Conference on Neural Networks*, vol.78, pp.1464-1480, 1990.

[Konolige et al., 1996] Konolige, K., and Myers, K., *The Saphira Architecture for Autonomous Mobile Robots*, MIT Press, 1996

[Kruskal et al., 1978] Kruskal, J. B. and Wish, M., *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA, United States, 1978.

[Kuniyoshi et al., 1994] Kuniyoshi, Y., Inaba, M., and Inoue, H., Learning by watching: Extracting reusable task knowledge from visual observation of human performance, *IEEE Transactions on Robotics and Automation*, 10(6), pp. 799–822, 1994.

[Laban, 1963] Laban, R., *Modern Educational Dance*, Macdonald & Evans Ltd., London, 1963.

[Laban et al., 1947] Laban, R. and Lawrence, F. C., *Effort*, Macdonald & Evans Ltd., London, 1947.

[Lim et al., 1999] Lim, H., Ishii, A. and Takanishi, A., "Basic Emotional Walking Using a Biped Humanoid Robot," *Proc. of 1999 IEEE International Conference on System, Man and Cybernetics*, Tokyo, pp. 383–389, 1999.

[Lindström et al., 1999] Lindström, L., Kotiaho, J. S., "Signalling and Reception," *Encyclopedia of Life Sciences*, Nature Publishing Group, London, 2002.

[Machover et al., 1989] Machover, T. and Chung, J., "Hyperinstruments : Musically intelligent and interactive performance and creativity systems," *Proc. of the International Computer Music Conference*, pp. 186-190, 1989.

[Mathews, 1970] Mathews, M. V. and Mooer, F. R., "GROOVE-a program to compose, store, and edit functions of time," *Communications of ACM*, Vol.13, No.12, 1970.

[Mehrabian, 1972] Mehrabian, A., *Nonverbal Communication*, Aldine-Atherton, Chicago, 1972.

[Minsky, 1986] Minsky, M., *The Society of Mind*, New York, Simon and Schuster, 1986.

[Miwa et al., 2001] Miwa, H., Umetsu, T., Takanishi, A., Takanobu, H., "Human-Like Robot Head that has Olfactory Sensation and Facial Color Expression," *Proc. of IEEE Intl. Conf. on Robot and Automation*, pp. 459-464, 2001.

[Moller, 1997] Moller, C., *Virtual Cage*, 2nd Prospect exhibition by Canon Art Lab. http://www.canon.co.jp/cast/artlab/pros2/, 1997.

[Morita et al., 1991] Morita, H., Hashimoto, S., and Ohteru, S., "A Computer Music System that Follows a Human Conductor," *IEEE Computer Magazine*, 24(7), pp. 44-53, 1991.

[Nagashima, 1999] Nagashima, Y., "'It's SHO time' — An Interactive Environment for SHO(Sheng) Performance," *Proc. of International Computer Music Conference*, Beijing, 1999.

[Nakamura et al., 1994] Nakamura, J., Kaku, T., Hyun, K., Noma, T. and Yoshida, S., "Automatic Background Music Generation based on Actors' Mood and Motions," *The Journal of Visualization and Computer Animation*, Vol.5, pp.247-264, 1994.

[Nakata et al., 1998] Nakata, T., Sato, T., Mori, T., "Expression of Emotion and Intention by Robot Body Movement," in *Proc. of International Autonomous Systems 5* (IAS-5), 1998.

[Nakatsu, 1997] Nakatsu, R., "Image/Speech Processing that Adopts an Artistic Approach," *Proc. of Intl. Conf. on Acoustics Speech and Signal Processing*, ICASSP, 1997.

[Nilsson, 1980] Nilsson, N., *Principles of Artificial Intelligence*, Tioga Pub. Co., Palo Alto, CA, United States, 1980.

[Onoe et al., 1996] Onoe, N., Chang, D., and Hashimoto, S., "Background Music Generation Based on Scene Analysis," *Proc. of 1996 International Computer Music Conference*, pp. 361–362, 1996.

[Ogata et al., 2000] Ogata, T., Matsuyama, Y., and Sugano, S., "Acquisition of the Internal Representation in Robots -Toward the Human Robot Communication Using the Primitive Language," *The Intl. Journal of the Robotics Society of Japan (Advanced Robotics)*, pp. 277-291, 2000.

[Padgett et al., 1997] Padgett, C. and Cottrell, G., "Representing face images for classifying emotions," *Advances in Neural Information Processing Systems 9*, Cambridge, MA, MIT Press, 1997.

[Paradiso et al., 1999] Paradiso, J., Hsiao, K.Y. and Hu, E., "Interactive Music for Instrumented Dancing Shoes," *Proc. of International Computer Music Conference*, pp. 453–456, 1999.

[Paradiso, 1999] Paradiso, J., "The Brain Opera Technology: New instruments and gestural sensors for musical interaction and performance," *Journal of New Music Research*, 28:2, pp. 130-149, 1999.

[Pfeifer et al., 1999] Pfeifer, R. and Scheier, C., *Understanding Intelligence*, MIT Press, 1999.

[Riecken, 1992] Riecken, D., "Wolfgang - A System Using Emoting Potentials to Manage Musical Design," *Understanding Music with AI: Perspective on Musical Cognition*, eds., Balaban, M., Ebcioglu, K., and Laske, O., AAAI Press/MIT Press, 1992.

[Roads, 1996] Curtis Roads, *Computer Music Tutorial*, MIT Press, BA, USA, 1996

[Rokeby, 1995] Rokeby, D., Lecture for "Info Art.", *Kwanju Biennale*, Kwangju, Korea, 1995.
(available at http://www.interlog.com/drokeby/install.html)

[Rowe, 1992] Rowe, R., *Interactive Music Systems - Machine Listening and Composing*, MIT Press, United States, 1992.

[Rumelhart, 1986] Rumelhart, D.E., "Learning Internal Representation by Error Propagation," *Parallel Distributed Processing*, Vol. 1, pp. 318–362, MIT Press, 1986.

[Russell et al., 1985] Russell, J.A. and Bullock, M., "Multidimensional scaling of emotional facial expressions: Similarity from preschoolers to adults," *Journal of Personality and Social Psychology*, 48, pp. 1290–1298, 1985.

[Sammon, 1969] Sammon, J. W., “Nonlinear mapping algorithm for data structure analysis,“ *IEEE Trans. Computer*, vol. C-18, pp. 401–09, 1969.

[Sawada et al., 1997] Sawada, H., Onoe, N., and Hashimoto, S., "Sounds in Hands - A Sound Modifier Using Datagloves and Twiddle Interface -", *Proc. of International Computer Music Conference*, pp.309-312, 1997.

[Sawada et al., 1999] Sawada, H., and Hashimoto, S., "A Haptic Device Driven by Grasping Force For Hand Gesture Tele-Communication," *Proc. of the ASME Dynamic Systems and Control Division*, pp. 437–444, 1999.

[Schlosberg, 1952] Schlosberg, H. "The description of facial expressions in terms of two dimensions," *Journal of Experimental Psychology*, 44, pp. 229–237, 1952.

[Scholsberg, 1954] Schlosberg, H. "Three dimension of emotion," *The Psychological Review*, vol.61, no.2, pp. 81–88, 1954.

[Shepherd, 1998] Shepherd, G. M., *Neurobiology*, Oxford Univ. Press, 2nd ed., 1988.

[Scholkopf et al., 1999] Scholkopf, B., Burges, C. and Smola, A. (Eds), *Advances in Kernel Method: Support Vector Learning*, MIT Press, 1999.

[Schomaker et al., 1995] Schomaker et al., "A Taxonomy of Multimodal Interaction in the Human Information Processing System," *A Report of the ESPRIT PROJECT 8579*, 1995.

[Shawe-Taylor et al., 2000] Shawe-Taylor, J. and Cristianini, N., *An Introduction to Support Vector Machines*, Cambridge Univ. Press, 2000.

[Siegel et al., 1999] Siegel, W. and Jacobsen, J., "Composing for the Digital Dance Interface," *Proc. of 1999 International Computer Music Conference*, pp. 276–277, 1999.

[Silbernagel, 1979] Silbernagel, D., *Taschenatlas der Physiologie*, Thieme, 1979.

[Squire, 1998] Squire, D., "Learning a similarity-based distance measure for image database organization from human partitionings of an image set," *Proc. of Fourth IEEE Workshop on Applications of Computer Vision*, Princeton, NJ, United States, pp. 88–93, 1998.

[Sutton, 1992] Sutton, R. S. (ed), Special Issue of Machine Learning, *Machine Learning*, Vol. 8, pp. 1–395, 1992.

[Sutton et al., 1998] Sutton, R. S. and Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, 1998.

[Steels, 1993] Steels, L., "The artificial life roots of artificial intelligence," *Artificial Life*, 1(1)., pp. 75–110, 1993.

[Suzuki et al., 1997] Suzuki, K. and Hashimoto, S., "Modeling of emotional sound space using neural networks," *Proc. AIMI Intl. Workshop on Kansei - the technology of emotion*, pp. 116–121, 1997.

[Suzuki et al., 1998] Suzuki, K., Camurri, A., Ferrentino, P., and Hashimoto, S., "Intelligent Agent System for Human-Robot Interaction through Artificial Emotion," *Proc. of 1998 IEEE Intl. Conf. on System, Man and Cybernetics*, USA, pp. 1055–1060, 1998.

[Suzuki et al., 1999] Suzuki, K., Ohashi, T. and Hashimoto, S., "Interactive Multimodal Mobile Robot for Musical Performance," *Proc. of 1999 International Computer Music Conference*, Beijing, pp. 407–410, 1999.

[Suzuki, 2000] Suzuki, K., "A Quantification of Kansei using Neural Network and its application for Human Collaborative Robot," *Master Thesis*, Waseda University, 2000.

[Suzuki et al., 2000] Suzuki, K., Tabe, K. and Hashimoto, S., "A Mobile Robot Platform for Music and Dance Performance," *Proc. of 2000 International Computer Music Conference*, Berlin, pp. 539–542, 2000.

[Suzuki et al., 2001] Suzuki, K., Yamada, H. and Hashimoto, S., "Interrelating physical feature of facial expression and its impression," *in Proc. of IEEE/INNS International Conference on Neural Networks 2001*, Washington D.C., United States, pp. 1864–1869, 2001.

[Suzuki et al., 2002] Suzuki, K., Taki, Y., Konagaya, H., Hartono, P. and Hashimoto, S., "Machine Listening for Autonomous Musical Performance Systems," *Proc. of 2002 International Computer Music Conference*, Berlin, pp. 61-64, 2002.

[Tabe et al., 2001] Tabe, K., Suzuki, K. and Hashimoto, S., "Survival Strategy for Autonomous Mobile Robot," *Proc. of Second IEEE-RAS International Conference on Humanoid Robots*, Japan, pp. 485–492, 2001.

[Takahashi et al., 1999] Takahashi, S., Suzuki, K., Sawada, H., and Hashimoto, S., "Music Creation from Moving Image and Environmental Sound," *Proc. of 1999 International Computer Music Conference*, Beijing, pp. 240–243, 1999.

[Taki et al., 2000] Taki, Y., Suzuki, K. and Hashimoto, S., "Real-time Initiative Exchange Algorithm for Interactive Music System," *Proc. of 2000 International Computer Music Conference*, Berlin, pp. 266–269, 2000.

[Tax et al., 2001] Tax, D.M.J., and Duin, R.P.W., "Combining one-class classifiers," *Proc. of the Second International Workshop Multiple Classifier systems*, Vol. 2096, Springer Verlag, Berlin, pp. 299–308, 2001.

[Tevatia et al., 2000] Tevatia, G. and Schaal, S., "Inverse Kinematics for Humanoid Robots," *Proc. of IEEE International Conference on Robotics and Automation*, vol. 3, pp. 294–299, United States, 2000.

[Toda, 1962] Toda, M., "Design of a Fungus-Eater," *Behavioral Science*, 7, pp. 164-183. (Reprinted in Toda, *Man, robot and society*, pp. 100–129, 1982.), Martinus Nijhoff Publishing, 1962.

[Tojo et al., 2000] Tojo, T., Matsusaka, Y., Ishii, T., Kobayashi, T., "A Conversational Robot Utilizing Facial and Body Expressions," *Proc. of IEEE Intl. Conf. on System, Man and Cybernetics*, pp. 858-863, 2000.

[Van Hulle, 1997] Van Hulle, M. M., "Topology-preserving Map Formation Achieved with a Purely Local Unsupervised Competitive Learning Rule," *Neural Networks*, vol.10, No.3, pp. 431-446, 1997.

[Vapnik et al., 1997] Vapnik, V., Golowich, S., and Smola, A., "Support Vector Method for Function Approximation, Regression Estimation and Signal Processing," In Mozer, M., Jordan, M. and Petsche, T. (Eds) *Advances in Neural Information Processing System*, Vol. 9, pp. 281–287, MIT Press, 1997.

[Vapnik, 2000] Vapnik, V.N., *The nature of statistical learning theory*, Springer Verlag, New York, 2000.

[Wasserman et al., 2000] Wasserman, K.C., Blanchard, M., Bernardet, U., Manzolli, J., and Verschure, P., "Roboser: An Autonomous Interactive Composition System," *Proc. of the International Computer Music Conference*, pp. 531-534. Berlin, 2000.

[Watson et al., 1999] Watson, R. A., Ficici, S. G. and Pollack, J. B., "Embodied Evolution: Embodying an Evolutionary Algorithm in a Population of Robots," *Proc. of 1999 Congress on Evolutionary Computation*, IEEE Press, p.335–342, 1999.

[White, 1990] White, H., "Connectionist Nonparametric Regression: Multilayer Feedforward Network that Can Learn Arbitrary Mappings," *Neural Networks*, Vol. 3, No. 5, pp. 535–549, 1990.

[Yamada, 1993a] Yamada, H., "Visual information for categorizing facial expression of emotion," *Japan Psychology Rev.*, 35, pp.172–181, 1993.

[Yamada, 1993b] Yamada, H., "A basic research on the process of facial expression perception," *Ph.D Thesis*, Nihon University (*in Japanese*), 1993.

[Yokono et al., 1998] Yokono, J., Hashimoto, S., "Motion Interface for Omni-Directional Vehicle," *Proc. of 7th International Workshop on Robot and Human Communication*, pp. 436–441, 1998.

[Zatsiorsky, 1998] Zatsiorsky, V., "Kinematics of Human Motion," *Human Kinetics*, 1998.

[Zhang et al., 1997] Zhang, C. X., and Mlynski, D.A., "Mapping and Hierarchical Self-Organizing Neural Networks for VLSI Placement," *IEEE trans. on Neural Networks*, vol.8, No.2, 1997.

[Zhang et al., 1998] Zhang, Z., Lyons, M., Schuster, M. and Akamatsu, S., "Comparison between geometry-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron," *Proc. of the Third IEEE Conference on Face and Gesture Recognition*, Nara Japan, 1998.

List of Figures

List of Tables

List of Publications

Journal Papers

1. Suzuki, K., Hikiji, R. and Hashimoto, S., Development of an Autonomous Humanoid Robot, iSHA, for Harmonized Human-Machine Environment, *Journal of Robotics and Mechatronics*, Vol.14, No.5, 2002, pp. 324-332.

2. Hai, Q., Hartono, P., Suzuki, K. and Hashimoto, S, Sound database retrieved by sound, *Acoustical Science and Technology*, Vol. 23, No. 6, 2002, pp. 293-300.

3. Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A., Suzuki, K., Trocca, R,, and Volpe, G., EyesWeb - Toward Gesture and Affect Recognition in Dance/Music Interactive Systems, *Computer Music Journal*, Vol.24, No.1, MIT Press, 2000, pp. 57-69.

4. 鈴木 健嗣，アントニオ カムリ，橋本 周司，マルチモーダル移動ロボットを用いた仮想音楽環境, 日本バーチャルリアリティ学会論文誌, Vol. 5, No. 1, 2000, pp. 763-769.

5. 鈴木 健嗣，橋本 周司，ニューラルネットワークを用いた感性情報の数量化, 電子情報通信学会論文誌，Vol.J-82-D-II, 1999, pp. 677-684.

International Conference Papers

1. Suzuki, K., Hashimoto, S., ”A Multiclass Classification Method by Distance Mapping Learning Network,” *Proc. of 9th International Conference on Neural Information Processing*, Singapore, 2002, Vol.1, pp. 393-397.

2. Kobori, N., Suzuki, K., Hartono, P. and Hashimoto, S. (2002). ” Learning to Control a Joint Driven Double Inverted Pendulum using Nested Actor/Critic Algorithm,” *Proc. of 9th International Conference on Neural Information Processing*, Singapore, 2002, Vol. 5, pp.2610-2614.

3. Doi, M., Suzuki, K. and Hashimoto, S., ”Integrated Communicative Robot - BUGNOID,” *Proc. of 11th IEEE International Workshop on Robot and Human Interactive Communication*, 2002, Berlin, Germany, pp. 259-265.
4. Suzuki, K., Taki, Y., Konagaya, H., Hartono, P. and Hashimoto, S., ”Machine Listening for Autonomous Musical Performance Systems,” *Proc. of 2002 International Computer Music Conference*, ICMA, San Francisco, 2002, pp. 61-64.
5. Hartono, P., Suzuki, K., Hai, Q. and Hashimoto, S., ”Subjective Preference Oriented Global Sound Database,” *Proc. of International Computer Music Conference*, ICMA, San Francisco, 2002, pp. 446-449.
6. Suzuki, K. and Hashimoto, S., ”Harmonized Human Machine Environment for Humanoid Robot,” *Proc. of IEEE-RAS International Conference on Humanoid Robots (Humanoids2001)*, Tokyo, Japan, 2001, pp. 43-50.
7. Tabe, K., Suzuki, K., Hartono, P. and Hashimoto, S., ”Survival Strategy Learning for Autonomous Mobile Robot,” *Proc. of 2nd IEEE-RAS International Conference on Humanoid Robots (Humanoids2001)*, Tokyo, Japan, 2001, pp. 485-492.
8. Suzuki, K. and Hashimoto, S., ”Robotic Interface for Embodied Interaction via Musical and Dance Performance,” *Proc. of International Workshop Human Supervision and Control in Engineering and Music*, Kassel, Germany, 2001, pp. 223-228.
9. Suzuki, K. and Hashimoto, S., ”Qualitative Musical Information Retrieval,” *Proc. of International Workshop Human Supervision and Control in Engineering and Music*, Kassel, Germany, 2001.
10. Suzuki, K., Yamada, H., Hartono, P. and Hashimoto, S., ”Modeling of Interrelationship between Physical Feature of Face and its Impression,” *Proc. of The First International Symposium on Measurement, Analysis and Modeling of Human Functions*, Sapporo, Japan, 2001, pp. 310-315.
11. Suzuki, K., Yamada, H. and Hashimoto, S., ”Interrelating Physical Feature of Facial Expression and Its Impression,” *Proc. of IEEE/INNS International Conference on Neural Networks*, Washington D.C., 2001, pp. 1864-1869.
12. Suzuki, K., Tabe, K. and Hashimoto, S., ”A Mobile Robot Platform for Music and Dance Performance,” *Proc. of International Computer Music Conference*, ICMA, San Francisco, 2000, pp. 539-542.

13. Taki, Y., Suzuki, K. and Hashimoto, S., "Real-time Initiative Exchange Algorithm for Interactive Music System," *Proc. of International Computer Music Conference*, ICMA, San Francisco, 2000, pp. 266-269.
14. Suzuki, K., Ohashi, T. and Hashimoto, S., "Interactive Multimodal Mobile Robot for Musical Performance," *Proc. of International Computer Music Conference*, ICMA, San Francisco, pp. 407-410, 1999, pp. 407-410.
15. Takahashi, S., Suzuki, K., Sawada, H. and Hashimoto, S., "Music Creation from Moving Image and Environmental Sound," *Proc. of International Computer Music Conference*, ICMA, San Francisco, 1999, pp. 240-243.
16. Camurri, A., Hashimoto, S., Suzuki, K. and Trocca, R., "Kansei Analysis of Movement in Dance/Music Interactive Systems," *Proc. of International Conference of HUmanoid and RObot (HURO99)*, Tokyo, Japan, 1999, pp. 9-14.
17. Camurri, A., Hashimoto, S., Ricchetti, M., Suzuki, K., Trocca, R., Volpe, G., "Kansei Analysis of Dance Performance," *Proc. of IEEE International Conference on System, Man and Cybernetics*, Tokyo, Japan, 1999, pp. IV- 327-332.
18. Suzuki, K., Camurri, A., Ferrentino, P. and Hashimoto, S., "Intelligent Agent System for Human-Robot Interaction through Artificial Emotion," *Proc. of IEEE Intl. Conf. on System, Man and Cybernetics*, San Diego, CA, 1998, pp. 1055-1060.
19. Suzuki, K., Hashimoto, S., "Emotional Sound Space Using Neural networks," *Proc. of AIMI Intl. Workshop on Kansei - Technology of emotion*, Genoa, Italy, 1997, pp. 116-121.

Oral Presentations

1. 鈴木 健嗣，橋本 周司，"ヒューマノイドインタフェースの構築とその評価," 精密機械工学会 第3回システムインテグレーション部門講演会 (SI2002)，2002.
2. 土居 正崇，鈴木 健嗣，橋本 周司，"自律移動型ロボットによる動的環境の認識," 精密機械工学会 第3回システムインテグレーション部門講演会 (SI2002)，2002.
3. 田部 圭史朗，鈴木 健嗣，橋本 周司，"環境で自活適応するロボット," 第 62 回情報処理学会全国大会論文集 (2), pp.143-144，東京工業大学，2001.

4. 田部 圭史朗，鈴木 健嗣，橋本 周司，”自己保存に基づいたロボットの行動獲得,” 第 6 回知能メカトロニクスワークショップ，pp. 43-48，香川大学，2001.
5. 高橋 正吾，鈴木 健嗣，橋本 周司，”動画像と環境音による音響生成を用いたインタラクティブシステム,” 情報処理学会第 60 回全国大会講演論文集 (2), pp.243-244，慶應義塾大学，2000.
6. 鈴木 健嗣，大橋 武史，田部 圭史朗，橋本 周司，””iDance” - マルチモーダル移動ロボットを用いた音楽生成環境,” 日本情報処理学会インタラクション 2000, 東京工業大学，2000.
7. 鈴木 健嗣，橋本 周司，”マルチモーダル情報を用いた仮想音楽環境,” 情報処理学会研究会報告音楽情報研究会 (SIGMUS)，香川大学，2000.
8. 鈴木 健嗣，山田 寛，橋本 周司，”ニューラルネットワークによる顔面表情の物理的パラメータと感性的パラメータの対応付け,” 電子情報通信学会ヒューマン情報処理基礎研究会 HCS2000-47, pp. 7-13，九州大学, 2000.
9. Suzuki, K., Camurri, A. and Hashimoto, S., ”A Multimodal/Multimedia Environment for Human-robot Interaction,” 1999 年電子情報通信学会総合大会，慶應義塾大学, 1999.
10. Suzuki, K., Camurri, A. and Hashimoto, S., ”An Interactive Agent for Human-robot Interaction through Artificial Emotion,” 日本情報処理学会インタラクション’99，東京工業大学, 1999.
11. Suzuki, K., Camurri, A. and Hashimoto, S., ”Toward Human-robot Interaction through Artificial Emotion”, 1998 年電子情報通信学会総合大会, 東海大学, 1998.
12. 鈴木 健嗣，橋本 周司，”ニューラルネットワークを用いた物理的特徴と感性情報の対応づけ,” 信学技法，電子情報通信学会ヒューマンインタフェース研究会 HIP97-32, 広島大学，1998.
13. 鈴木 健嗣，橋本 周司，”ニューラルネットワークによる数量化理論への応用,” 第 54 回情報処理学会全国大会，千葉工科大学，1997.
14. 鈴木 健嗣，橋本 周司，”ホログラトロンとアソシアトロンの結合による連想記憶,” 1997 年電子情報通信学会総合大会, 関西大学，1997.

Others (selected)

Invited Talks

1. CEATEC2001 ジュニア・サイエンス・スクール，幕張メッセ国際会議場, 千葉県, 10 月 6 日，2001.
2. Public seminar, "Kansei" - Technology of Emotion -, Teatro di Carlo Felice, Genoa, Italy, 1998.

Invited Demonstrations

1. ロボット特別展，千葉県立現代産業科学館，千葉県，8 月 8 日〜8 月 11 日，2002.
2. ROBODEX2002，パシフィコ横浜，神奈川県，3 月 27 日〜3 月 31 日，2002.
3. 北陸技術交流テクノフェア 2001，産業市民会館，福井県，10 月 25 日〜10 月 26 日，2001.
4. ロボフェスタ 2001 神奈川，川崎市民アリーナ，神奈川県，9 月 8 日〜9 月 16 日，2001.
5. ザ・ロボット博 21，ナガシマスパーランド，三重県，4 月 27 日〜5 月 6 日，2001.
6. Arti Visive 2, Palazzo Ducale, Genova, Italy, 1998.
7. Music Planet - Il giardino della musica, Palazzina Liberty, Milano, Italy, 1998.

Magazine Review

1. "Topics 北京発音分析合成／インタラクティブシステム (ICMC'99 より)," 鈴木健嗣 他，コンピュータサイエンス誌 bit5 月号，vol.32, No.5, 2000, 共立出版.

Newspapers

1. 人工知能ー感情　「顔の特徴点から感情を分析」，産経新聞, 2001.

Other Media

1. テレビ放映,『NHK インターネット情報局』NHK 教育、2002 年 2 月 12 日, 19 日放送、14:00〜14:20.
2. テレビ放映,『21 世紀！ ドラえもんのひみつ道具大発明スペシャル！』テレビ朝日系、2000 年 12 月 17 日放送、14:00〜15:25.