論文 / 著書情報
Article / Book Information

題目(和文)	さりげないヒューマン・ロボットコミュニケーションにおける深層的状況理解
Title(English)	Deep Level Situation Understanding in Casual Communication between Humans and Robots
著者(和文)	湯永康
Author(English)	Yongkang TANG
出典(和文)	学位:博士(学術), 学位授与機関:東京工業大学, 報告番号:甲第9579号, 授与年月日:2014年3月26日, 学位の種別:課程博士, 審査員:廣田　薫,寺野　隆雄,室伏　俊明,長谷川　修,小野　功
Citation(English)	Degree:Doctor (Academic), Conferring organization: Tokyo Institute of Technology, Report number:甲第9579号, Conferred date:2014/3/26, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Deep Level Situation Understanding in Casual Communication between Humans and Robots

Yongkang TANG
湯　永康

Supervisor: Prof. Kaoru HIROTA

Doctoral Thesis

Tokyo Institute of Technology
東京工業大学
Interdisciplinary Graduate School of Science and Engineering
大学院総合理工学研究科
Department of Computational Intelligence and Systems Science
知能システム科学専攻

March 2014

Abstract

The concept and inference system of deep level situation understanding are proposed to realize human-like natural communication among agents (e.g., humans and robots/machines). The human robot communications based on visible and audible information are called surface level communications, such as gesture/posture understanding, facial expression understanding, and speech/voice understanding. The deep level situation understanding is characterized as unifying the surface level understanding, emotion understanding, intention understanding, and atmosphere understanding by applying universal and customized knowledge of each agent.

Gesture communication is an important communication way in surface level understanding. A multimodal gesture recognition method for Mascot Robot System is also proposed based on Choquet integral by fusing camera and 3D accelerometer data. By calculating two fuzzy measures in the training phase for camera based and accelerometers based units, the proposed system obtains enough recognition rate of 96.0% for 8 types of gestures by improving the recognition rate approximate 20%

compared with that of each unit.

The proposed deep level situation understanding aims to smooth the communications between human and robot, to realize harmonious communication by excluding unnecessary troubles or misunderstandings among agents, and finally to create a peaceful, happy, and prosperous humans-robots society.

A simulated experiment is established to implement the proposed deep level situation understanding system where meeting-room reservation in a company is done between a human employee and a secretary-robot. Twelve subjects are asked by questionnaires to evaluate the response of the proposed inference system comparing to the responses from familiar people. The proposed deep level inference system achieves a naturalness value of 0.84 which is between the ranks of “natural (=1.0)” and “a little natural (=0.75)” comparing to communicate with familiar people.

The proposed deep level situation understanding may be applied in robot systems for casual communications such as restaurant service robot systems, secretary robot systems, domestic robot system, and therapy robot systems.

Acknowledgements

At first, I would like to express my sincere gratitude to my supervisor, Professor Kaoru HIROTA, for enrolling me in his lab and willing to supervise my master and doctoral studies. His guidance, encouragement, and inspiration were invaluable in the last five years. I am extremely grateful to my supervisor for guiding me to an interesting research topics. I also would like to express my sincere gratitude to Professor Takao TERANO, Professor Toshiaki MUROFUSHI, Professor Osamu HASEGAWA, and Professor Isao ONO for their precious time, valuable comments that have led me to improve and perfect my master and doctoral thesis.

Meanwhile, I would like to thank Dr. Fang-Yan Dong, assistant professor at Hirota Lab for the grateful advice to me for both research and daily-life. Additionally, my sincere gratitude goes to Ms. Harumi Hoshino for all the irreplaceable help and encouragement.

I am also grateful to my friends at the Hirota lab. I am thankful to my tutor Dr. Hai An VU, Dr. Phuc. Q. LE, Dr. Chastine Fatichah, and Dr. A. M. ILIYASU for valuable

advices about research.

I would also like to thank Atsushi SHIBATA, Kazuhiro OHNISHI, Fei YAN, Jiajun LU, Tianyu LI, Jesus A. GARCIA SANCHEZ, Masakazu FUNAZUKURI, Mina YUHKI, Takahiro KAWABUCHI, Mariko SHOZAWA, and Li BAI for their helping on taking the demonstrate video. I would also like to thank all Hirota Lab members for their continuing support and never ending patience with my constant requests for assistance.

I would also like to thanks the Monbukagakusho Honors Scholarship from Monbukagakusho and Rotary Yoneyama Scholarship from the Rotary Yoneyama Memorial Foundation for financial support. Thanks the encouragement and helps from my counselors (Ms. Akio ISHIYAMA and Ms. Hideo ISHII) and the other members in the Yokohama-Seya Rotary Clubs.

At last, I would like to thanks my mother (Caiye YANG), father (Zhongdong TANG), wife (Yajun NAN), and my daughter (Nanyue TANG) for their continual selfless encouragements and supports. I am very appreciative of my mother' encouragement on my study. I'm greatly thanks my wife for encouraging me to conquer the doctor research. I also appreciate the innocent smiles from my daughter which always heals me and encourages me to overcome difficulties. I could not finish my doctoral research without the encouragements and supports from my family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Robots may exist everywhere in future. They may work in factories as industrial robots, or do housework in the families as household robots, or serve in restaurants as waitress robots, or work with human as colleague. It's a big challenge for a robot to understand the action (e.g., speech, gestures, facial expression) of human. In the human-robot co-existing society, the abilities of human like communications are greatly needed to robot. There are so many challenges during human robots interaction. For example, misunderstanding caused by poor recognition rate of human actions, unnatural communication between humans and robots.

Hospitality, in general, is not only from consideration of utterance, but also from attention to nonverbal information. Recognition of human motions is greatly needed for communication between humans and robots in social robot systems such as service robots,

entertainment robots, and household robots, which need to assist human with hospitality attitude in daily life. The casual communication robotic system, e.g., Mascot Robot System [1][2], has been proposed based on verbal information to assist human in living environment. As for nonverbal approaches, gesture recognition has recently become attractive research themes in the field of Human-Computer Interaction (HCI) for robotics [3], human behavior studies [4], emotion [5] and sign language [6] recognition, and virtual environment navigation [7]. Most works on the gesture recognition for HCI have been done based on visual information, in terms of Hidden Markov Models(HMM) [8], Dynamic Time Warping(DTW) [9], and Self Organizing Markov Map(SOMM) [10]. Camera based human gesture recognition system usually capture motion information of body parts (hands and head) by skin color. It is, however, easy to have a noise effect by the object that has similar color of the skin [11]. On the other hand, application of accelerometer based gesture recognition is an emerging technique to improve recognition performance. A wearable acceleration sensor based air writing system that recognizes the gestures of writing alphabets is proposed in [12].

A multimodal gesture recognition system is proposed in [13], where information of both 3D acceleration and camera sensors are combined by fuzzy logic. In the study, when the similarity calculated from the acceleration recognition units is greater than a given threshold value, the results of the accelerometer unit are taken as the final results of the gesture recognition system. In reverse, image recognition unit processes the candidate gestures that come from acceleration unit to get final result. They, however, have not discussed on how to decide the given threshold. So, it is difficult to apply this method for other general cases of gesture recognition.

To deal with this problems, a Choquet integral [14] based multimodal gesture recognition method is proposed, where human gestures are recognized based on the fusion of video images and 3D acceleration sensors. The Angular Metrics for Shape Similarity (AMSS) [15] algorithm is employed to evaluate the similarity between gesture templates and input data from above two units separately. Next, the optimal fuzzy measures of Choquet integral are calculated from the training data by a hill climbing algorithm. In the fusion processing, the similarities of both units are fused by Choquet integral. Finally, the gesture with maximum fused similarity is chosen as the recognition result of the multimodal gesture recognition system.

To validate the proposal, a multimodal gesture recognition system is implemented with a Logicool Qcam(R) camera and two wearable 3D acceleration sensors (manufactured by Microstone Inc.)[16]. The recognizing gestures in the system are eight types of typical human emotional-gestures of the Mascot Robot System, i.e., "toast", "throw dart", "victory", "banzai", "squatting with hands over the head", "face covering", "guiding" ,and "bye-bye".

Robots are increasingly capable of co-existing with humans in environment, such as in factories, offices, restaurants, hospitals, elder care facilities, and homes. The ability of comprehending human activities, e.g., gesture/posture, speech, and emotion, is required for robots in casual communication, i.e., human-human like communication. Verbal and non-verbal communications are the two basic ways in casual communications, which transmit among various agents such as humans and robots/machines. Several Spoken Dialog Systems are proposed for verbal communication [17] [18]. As for nonverbal approaches, gesture recognition has become an attractive research theme in the field of Robot Control [19], Virtual Game Control [20]. Most works on gesture

recognition for Human Robot Interaction (HRI) have been done based on visual information [19] [21]. To improve the robustness of gesture recognition system, a Choquet integral based multimodal gesture recognition system [22] is proposed. Besides the verbal and non-verbal communication, the atmosphere of the communication environment are also important factor in Human Robot Communication. The Fuzzy Atmosfield [23] is proposed for representing the visualizing the atmosphere.

In casual communication among humans, human may hide their real emotions, intentions, and opinions. But other humans may be able to understand them to some extent by understanding the spoken contents, voice tones, and facial expression changes. The communications only based on audible information (e.g., speech and voice) and visible information (e.g., gesture, posture, and facial expression) are called surface level understanding in this thesis, while deep level situation understanding is characterized by unifying the surface level understanding, emotion understanding, intention understanding, and atmosphere understanding by applying careful attention to both universal and agent dependent customized knowledge. The deep level situation understanding framework consists of a gesture/posture recognition module, speech/voice recognition module, emotion recognition module, intention understanding module, atmosphere understanding module, and knowledge (including universal knowledge and customized agent-dependent knowledge) of the interlocutor. Appropriate responses (e.g., speech, gesture, and facial expression) will produced as the output of the deep level situation understanding framework.

The deep level situation understanding in casual communication among various agents, e.g., humans and robots/machines, aims at three issues. Firstly, humans must pay special attention to robots in the ordinary human-machine communication systems, but

such burden may be reduced if robots have deep level situation understanding abilities. Secondly, in the real world, unnecessary troubles or misunderstandings in human-human communications sometimes may happen but the deep level situation understanding can make it possible to avoid such lower level troubles. The customized agent-dependent knowledge will help to comprehend and avoid miscommunication. Thirdly, with the consideration of surface level understanding, emotions, intentions, atmospheres, universal knowledge, and customized agent-dependent knowledge, it will also help to understand the background, habits, and intention of the agent for smoothing natural Human Robot Interaction, so as to create a peaceful, happy, and prosperous society which is consisted of humans and various specification robots/machines. To illustrate such a peaceful, happy, and prosperous humans-robots society, a short story is demonstrated by four persons, two eye robots, and a therapy-robot PARO.

In this research, a concept and an inference system of Deep Level Situation Understanding is proposed to realize human-human like communication among humans and robot.

This dissertation is organized as follows:

In chapter 2, a multi-modal gesture recognition method based on Choquet integral, is proposed for improving the accuracy of recognizing human actions in Human Robot Interaction. A camera sensor and two accelerometer sensors are employed for capturing the motion of gestures in the proposed multimodal system. Six subjects are invited to perform the eight typical gesture of the Mascot Robot System. By calculating the optimal fuzzy measures for the camera recognition and accelerometer recognition units in the training process, the proposed Choquet integral based fusion method greatly

improved the accuracy of the multimodal system comparing to the accuracy of camera based recognition unit and accelerometer based recognition unit.

In chapter 3, the concept and inference system of deep level situation understanding are introduced for achieving human-human like natural communications among humans and robots. The proposed deep level situation understanding aims to smooth the communications between human and robot, so as to realize harmonious communications and finally to create a peaceful, happy, and prosperous humans-robots society. A simulated experiment is established to implement the proposed deep level situation understanding system where meeting-room reservation in a company is done between a human employee and a secretary-robot. Twelve subjects are asked by questionnaires to evaluate the response of the proposed inference system comparing to the responses from familiar people.

Finally, this research is summarized in chapter 4.

The roadmap (Figure 1.1) visualizes the relation of each chapter and summarizes the organization of this thesis.

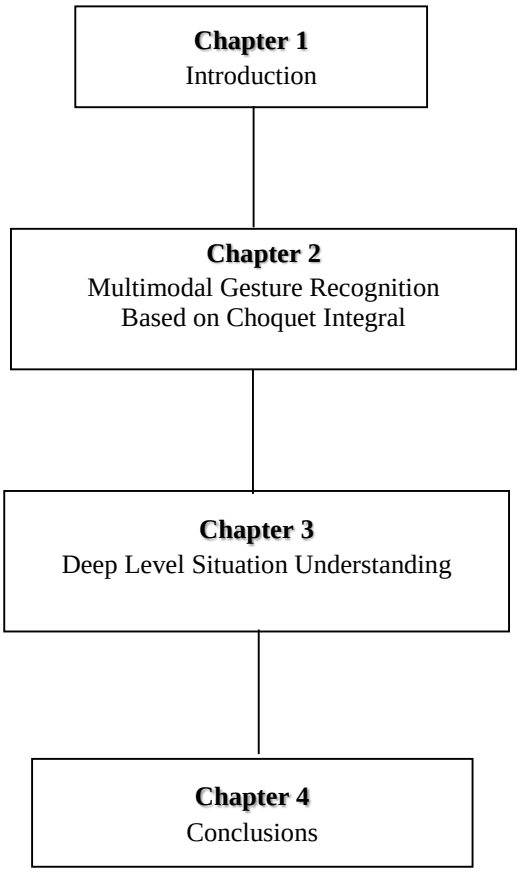


Figure 1.1. Research Roadmap

Chapter 2

Multimodal Gesture Recognition Based on Choquet Integral

Gesture communication is an important communication channel in daily life. Understanding human gestures is also a big challenge for machines. As the eye of machine, camera based human gesture recognition system usually tracking motion of body parts (hands and head) by skin color cue. It is, however, easy to be effected by the object that has similar color of the skin [11]. On the other hand, the accelerometer sensor based is an emerging technique for acquiring motion information of gesture. A wearable acceleration sensor based air writing system that recognizes the gestures of writing alphabets is proposed in [12].

A multimodal gesture recognition system is proposed in [13], where information of both 3D acceleration and camera sensors are combined based on fuzzy logic. In the study, when the similarity calculated from the acceleration recognition units is greater than a given threshold value, the recognition result from the accelerometer unit is taken as the final result. In reverse, image recognition unit processes the candidate gestures that come from acceleration unit to get final result. They, however, have not discussed about how to decide the given threshold. So, it is difficult to apply this method for general cases of gesture recognition.

To realize robust gesture recognition, a Choquet integral [14] based multimodal gesture recognition method is proposed, where human gestures are recognized by fusing the results of vision and 3D acceleration based recognition units. The Angular Metrics for Shape Similarity (AMSS) [15] algorithm is employed to evaluate the similarity between gesture templates and input data from the two units separately. Next, the optimal fuzzy measures of Choquet integral are calculated from the training data by a hill climbing algorithm. In the fusion processing, the similarities of the two units are fused by Choquet integral. Finally, the gesture with maximum fused similarity is chosen as the recognition result of the multimodal gesture recognition system.

To demonstrate the validity, the proposed gesture recognition system is implemented with a Logicool Qcam(R) camera and wearable 3D acceleration sensors (manufactured by Microstone Inc.). The recognizing gestures in the system are eight types of typical human emotional-gestures of the Mascot Robot System, i.e., "toast", "throw dart", "victory", "banzai", "squatting with hands over the head", "face covering", "guiding" ,and "bye-bye"

2.1 Mascot Robot System

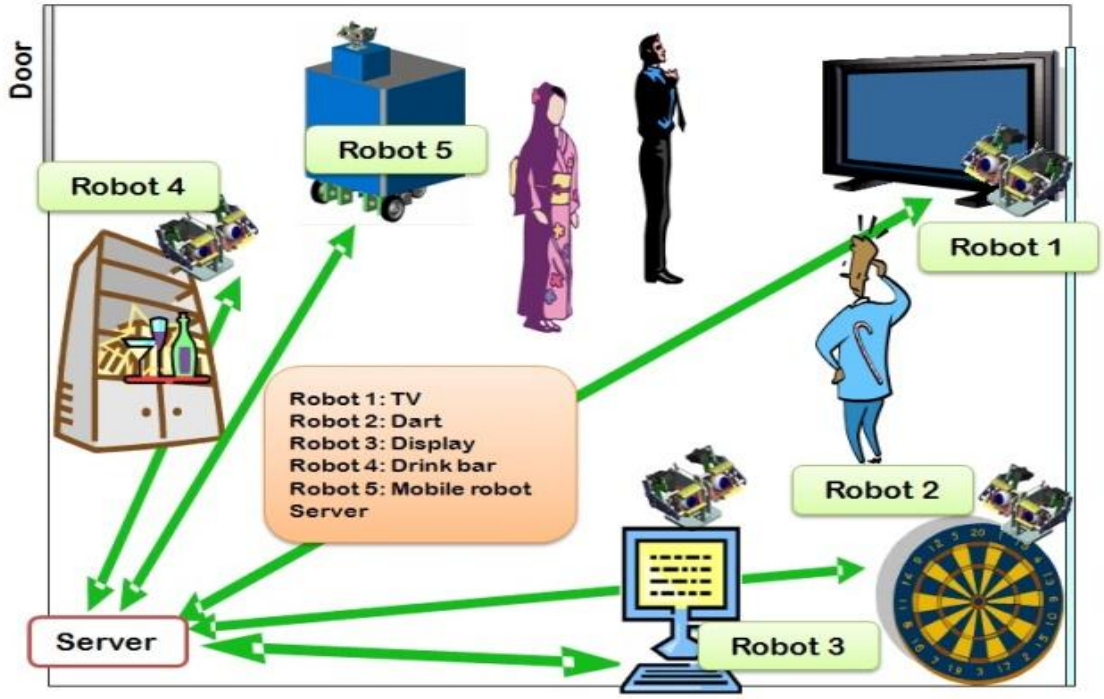


Figure 2.1. Mascot Robot System in Home Party Environment

The focus of Human Robot Interactions has been concentrated on social robot systems such as service robots, entertainment robots, and domestic robots in home environments, where robots are demanded to communicate with humans naturally and smoothly. The Japanese government has promoted such research directions, and the Mascot Robot System has been developed supported by New Energy and Industrial Technology Development Organization (NEDO) and Japan Society for the Promotion of Science (JSPS). The Mascot Robot System is proposed in household environments, which provides casual communication between humans and robots through human speeches and gestures.

As an example, the architecture of Mascot Robot System used in living environment is shown in Figure 2.1. There are four fixed type eye robots placed on furniture or appliances such as TV, dart game board, information monitor display, drink mini bar, and an autonomous mobile eye robot in the living room. Each robot is controlled by a networked laptop. The four fixed robots are connected to the common server via LAN cable, and the mobile robot are wirelessly connected to the server. The whole system is controlled and supervised by the common server, which also coordinate the action and responses of the robots. In the Mascot Robot System, a home party scenario, e.g., greeting visitors, playing darts game, and retrieving drink information, is implemented to show the responsiveness of the system by considering the emotions of eye-robots and the atmosphere in the living room. There are four humans (a host and three guests including one unexpected guest) and five robots in the scenario.

Every robot is composed of a laptop computer, a microphone, a web camera, and an eye mechanism. Firstly, the speech information, images information and acceleration information are captured from microphone, web camera, and 3D accelerometers. Then these information are processed and the results are shared with the common server. The eye robot will express their emotions for responding to the human gestures which is recognized by the proposed multimodal gesture recognition module. Eight types of typical emotional-gestures in a home party scenario are used in the multimodal gesture recognition module. There also exist some other types of gestures that control the mobile robot moving.

In order to integrate all information in the system, Robot Technology Middleware (RTM) is used to construct the network system, called RTM-Network. In the RTM-Network, each robot can be viewed as a network component, and each function unit of

robots is called a Robot Technology Component (RTC). There are 8 kinds of modules constructed in the RTM network of Mascot Robot System:

1) A speech recognition module (SRM): This module recognizes human speech and controls the output of suitable word spoken by the robot.

2) A gesture recognition module (GRM): This module recognizes human gestures by fusing the recognition result of camera and 3D accelerometers recognition units for robust and accurate recognition.

3) An eye-robot control module: This module controls eye robot movement according to recognized emotion from emotion processing module.

4) An emotion processing module: This module has two tasks. The first one is to transfer the recognition result from speech recognition module to the server module for the further processing. The second one is to receive the data from server module and to send them to eye-robot control module.

5) A display module: This module shows the recommended information on display.

6) A mobile robot control module: This module controls the movement of the mobile robot.

7) An information retrieval module: This module replies the information in correspond to the request given by the scenario server module.

8) A scenario server module: This module supervises whole system and communicates with information retrieval engine module.

The relation of the 8 kinds of modules is illustrated in Figure 2.2.

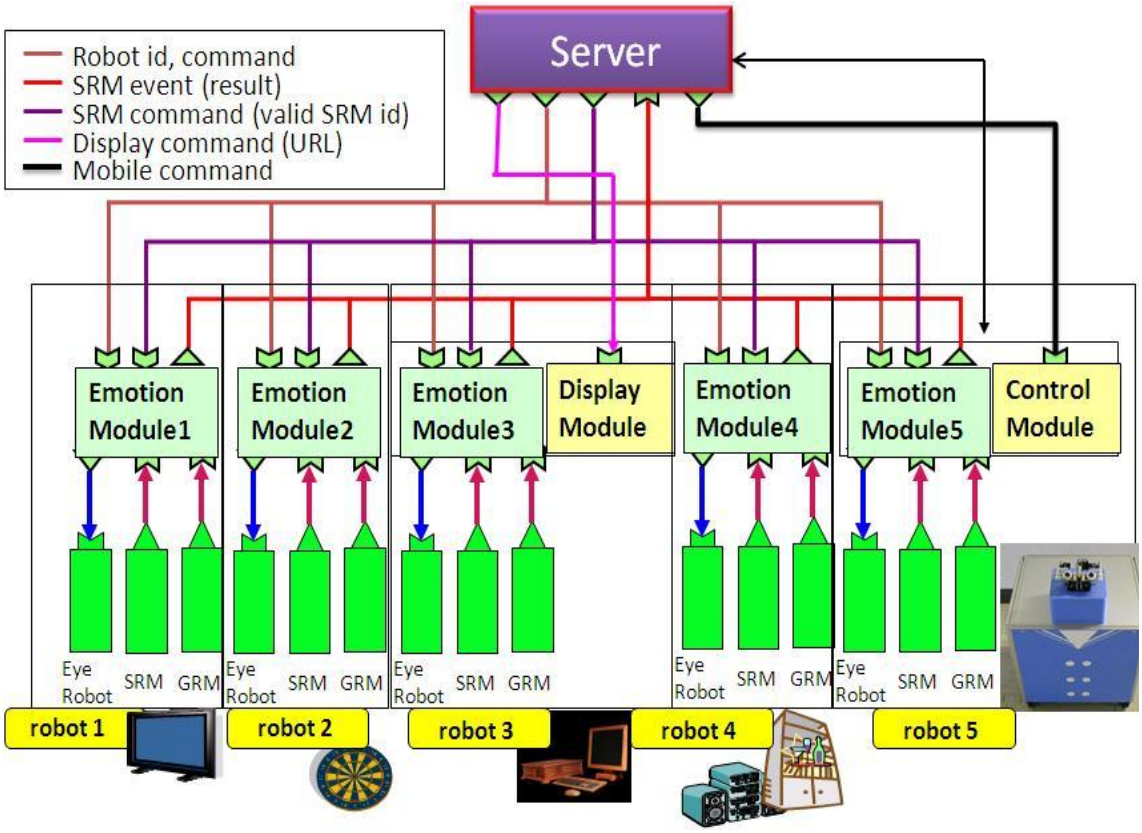


Figure 2.2. RTM Network for Mascot Robot System

In each robot, the recognized result of GRM is the one of the inputs to the Emotion Module. When an input comes, the Emotion Module will pre-process it, pass the processed data to server, and then the server will send its responding instruction to other modules.

The eye robot has a pair of eyeballs and eyelids. The motions of the eye robot reflect the fuzzy mentality expression, where eye robots express the emotional expressions such as happiness, sadness, surprise, and anger. The appearance of an eye robot and mobile robot is shown in Figure 2.3 and 2.4.

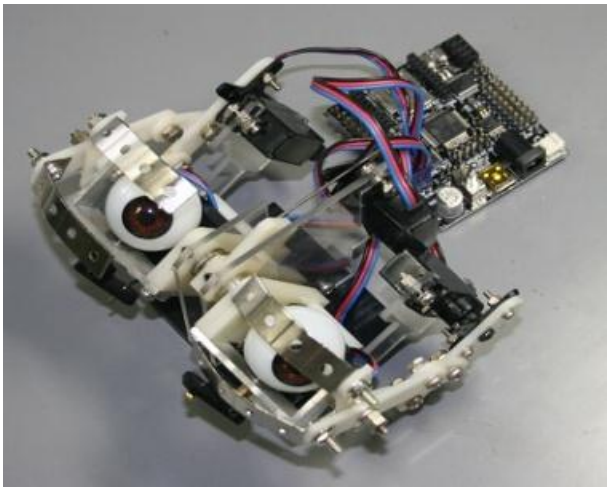


Figure 2.3. The Eye Robot in Mascot Robot System

Figure 2.4. The Mobile Robot

2.2. Multimodal Gesture Recognition System

The architecture of the proposed gesture recognition system is shown in Figure 2.5. The system consists of two units, i.e., camera based gesture recognition unit and accelerometers based gesture recognition unit.

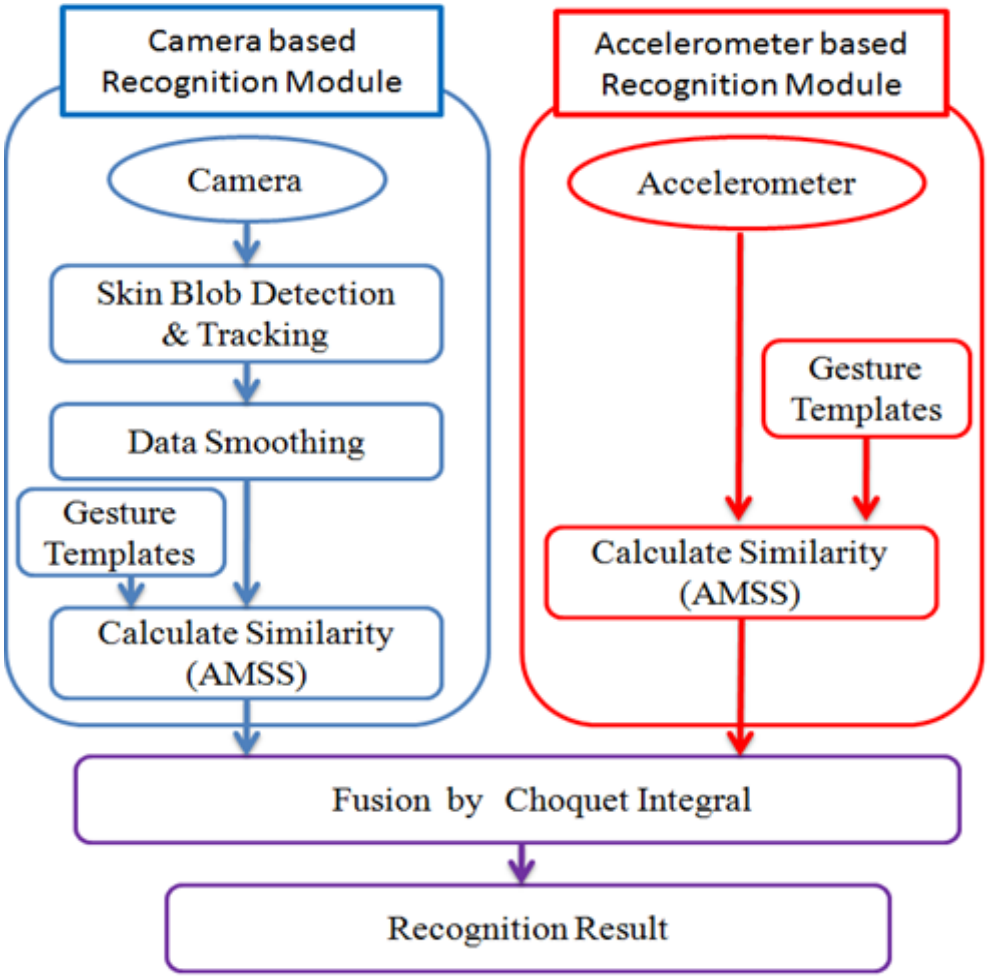


Figure 2.5. The Architecture of Multimodal Gesture Recognition System

2.2.1 Camera based Gesture Recognition Unit

The video images captured from the camera are converted from RGB color space to Hue-Saturation-Value (HSV) color space to obtain the human skin information smoothly. After that, the skin blob detection and tracking sub-unit is used to detect and to track the skin color blobs (both hands and head). The trajectory information of these three blobs is smoothed in the data smoothing sub-unit. Finally, AMSS algorithm is applied to calculate the similarity between the tracked trajectory and the gesture template.

2.2.1.1 Skin Blob Detection and Tracking

The skin blob detection and tracking sub-unit uses the HSV color space because the HSV color space is suitable for detecting human's skin color. A pixel's color is identified as the skin color if the HSV value of this pixel exists in certain color ranges set in advance. The possible range of skin color, however, is changed according to the environment, and the final assigned range in the experiment is $2 \leq \mathrm{H} \leq 14$, $65 \leq \mathrm{S} \leq 180$, and $20 \leq \mathrm{V} \leq 250$ by using OpenCV Library. This skin color range is applied to all participants in experiments. It is confirmed that the three regions of the head and two hands parts are detected in the skin blob detection processing sub-unit.

The detected blobs are labeled from the first frame, i.e., the blobs in the left side and right side of the frame are labeled as hands, and the blob between two hands is labeled as head. From the second frame on, the nearest blob to the blob in the previous frame will be marked as the same label. The sequence of coordinates of the three body parts are obtained as the trajectories of left hand, head, and right hand.

2.2.1.2 Feature Extraction and Data Smoothing

To recognize gestures, the 2D locations of blobs' centers are used as the features. The locations of left hand, head, and right hand are denoted as $\{x_l, y_l\}$, $\{x_h, y_h\}$, and $\{x_r, y_r\}$, respectively.

In the feature extractions part, the following two processes are done on the raw data:

(1) **Data Smoothing**: Simple moving average of 3 adjacent points is used to smooth raw data, i.e.,

$$x_{smoothing} = (x_{i-1} + x_i + x_{i+1}) /3, \quad (2.1)$$

where x_i stands for the raw locations of blobs.

(2) **Quantization**: The Dynamic Time Warping (DTW) algorithm mentioned in 2.2.1.4 is used for recognizing gestures. The complexity of the DTW algorithm is dependent on the length of the input sequence and template sequence. The short length of sequences contributes to the performance improvement of recognition. If the distance between a point and its nearest point of the trajectory is shorter than 10 pixels, the point is ignored; otherwise the point is added to the points' sequence of trajectory.

2.2.1.3 Template Selection

The DTW algorithm is a template-dependent method. The recognition rate of DTW based system is greatly dependent on the quality of the reference template. Accordingly selecting reliable template is one of the important tasks in the training process. There are several methods for template selection such as random selection,

minimum selection, normal selection, and multiple selection [24]. After checking from the cost performance viewpoint, minimum selection is used for selecting the appropriate template for the gesture recognition.

2.2.1.4 Camera based Gesture Recognition by AMSS

The Dynamic Time Warping (DTW) is widely used for matching time series by calculating the optimal distance between two time series based on dynamic programming. To calculate the DTW value of two time series, a distance matrix, which represents the distance of each element pair in the two time series, is built. Dynamic programming is then used to find the optimal path, called warping path. The accumulated value of the warping path represents the similarity (distance) of two time series.

The Angular Metrics for Shape Similarity (AMSS) [15] algorithm is used to calculate the similarity between gesture sequence and template sequence, which is a robust improvement of DTW algorithm. Instead of calculating the distance of elements in conventional DTW, the AMSS takes the cosine value of the intersection angular of two trend vectors, which is computed by neighboring element of each time series, as the distance measurement. The warping path is computed through the whole distance matrix. The similarity of two time series is calculated as the average value of accumulated value on optimal warping path.

The AMSS algorithm is based on trend vectors which are scale-invariant, so it has the power for recognizing user-independent gestures. Besides, due to its efficiency and simplicity, the AMSS algorithm is selected as the main recognition algorithm.

2.2.2 Accelerometer based Gesture Recognition Unit

Two 3D acceleration sensors are used for capturing the motion of human's two hands.

The sequence of feature values are taken as the input of the accelerometers based gesture recognition unit. The AMSS algorithm is used for measuring the similarity of input sequence and template sequences. The average similarity of the 3-axes is taken as the result of this unit.

2.2.3 Proposed Fusion Method based on Choquet Integral

The information obtained from two accelerometers and a camera sensors are different in accuracies and frequencies. These information are processed by different preprocessing and recognition algorithms. The recognition result comes from accelerometers and camera is not additive. Thus, a Choquet integral based method is proposed for fusing the similarity of accelerometers and camera based recognition units.

The Choquet integral in terms of fuzzy measure is studied by Murofushi and Sugeno[14]. Let $S^{a}_{i,g,t}$ denotes the similarity between i-th training data of g-th gesture and the t-th template sequence by the accelerometers based unit, and $S^{c}_{i,g,t}$ is the similarity calculated from the camera based unit. The symbol M^{c}_{j} and M^{a}_{j} stand for the fuzzy measures of camera recognition unit and the accelerometer unit of j-th gesture template separately. Then, the fused similarity calculated by Choquet integral [14] is shown as

$$C_{i,g,t}\left(S^{c}_{i,g,t}, S^{a}_{i,g,t}, M^{c}_{j}, M^{a}_{j}\right) = \begin{cases} S^{a}_{i,g,t} + \left(S^{c}_{i,g,t} - S^{a}_{i,g,t}\right) \times M^{c}_{j}, \text{if}(S^{c}_{i,g,t} \geq S^{a}_{i,g,t}) \\ S^{c}_{i,g,t} + \left(S^{a}_{i,g,t} - S^{c}_{i,g,t}\right) \times M^{a}_{j}, \text{if}(S^{c}_{i,g,t} < S^{a}_{i,g,t}) \end{cases} . \quad (2.2)$$

The difference of similarity error between i-th training data of g-th gesture and the misrecognized gesture is calculated as

$$f(i, g, M^{a}_{1}, M^{c}_{1}, \dots, M^{a}_{G}, M^{c}_{G}) = \max_{k \in \{1,\dots,T\}} \{C_{i,g,k}\} - C_{i,g,g} \quad (2.3) ,$$

where G denotes the number of gesture, and T is the number of templates.

Then, the problem of estimating the optimal fuzzy measure $({}^{*}M^{a}_{1}, {}^{*}M^{c}_{1}, \dots, {}^{*}M^{a}_{G}, {}^{*}M^{c}_{G})$ from all training data can be modeled as

$$\min_{(M^{a}_{1}, M^{c}_{1}, \dots, M^{a}_{G}, M^{c}_{G})} \sum_{g \in \{1,\dots,G\}} \sum_{i \in \{1,\dots,TN\}} f(i, g, M^{a}_{1}, M^{c}_{1}, \dots, M^{a}_{G}, M^{c}_{G}) , \quad (2.4),$$

where TN is the number of training data for each gesture.

The optimal fuzzy measure value of both units cannot be calculated exactly, but it can be found approximately. A hill climbing based algorithm is employed to find the optimal fuzzy measure. The algorithm is demonstrated as follows:

Step 0: Initialize $(M^{a}_{1}, M^{c}_{1}, M^{a}_{2}, M^{c}_{2}, \dots, M^{a}_{G}, M^{c}_{G})$ randomly in the range of between 0 and 1 by uniform distribution.

Step 1: Calculate the fused similarity from the two units by equation (2.2), and then calculate the recognition rate from the training data.

Step 2: Preserve current parameters if the recognition rate in step 1 is better than the best training rate before.

Step 3: For each gesture g in {1, 2, ..., G}, similarity error of all training data is calculated by adjusting the parameters M_g^a and M_g^c one step forward and backward as in equation (2.3). And then the minimum of these similarity errors is found.

Step 4: If the minimum similarity error calculated in step 3 is greater than the similarity error before parameter adjustment, then it has reached to the local minimum similarity error, so exit the loop.

Step 5: Update the changes if the minimum similarity error is less than the current error, and then go back to Step 1.

2.3. Experiment on Multimodal Gesture Recognition

2.3.1 Experimental Environment

To demonstrate the validity and applicability, the proposed method is confirmed by the typical gestures of the Mascot Robot System. The experiments on the Mascot Robot System are complex and require a lot of time, therefore, the authors here did the offline experiments.

The proposed method is implemented on each of notebook PCs with Intel® Core2Duo 2.54GHz, 2GB of RAM by using Visual C++ 2008 and OpenCV 2.1 library for video processing on Windows system. Two 3D accelerometers and a web camera are used to track movements of the body parts in experiments. The wearable accelerometers (W=45 ×D=45 ×H=20mm), manufactured by Microstone Inc., are used to get three dimensional acceleration data of human body movements in 3D Euclidean space, and the accelerometer transmits these 3D movement data via a Bluetooth connection. The two 3D accelerometers are attached on the wrist of participants (shown in Figure 2.6). The video sequences are captured by a Logicool Qcam(R) Connect camera mounted on a notebook PC with 30 fps rate and 320x240 pixels resolution.

Figure 2.6. 3D Wearable Wireless Accelerometer

Eight types of typical emotional-gestures are employed in the experiments. Each name and description of the eight gestures is;

(**G1**) Toast: Raising a hand with a glass upwards from the waist level.

(**G2**) Throw dart: Moving a hand with a dart forward speedily from the head level.

(**G3**) Victory: From the relaxed position besides legs, the right hand is raised up to a level between the chest and the head. (This motion expresses happiness when the dart hits the target.)

(**G4**) Banzai: Both hands are first put down besides legs and then both are held right up. (This motion expresses happiness when the dart hits the target.)

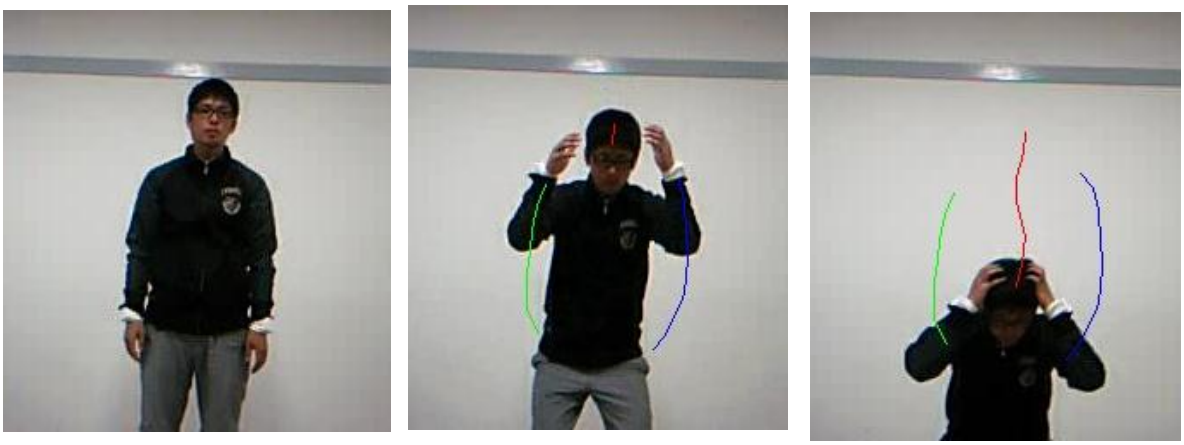


Figure 2.7. Motion of G5 (Squatting with hands over the head).

(The three lines are the trajectory of the three body parts.)

(**G5**) Squatting with hands over the head: From the relaxed position the besides legs, both hands are put on the head and human participant squats. (This expresses disappointment when the dart loses the target.)(Figure 2.7)

(**G6**) Face covering: Both hands are first put down besides legs and then right hand is held at face. (This expresses sadness when the dart loses the target.) (Figure 2.8)

(**G7**) Guiding: From their relaxed position, the right hand beside legs is swung up to the point towards the right direction.

(**G8**) Bye-bye: Both hands are first put down besides legs and then right hand is waved iteratively at right side.

Figure 2.8. Motion of G6 (Face covering Gesture).

(The line shows the trajectory of right hand)

The eight types of gestures are performed by 6 experimenters with five times' repeats. Finally 240 gesture data are collected.

2.3.2 Training the Value of Fuzzy Measure

Cross Validation [25] is widely used for evaluating and comparing learning algorithms. In this experiment, the 5-fold Cross Validation, i.e., the dataset is divided into five parts where each part is taken as the test data and the rest parts are taken as the training data, is employed for train and test the proposed system. As mentioned in 2.2.1.3, the minimum selection is used to choose the suitable template. To calculate the optimal fuzzy measure values, the following three steps are done;

(1) Similarity calculation: Calculate the similarities of camera and accelerometers units separately for all of the training data.

(2) Similarity normalization: Each training data is compared to the eight templates. That is, it will get eight similarities for each input data. Then these eight similarities are normalized in the range of [0, 1], i.e., divided by the maximum of the eight similarities.

(3) Finding optimal measure value: The algorithm illustrated in 2.2.3 is implemented for calculating the optimal fuzzy measures for each gesture.

Because the algorithm mentioned in 2.2.3 is the one relied on initial values, it is executed for many times to get the approximate global optimal fuzzy measures. The fuzzy measures which achieve best recognition rates from the training data are considered as the optimal fuzzy measures for evaluating test data.

Table 2.1. The Initial and Optimal Values of Fuzzy Measures

Gesture	Camera Unit		Accelerometer Unit	
	Initial	Optimal	Initial	Optimal
G1	0.5	0.6	0.9	0.85
G2	0.3	0.4	0.9	0.6
G3	0.5	0.45	0.8	0.3
G4	0.5	0.8	0.4	0.3
G5	0.4	0.7	0.3	0.5
G6	0.4	0.25	0.1	0.75
G7	0.1	0.25	0.9	0.55
G8	0.5	0.5	0.9	0.75

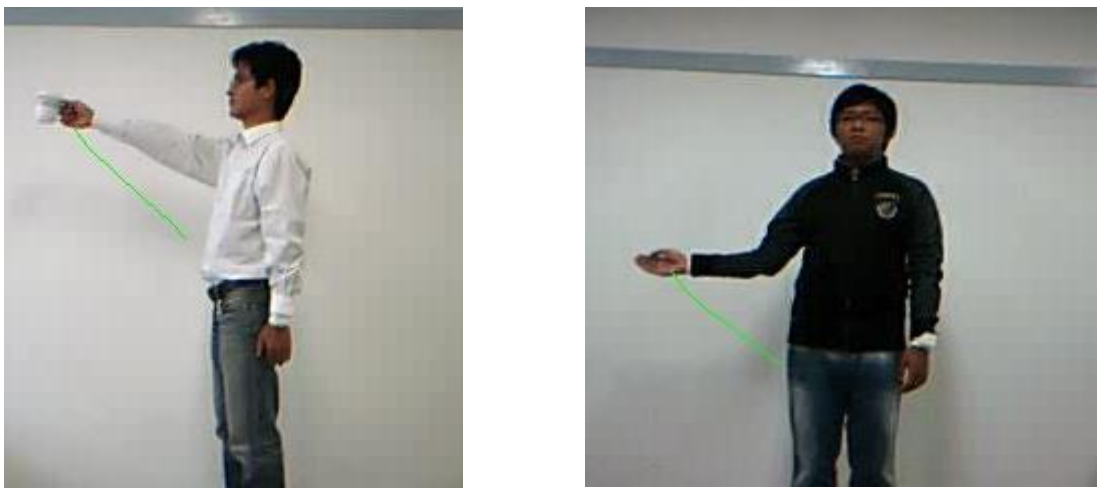
The 5-fold Cross Validation is applied for 100 times with step value of 0.05. Table 2.1 demonstrates the initial and estimated optimal fuzzy measures of the two units in the training stage.

2.3.3 Result of Multimodal Gesture Recognition and Discussion

The recognition result of the camera based and the accelerometers based units are given in Table 2 and Table 3. According to Table 2, G1 (toast gesture) shows the worst result of the camera based recognition unit. Because G7 (guide gesture) has the similar trajectory to G1 (Figure 2.9), 10 data of G1 are misrecognized as G7 and 6 data of G7 are misrecognized as G1. For some gesture (i.e. G6) obtained better recognition result in the camera based recognition unit, while some gestures (i.e. G7) got better recognition result in the accelerometers unit.

Table 2.2. Gesture Recognition Results of Camera based Recognition Unit

Gesture	Gesture category								
	G1	G2	G3	G4	G5	G6	G7	G8	Total
G1	**18**	0	1	0	0	1	10	0	30
G2	2	**19**	0	0	0	0	9	0	30
G3	0	0	**20**	0	0	2	8	0	30
G4	0	0	5	**24**	0	1	0	0	30
G5	0	0	0	0	**30**	0	0	0	30
G6	2	0	0	0	0	**25**	3	0	30
G7	6	0	4	0	0	0	**20**	0	30
G8	0	0	1	0	0	0	1	**28**	30



(a) G1 (toast gesture) (b) G7(guide gesture)

Figure 2.9. The Trajectories of G1 and G7 (shown in green)

Table 2.3. Recognition Results of Accelerometers-based Recognition Unit

Gestures	Gesture category								
	G1	G2	G3	G4	G5	G6	G7	G8	Total
G1	**14**	0	3	10	1	1	0	1	30
G2	0	**30**	0	0	0	0	0	0	30
G3	0	0	**9**	5	13	2	0	1	30
G4	0	0	0	**16**	14	0	0	0	30
G5	0	0	0	6	**24**	0	0	0	30
G6	0	0	4	4	7	**15**	0	0	30
G7	0	0	0	0	0	0	**30**	0	30
G8	0	0	0	0	0	0	0	**30**	30

An example of fusing similarities by Choquet integral is illustrated in table 4. The input gesture is G4.The second and third rows show the normalized similarities comparing with the eight templates by the two recognition unit. It will be misrecognized as G5 and G3 separately. The forth row present the fused similarity of the proposal based on trained fuzzy measure and equation (2.2). After the fusion process, it obtains biggest similarity (0.89) comparing with the forth template. So the input gesture is recognized as G4 by the proposed method though it is misrecognized by the accelerometer recognition unit and the camera recognition unit.

Table 2.4. Example of Fusing Similarities by Choquet Integral

	G1	G2	G3	G4	G5	G6	G7	G8
Accele-rometer Only	0.44	0.62	0.48	0.99	1.00	0.46	0.36	0.55
Camera Only	0.93	0.05	1.00	0.84	0.23	0.82	0.93	0.66
Fused Similarity	0.73	0.39	0.71	0.89	0.62	0.55	0.50	0.61

The recognition rate of training data and test data is shown in Table 2.5 with its graph illustration in Figure 2.10. By calculating the optimal fuzzy measure of Choquet integral, the proposed fusion method achieves an average recognition rate of 96.0%, which is 22.7% higher than the average recognition rate of both units. Six of all eight gestures achieve higher recognition rate than the recognition rate of each unit. The proposed multi-modal recognition method obtains the average evaluation value of 96.0% among the eight types of gestures.

This result confirms the practical applicability of the proposed multi-modal recognition method compared with the accuracy of the camera recognition unit and the accelerometer recognition unit.

Table 2.5. Fusing Result of Choquet Integral Comparing with Each Unit

	Camera Unit	Accelerometer Unit	Choquet Integral	
			Train	(Test)
G1	60.0%	46.7%	90.8%	92.6%
G2	63.3%	100.0%	100.0%	97.0%
G3	66.7%	30.0%	99.9%	90.0%
G4	80.0%	53.3%	100.0%	99.7%
G5	100.0%	80.0%	100.0%	100.0%
G6	83.3%	50.0%	94.3%	92.1%
G7	66.7%	100.0%	100.0%	100.0%
G8	93.3%	100.0%	95.9%	96.7%
Average	76.7%	70.0%	97.6%	96.0%

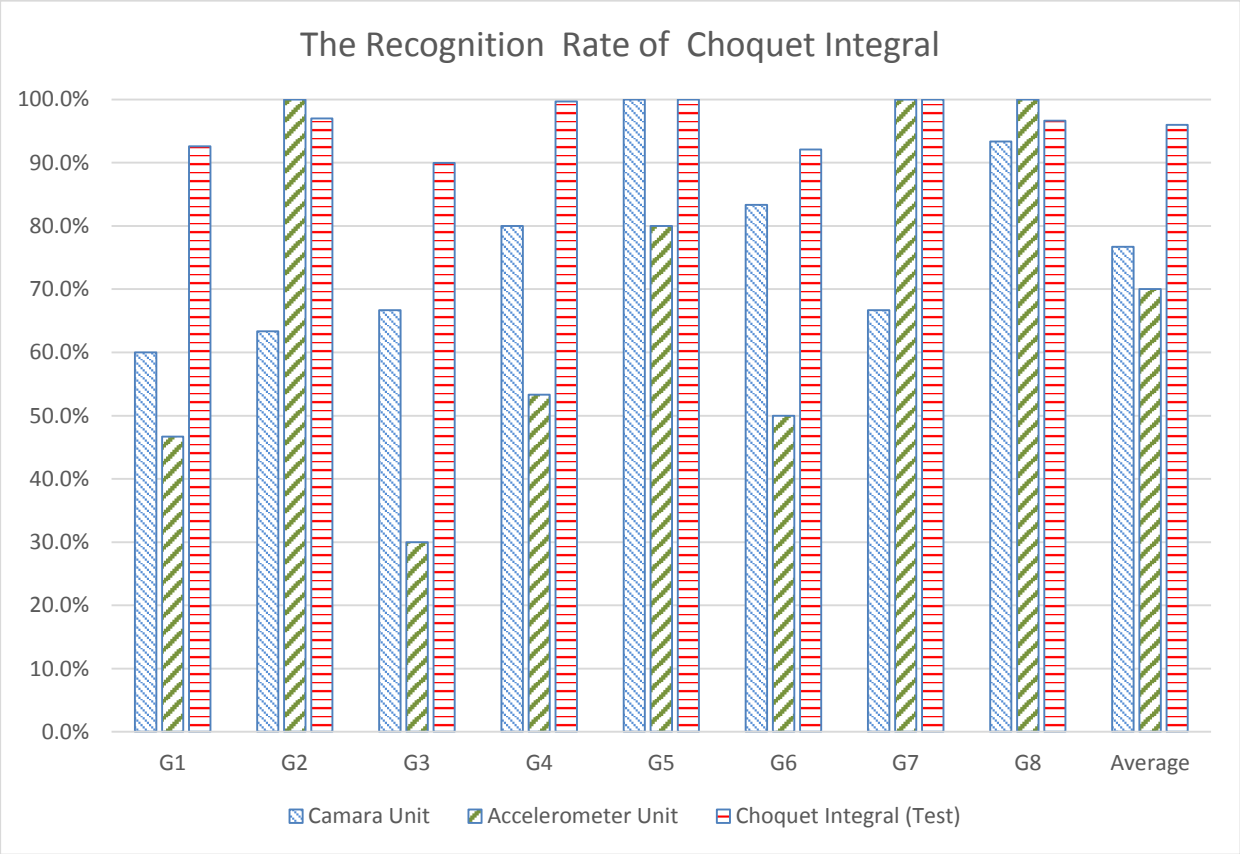


Figure 2.10. Result of Choquet Integral Based Fusion

Sugeno Integral [26] is another famous integral method in fuzzy. The Sugeno Integral is also implemented and compared with the result of each recognition unit. The recognition rate of Sugeno Integral based fusion is shown in Table 2.6 and its graph illustration is shown in Figure 2.11. As shown in the table, Sugeno Integral based fusion method also achieves a high accuracy of 94.3% which is 21% higher than the average recognition rate of camera based gesture recognition unit and accelerometer based gesture recognition unit. Because the operations of Sugeno Integral are only Max and Min operations, it bring out a low recognition rate which is 1.7% lower than comparing with Choquet integral based fusion method.

The Opposite-Sugeno Integral [27] is also implement in the proposed multi-modal gesture recognition system. The result of Opposite-Sugeno Integral based fusion is shown in Table 2.8 and Figure 2.12. The Opposite-Sugeno Integral based fusion method achieves an accuracy of 95.3% which is 1% higher than the Sugeno Integral based fusion method and is 0.7% lower than the recognition rate of Choquet integral based fusion method. This is because that the mathematic operations of Opposite-Sugeno Integral is more abundant than the Sugeno Integral in which only the the min operation and max operation is used for computing, while the Choquet integral use the mathematics operation like addition operation, minus operation and multiply operation which is more precise and powerful than the operations in Opposite-Sugeno Integrals.

Table 2.6. Result of Sugeno Integral Based Fusion

	Camera Unit	Accelerometer Unit	Sugeno Integral	
			Train	Test
G1	60.0%	46.7%	81.5%	76.6%
G2	63.3%	100.0%	99.9%	100.0%
G3	66.7%	30.0%	93.3%	88.4%
G4	80.0%	53.3%	99.2%	98.3%
G5	100.0%	80.0%	100.0%	100.0%
G6	83.3%	50.0%	94.6%	92.3%
G7	66.7%	100.0%	100.0%	100.0%
G8	93.3%	100.0%	98.7%	98.9%
Average	76.7%	70.0%	95.9%	94.3%

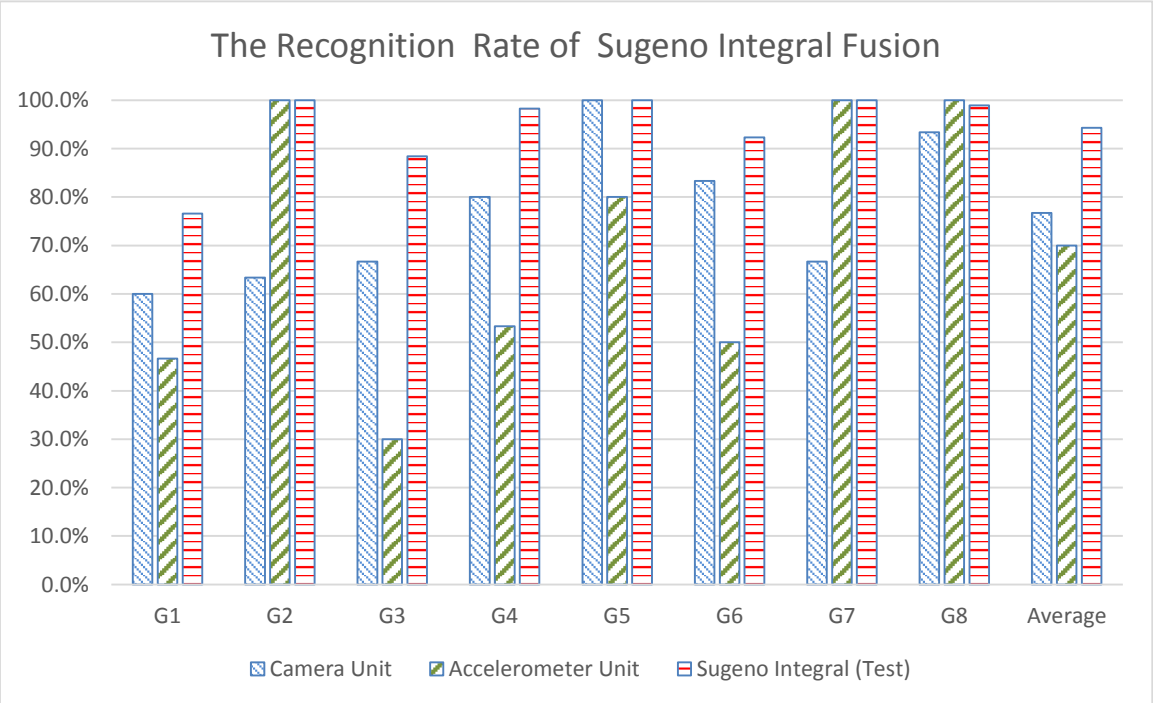


Figure 2.11. Result of Sugeno Integral Based Fusion

Table 2.7. Result of Opposite-Sugeno Integral Based Fusion

	Camera Unit	Accelerometer Unit	Opposite-Sugeno Integral	
			Train	Test
G1	60.0%	46.7%	90.4%	79.9%
G2	63.3%	100.0%	100.0%	100.0%
G3	66.7%	30.0%	98.6%	99.0%
G4	80.0%	53.3%	99.6%	99.7%
G5	100.0%	80.0%	100.0%	100.0%
G6	83.3%	50.0%	87.9%	87.1%
G7	66.7%	100.0%	100.0%	100.0%
G8	93.3%	100.0%	96.2%	96.7%
Average	76.7%	70.0%	96.6%	95.3%

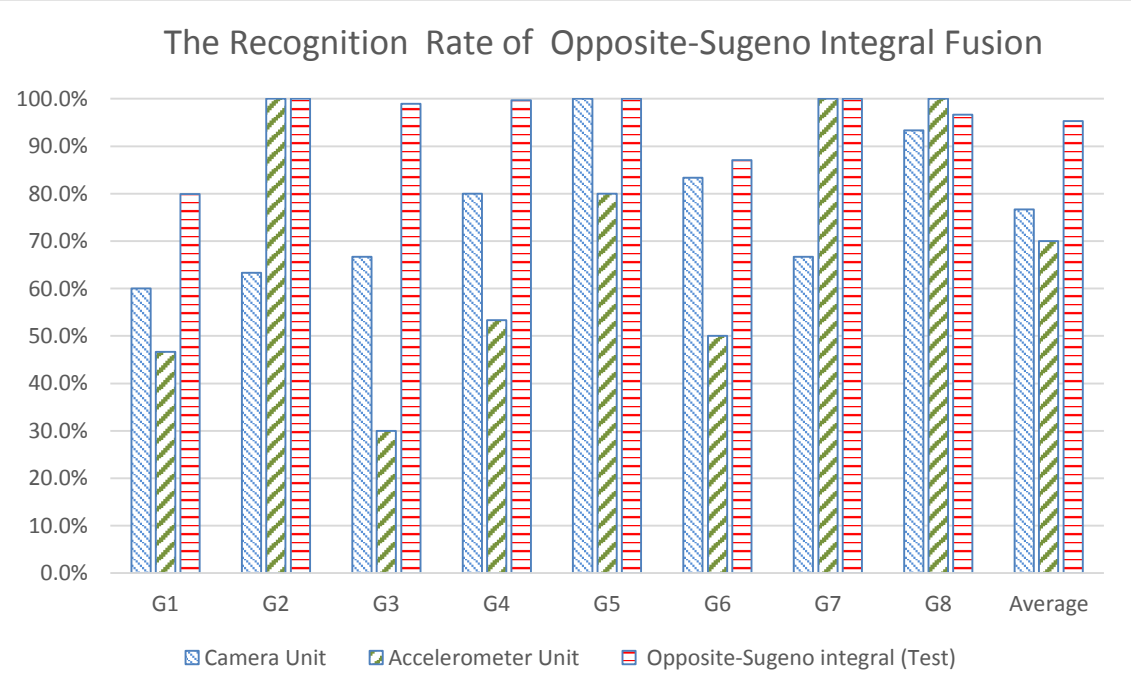


Figure 2.12. Result of Opposite-Sugeno Integral Based Fusion

Table 2.8. Result of Linear Weight Based Fusion

	Camera Unit	Accelerometer Unit	Linear Weighted	
			Train	Test
G1	60.0%	46.7%	83.7%	76.6%
G2	63.3%	100.0%	99.9%	98.7%
G3	66.7%	30.0%	93.2%	93.9%
G4	80.0%	53.3%	90.1%	90.7%
G5	100.0%	80.0%	100.0%	100.0%
G6	83.3%	50.0%	89.5%	87.5%
G7	66.7%	100.0%	98.7%	99.0%
G8	93.3%	100.0%	96.0%	96.6%
AVG	76.7%	70.0%	93.9%	92.9%

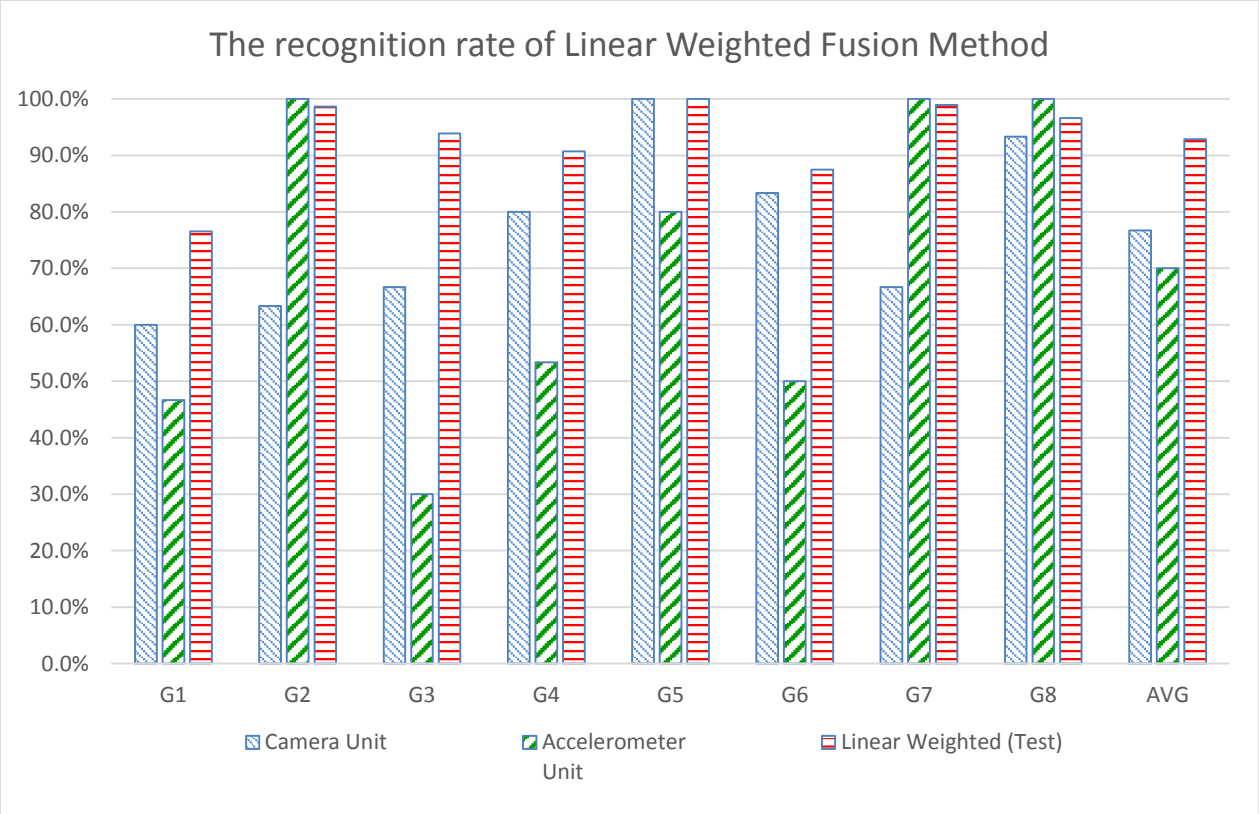


Figure 2.13. Result of Linear Weight Based Fusion

Linear weighted fusion is one of the efficient and widely used methods. The weight of each gesture is trained by the same hill climbing algorithm. The fusion result is shown in Table 2.8 and Figure 2.13. This results show that Choquet integral based fusion is more powerful than the linear weighted fusion. It can be concluded that the Choquet integral based fusion method is more robust than the Sugeno Integral based fusion method, the Opposite-Sugeno Integral based fusion methods and the linear weighted based fusion method.

2.4. Chapter Summary

A Choquet integral based multimodal gesture recognition method is proposed by fusing the information of camera and 3D accelerometer sensors. The proposed multimodal recognition system consists two gesture recognition units which are the camera recognition unit and the 3D accelerometers recognition unit. First, temporal sequence data of each unit are extracted, and then the gesture similarities are calculated by AMSS algorithm. Finally, the similarities calculated from the two recognition unit are fused by Choquet integral fusion method.

To demonstrate the validity, the proposed method is confirmed by the eight types of typical gestures in the Mascot Robot System research project, where six participants are invited to repeat the gestures for five times. The recognized gestures in the system are eight types of typical human emotional-gestures in the scenario. Consequently, the proposed multimodal gesture recognition system achieves an accuracy of 96.0%, which are about 20% improved comparing to the average accuracy of the two recognition unit. To show the advantage of the Choquet integral fusion method, the result of Choquet integral fusion method are compare with the Sugeno Integral based fusion

method, the Opposite-Sugeno Integral based fusion methods and the linear weighted based fusion method. It is obviously that the Choquet integral based fusion method is more powerful than the Sugeno Integral based fusion method, the Opposite-Sugeno Integral based fusion methods and the linear weighted based fusion method.

Chapter 3

Deep Level Situation Understanding

Human may hide their real emotions and intentions in casual communication among humans. But others may be able to understand their real emotions and intentions to some extent by understanding the spoken contents, voice tones, and facial expression changes. Robots are also expected to be competent to these kinds of deep level communications. Although speech recognition, gesture/posture recognition, emotion recognition, intention estimation, and atmosphere estimation can help robot to comprehend parts of human activities, these approaches are still not sufficient for casual Human Robot Interaction. The audible information (e.g., speech and voice) and visible information (e.g., gesture, posture, and facial expression) are called surface level communication in this thesis, while deep level situation understanding is characterized by unifying the surface level understanding, emotion understanding, intention understanding, and atmosphere understanding by applying thoughtfulness to both universal and agent dependent customized knowledge. The deep level situation understanding framework consists of a gesture/posture recognition module, speech/voice recognition module,

emotion recognition module, intention estimation module, atmosphere understanding module, and knowledge (including universal knowledge and customized agent-dependent knowledge) of the interlocutor.

The deep level situation understanding in casual communication among various agents, e.g., humans and robots/machines, aims at three issues. Firstly, humans must pay special attention to robots in the ordinary human-machine communication systems, but such burden may be reduced if robots have deep level situation understanding abilities. Secondly, in the real world, unnecessary troubles or misunderstandings in human to human communications may sometimes happen but the deep level situation understanding can make it possible to avoid such lower level troubles. The customized agent-dependent knowledge will help to comprehend and avoid misunderstanding. Thirdly, with the consideration of surface level information, emotions, intentions, atmospheres, universal knowledge, and customized agent-dependent knowledge, it will also help to understand the background, habits, and intention of the agent for smoothing natural Human-Robot Interaction, so as to create a peaceful, happy, and prosperous society which consists of humans and various specification robots/machines.

A simulated experiment is established to implement the proposed deep level situation understanding system where meeting-room reservation in a company is done between a human employee and a secretary-robot. Twelve subjects are asked by questionnaires to evaluate the response of the proposed inference system comparing to the responses from familiar people.

3.1 Related Works for Human Robot Interaction

3.1.1 Speech Understanding

Spoken Dialog System (SDS) provides a communication interface between the user and a computer-based system with the limited domain in the manner of speech. Many SDS has been developed in various applications. Most of the SDS are designed in restricted domains such as the telephone based "Let's Go Public" bus information system [28], JUPITER weather information system [29], DARPA travel planning system [30]. Some of the SDS are also proposed for multi domain such as SENECA [31] system for entertainment, navigation and communication, CHAT [32] system for multi-task driving helper system. Multi domain SDSs also have been used in real environments [33][34][35] [36] [37].

In the SDS, the dialogue management plays an important control roles in the spoken dialogue system. Early Dialogue systems are conducted by predefined rules [38][39]. But it is not flexible to adapt the natural dialog flow. A Markov Decision Process based dialog system [40] is proposed. They argued that a dialog system can be mapped to a Markov Decision Process with additional assumption about the state transition probabilities. A reinforcement learning algorithm is employed to find the optimal strategy. The Partially Observable Markov Decision Processes (POMDP) for dialogue modelling [41] is proposed by extending Markov Decision Processes with providing a principled account of noisy observations, and the result outperforms that of Markov Decision Processes based method. The POMDP based dialog systems is more robust because they can handling the errors of speech recognition. Some researchers also tried to combine the traditional knowledge based dialog management design with reinforcement learning

based dialog management to reflect domain dependent business rules and to reduce the policy space [42][43]. Traditional reinforcement learning based dialogue management systems require a large amount of training data to learn the optimal policies. To tackle this problem, methods for generate simulated dialogues are also proposed [44] [45]. An approach for optimize dialogue policies are proposed by integrating the reinforcement and supervised learning [46]. The advantage of this approach is that it could eliminate the need for large amounts of dialogue data. Some researcher also proposed methods by restricting the possible actions based on conventional rules in the POMDP framework [47]. In this approach, the optimization process works faster and are more reliable than the classical POMDP.

3.1.2 Gesture/Posture Understanding

Computer vision based human gesture recognition system often gets moving information of body parts (hands and head) by applying skin color tracking method. It is, however, easy to have a noise effect by the objects which have similar color of the skin. On the other hand, application of accelerometer data to gesture recognition is an emerging technique to improve recognition performance, e.g., accelerometer based control system [48], and accelerometer based recognition system for recognizing personalized gesture [49].

A multimodal gesture recognition system is proposed in [13], where information of both 3D acceleration and camera sensors are combined based on fuzzy logic. In their study, when the similarity calculated from the acceleration recognition units is greater than a given threshold value, the recognition result from the accelerometer unit is taken

as the final result. In reverse, image recognition unit processes the candidate gestures that come from acceleration unit to get final result. How to decide the given threshold, however, should be investigated to apply the method for other cases of gestures in general.

To deal with this problems, a Choquet integral based multimodal gesture recognition method is proposed [22], where Choquet integral is employed for fusing the similarities of the camera and 3D acceleration recognition units. By calculating the optimal fuzzy measures of the camera-based recognition unit and the accelerometer-based recognition unit, the gesture recognition system achieves a high recognition rate for eight types of gestures.

3.1.3 Emotion Understanding

An automatic real-time capable continual facial expression recognition system is proposed [50] based on Active Appearance Models (AAMs) and Support Vector Machines (SVMs) where face images are categorized to seven emotion states (neutral, happy, sad, disgust, surprise, fear, and anger). An individual mean face is estimated over time to reduce the influence of individual features.

In the casual communication, emotion may be expressed on both facial expression and voice. A multimodal emotion recognition system is proposed [51] to recognize emotions from audio sequence and static images.

3.1.4 Intention Understanding

Estimating the intention of human is also important in Human-Robot Interaction. An intention reason algorithm [52] is proposed based on bidirectional associative

memories driving support system. A maximum entropy based intention understanding method [53] is proposed for understanding the intention of speech in a dialog system.

3.1.5 Atmosphere Understanding

In many-to-many communication, e.g., a conference with twenty participants, it may be not easy to identify the attitude, mood, and emotion of each individual, and instead the atmosphere of the whole gathering becomes an important issue for smooth communication. To reflect the uncertainty and subjectivity of the atmosphere as well as its effect on the emotions of the individuals in many-to-many communication, a concept of Fuzzy Atmosfield (FA) is proposed [54] to represent the atmosphere being created in the process of interactive communication.

To adapt robot's behaviour for smooth communication in human robot interaction, a Fuzzy Production Rule based Friend-Q learning method (FPRFQ) is introduced [55]. Based on the FPRFQ, a behaviour adaptation mechanism is proposed to solve the robots' behaviour adaptation problem.

3.2 Concept of Deep Level Situation Understanding

Although speech understanding, gesture/posture understanding, emotion understanding, intention understanding, and atmosphere understanding can help robot to comprehend parts of human activities. These approaches are still difficult to understand human activity deeply in casual Human Robot Interaction. People usually hide their real emotions, intentions, and opinions and show them in another indirect/different way. These kinds of information are just a reflection of the real emotions, intentions, and feelings.

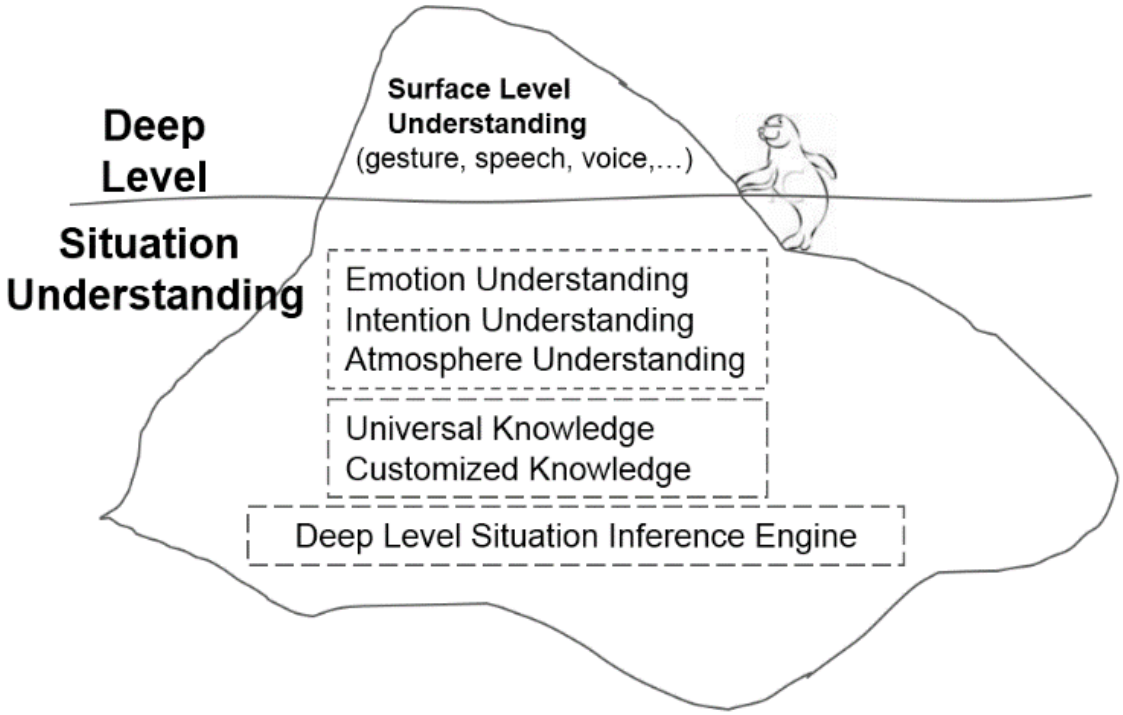


Figure 3.1. The Relationship of the Surface Level Understanding and the Deep Level Situation Understanding.

The audible information (e.g., speech and voice) and visible information (e.g., gesture, posture, and facial expression) are just the surface information of humans. Thus the understanding of such surface information is called surface level understanding in this thesis. If the understanding level is illustrated by an iceberg, the audible and visible information is just like a tip of the whole iceberg above the sea level, while there still remains more information hidden under the sea level such as emotion, intention, and atmosphere. In contrast with surface level understanding, the deep level situation understanding is characterized by unifying the surface level understanding, emotion understanding, intention understanding, and atmosphere understanding by adding a thoughtfulness function to the inference engine on the situation inference system consisted of both universal knowledge and agent dependent customized knowledge. The

relationship between the surface level understanding and the deep level situation understanding is illustrated in Figure 3.1.

Moreover, customized knowledge and thoughtfulness should also be considered for casual human robot communication. Customized knowledge and thoughtfulness are detailed in 3.2.1 and 3.2.2 respectively.

3.2.1 Customized Knowledge

Why the communications between friends are usually smoother than the communications between strangers? It's because friends usually know each other very well. Friends have special knowledge, e.g., tempers, habits, and means of expression, of each other. These special knowledge may help to avoid misunderstanding in human-human communications.

These special customized knowledge should also be considered in the humans-robots communication for realizing the smooth communications as human-human. There are two kinds of customized knowledge data. (1) The data that characterizing the normal state (including the normal tones, normal facial expression) of a people. Because people may show their pleasure and anger in different ways. Some people may keep smiling face all days. When angry, they may just keep silent. For these people, smiling is their normal state. (2) The data featuring people's habits (e.g., his/her favorite, frequency of doing something). The habit data is obtained from the history communications. Usually these kinds of data is known to friends.

3.2.2 Thoughtfulness Communications

Human usually consider emotion and intention of their conversation partners. There are many instances of deep level situation understanding in daily life. Suppose you visited a convenience store and want to buy a fountain pen. In this case, you may ask the shop assistant that “Do you have a fountain pen?” The shop assistant may guess that you want to buy this kind of pen. Even if it is a yes-no question, neither “yes” nor “no” is expected to end the conversation. If there are fountain pens in the shop, the shop assistant will guide the customer to the specific location of the fountain pens. If they do not have this kind of pen, in order to provide satisfactory service to the customer, they will guide the customer to the shop where the fountain pen can be purchased. Imagine a lady usually goes to a cafe for her favorite coffee and dessert. The waiter/waitress in the cafe knows the preference of their regular customer. When this lady just orders “the usual one” it is no doubt that the waiter/waitress will understand the meaning and bring the desired drink and dessert to her.

Because thoughtful communication can make the interlocutor feel comfortable, the robots should also have the ability of thoughtfulness inference for human-human like communication.

3.3 Inference Framework for Deep Level Situation Understanding

Since people usually hide their real emotions, intentions, and opinions and show them in another indirect/different way. Not only the surface level information, i.e. visible and audible information, is important for human-robot communication, but also emotion

information, intention information, and atmosphere information should be considered for human-like reactions.

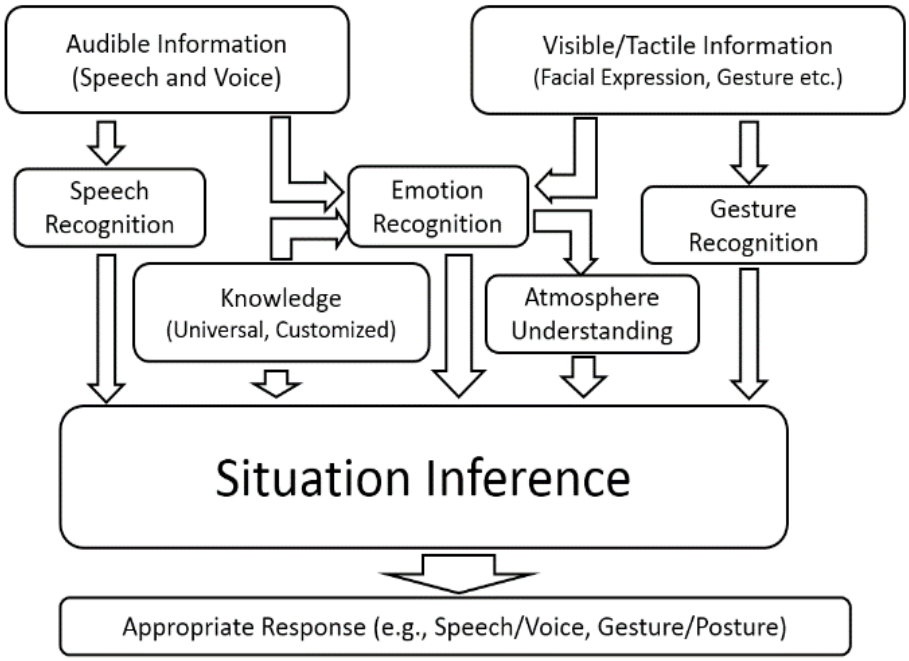


Figure 3.2. Multimodal Framework for Deep Level Situation Understanding.

The illustration of multi-modal framework for deep level situation understanding is shown in Figure 3.2. The audible information and visible/tactile information are obtained by microphones and cameras/tactile sensors. The speech content is recognized by speech recognition method (e.g., Julius Library [56]). People may express their emotion in different ways. To estimate the real emotion of a people, the friend-level knowledge, i.e., customized knowledge of the people, is necessary. The face features and voice features of normal state is used to train the classifier. Then the real emotion state of the interlocutor is estimated by the trained classifier. Atmosphere is estimated based on the emotion state of the agents. Gestures/Postures is recognized by gesture recognition algorithms [22] from sensors like cameras and accelerometers. The speech contents,

universal and customized knowledge, emotions, atmospheres, and gestures/postures are important input for the deep level situation inference.

The inference flowchart of deep level situation understanding is shown in Figure 3.3. The meaning of the interlocutor is analyzed from the verbal information (speech contents) and the non-verbal information (gestures/postures). Then the intention is estimated based on the analyzed meaning and knowledge from historical dialogs. Thoughtfulness is inferred based on the intention and the thoughtfulness knowledge. Finally with the comprehensive consideration of the current emotion state, atmosphere, universal and customized agent-dependent knowledge, and thoughtfulness, suitable response (speech, voice, and gesture/posture) is selected from a lookup table and then the system output the final reaction. The detail of situation inference is mentioned in 3.4.

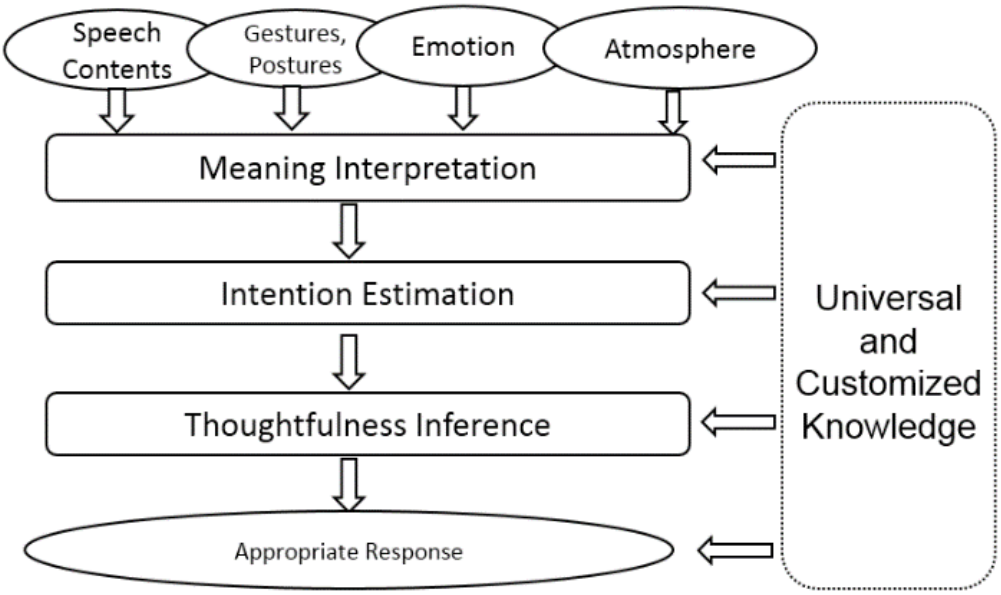


Figure 3.3. The Flowchart of Deep Level Situation Inference.

3.4 Situation Inference System

This research is a part of the project called "Multi-Agent Fuzzy Atmosfield", which is supported by the Japan Ministry of Education, Culture, Sports, Science and Technology. The project is aimed at realizing human-like natural communication (called casual communication) among multi-agents (e.g., humans and robots/machines). And the project also contains several sub research themes, such as research of deep level situation understanding, atmosphere understanding for human robot interaction, deep level emotion understanding and so on. This research mainly focus on inferring based on the text of utterance.

3.4.1 Meaning Interpretation

Verbal and non-verbal communications are two natural ways in humans-robots communications. The meaning of the non-verbal activities can be recognized directly, e.g., multimodal gesture recognition system [22], while analyzing the meaning of verbal information is complicated. The verbal information (i.e., speech) is able to be transferred into text sentence by means of speech recognition library (e.g., Julius [56] for Japanese speech recognition).

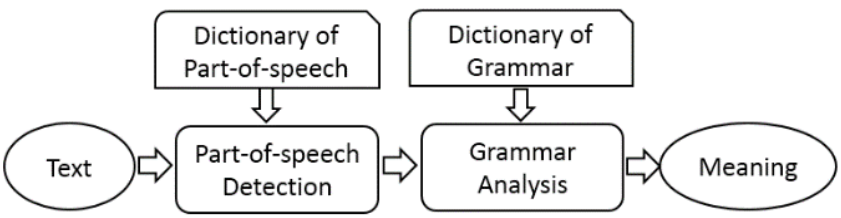


Figure 3.4. The Flowchart of Meaning Analysis.

The flowchart of meaning analysis is shown in Figure 3.4. Firstly, the text of utterance is divided into words list with part-of-speech tags. The boundary of Japanese utterance is determined by a conditional random field method [57].

Secondly, a dictionary of grammar is employed to transfer the words list into meaning string. For example, the utterance "Is there a meeting-room available from 15:00 PM?" is converted to the meaning string, "Query (subject: meeting-room, start-time: 15:00, status: available)".

After getting the meaning of utterances, it is easy to transfer the meaning string into a Structured Query Language (SQL) statement and execute on the knowledge database. The corresponding SQL sentence of previous example may be "Select id from knowledges where note_type = 'meeting-room' and available_time = '15:00' ".

3.4.2 Intention Understanding

The intention of utterances is able to be understood from customized knowledge data which is extracted from the history data of communications. A general communication process between agent A and agent B is illustrated in Figure 5. For the question from agent A, Agent B may reply many kinds of responses (e.g., Response 1, Response 2 … Response N). Agent A may intend different intentions with some frequency for each response of agent B. For example, agent A may purport intention 1, intention 2, and intention 3 with a frequency of P1, P2, and P3. Suppose P2 and P1 were the biggest and second biggest among P1, P2, and P3. If the difference between P2 and

P1 is significantly big, that means agent A purport Intention 2 definitely. Agent B should response to intention 2 directly. If the difference between P2 and P1 is very small, that means agent A may purport Intention 1 or intention 2. Agent B may respond to the question from agent A by "Do you mean intention 2" because P2 is a little bigger than P1. Then agent B could respond to intention 2 directly when agent A asks this question.

Utterances usually contain some important information, e.g., people usually order their favorite food more frequently than the others. These kinds of habit information of a person are some kinds of deep level information which is only known by their friends. These information may be extracted from the utterances.

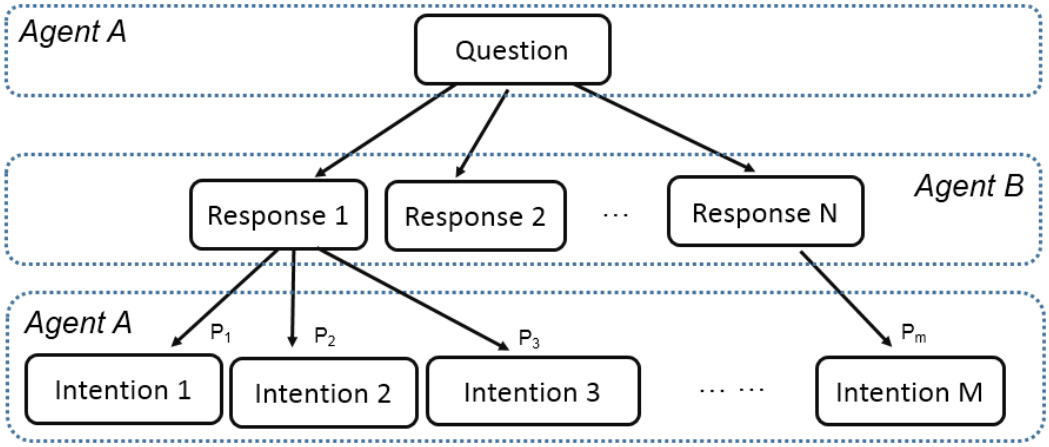


Figure 3.5. Illustration of Conversation between Agent A and Agent B

People may express one meaning in different utterances. For example, utterance "are you busy?" or "are you available" have the same meaning "query (subject: you, status: available)". The meaning strings of dialogs are stored in the database as the customized knowledge for inferring the intention of utterance.

An example is shown in Figure 3.6. Agent A asks "Are you busy?" to agent B which is convert to its meaning string as "query (subject: you, status: available)". As

illustrated in the figure, when agent B is busy and replies “Yes”, agent A may response “sorry to bothering you” .When agent B is available and replies “No”, the frequency of agent A asks for helping is 90% and the frequency of agent A ask for others is 10%. In this situation, if agent B is available, then agent B almost definite that agent A intend to ask for helping. After understanding the intention of agent A, the conversation between two agents may moves forward smoothly.

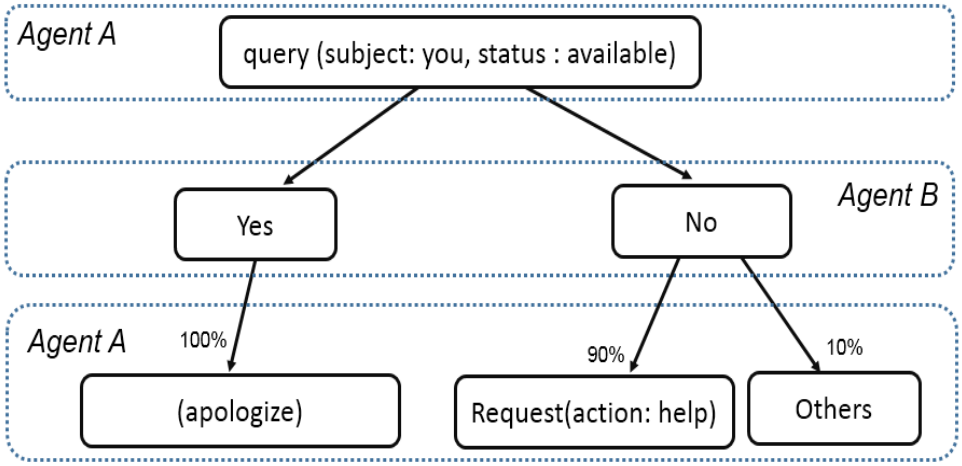


Figure 3.6. An Example of Communication between Agent A and Agent B.

3.4.3 Thoughtfulness Inference

Thoughtfulness, i.e., showing consideration for others, is a high level intelligent action of human that is usually performed between familiars.

The thoughtful response is able to be reasoned by the intention and customized knowledge. Assuming in some company, the TV conference system should be reserved with the meeting room together when holding a video conference with branch companies.

If an employee intend to hold a TV conference with branch companies and just reserves the meeting room, it will be very helpful to remind the user to reserve the TV conference system with the meeting room together.

Table 3.1. Example of Thoughtfulness Knowledge.

Intention	Condition	Response
Reserve room	No room available at that time	Try to revere at other time
Reserve room	The room is not available	Try to reserve other room
Reserve room for remote conference	Only reserved meeting room	Reserve the video conference system
Reserve room for remote conference	Only reserved conference system	Reserve the meeting room

Thoughtfulness knowledge is a known common knowledge of human. For robots, thoughtfulness response may be inferred based on the estimated intention and thoughtfulness knowledgebase. An example of thoughtfulness knowledge is shown in table 3.1.

After intention estimation and thoughtfulness inference processing, response string is arranged. The utterance is generated based on grammars. For simple, a database of responding utterances is used to map the response string into utterance.

In the conversation between an employee and a secretary-robot, if the employee just asks “are you busy?”. The secretary guesses that he is intended to reserve meeting-rooms. If the emotion state of the employee is as usual, the response from the secretary-robot may be “Are you going to reserve a meeting-room?”. If the employee looks sad and abnormal, the secretary-robot may response to his latent request (intention) directly by “Please!”. An example of utterances database is shown in table 3.2.

Table 3.2. An Example of Utterances Database.

Response string	Emotion state of interlocutor	Response Utterance
confirm(action: reservation, object: meeting-room)	Normal	Are you going to reserve a meeting-room?
confirm(action: reservation, object: meeting-room)	Sad	Please!

3.5 Demonstration Scenario of “Routine of a Business Man”

To illustrate the applicability of the proposed deep level situation understanding mechanism, a demonstration scenario, entitled “One day of a businessman”, is established

to narrate several communication activities among five human (i.e., a businessman, a manager, a colleague, a new customer and a wife) and five robot (i.e., a secretary-robot, a colleague robot, a waitress-robot, a therapy-robot and a child robot) in one day. Six episodes of deep level situation communication are created with comprehensive consideration of speech content, gesture, emotion, intention, atmosphere, universal knowledge and customized agent-dependent knowledge, and careful attention function. In the scenario, the businessman is asked to reserve a meeting room with a secretary-robot for remote TV meeting. After reported to his boss, he notices that he made a mistake of reserving meeting schedule. He asks the secretary-robot to help him to change the meeting schedule. After one day's hard working, he goes to a small Japanese-style restaurant with his colleague. Both of the case that a businessman goes to the restaurant as a new customer and the case that the businessman goes to the restaurant as a regular customer are demonstrated for comparing the surface level communication and deep level communication. The communication between the businessman and his wife is narrated in the last scene 6. The eye robots and the therapy robot PARO are shown in Figure 3.7 and Figure 3.8.

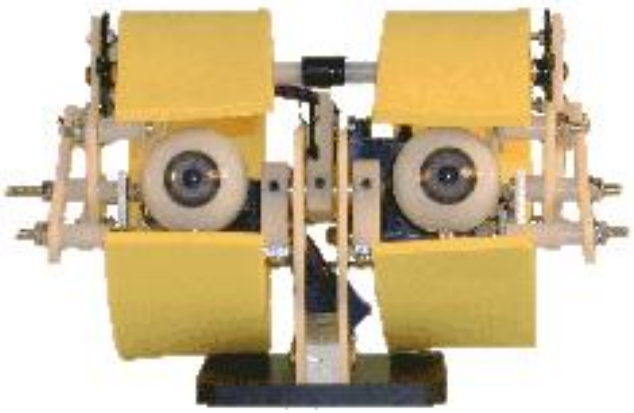


Figure 3.7. The Eye Robot.

Figure 3.8. The Therapy Robot: PARO.

The scripts of scenarios are shown in Table 3.3-3.8.

Table 3.3. Scenario of Reserving Meeting Room

Employee: Are you busy?
Secretary: No, would you like to reserve a room?
Employee: Is the meeting room for 10 people vacant at 3 o'clock this Thursday?
Secretary: They are available from 15:30.
Employee: Great! A quiet room is preferable.
Secretary: How about the regular conference room on the 17th floor?
Employee: Sounds good! It's for a remote conference with the branch office, please reserve it until 17 o'clock.
Secretary: In addition, I will reserve the video conference system, too.
Employee: Thanks!
Secretary: You're welcome. Good luck.

Table 3.4. Scenario of Reporting to Boss.

Employee: Are you busy?
Manager: What's up?
Employee: I reserved the conference room on Thursday from 15:30 for our meeting with the branch office.
Manager: Thanks! Is it on Thursday next week?
Employee: Argh! Sorry, I was wrong.
Manager: Again? You've made so many mistakes recently!
Employee: My apologies. I will correct it immediately.
(The Manager is leaved)
Manager: Be more careful next time. I am leaving for a meeting now. Do your best.
Employee: Oh, Darn it!
Colleague B: Is it all right? Well, it's easy to make mistakes on a date.
Employee: Recently, I have not been able to focus on my work as before.
Colleague B: Please be more careful next time. Let's go to the bar you mentioned after work.
Employee: Is that okay? I still have some data to input.
Colleague A: OK, I'll do it instead. Please take a break later.
Employee: Really? Thank you. I'm going to reserve the conference room again.

Table 3.5. Scenario of Changing the Schedule of Reserved Meeting Room.

Employee: Are you busy now? Secretary: Go ahead! Employee: Excuse me, can you change the meeting with the branch office to Thursday next week? Secretary: Sure! I will check it now. Employee: Yes, please. Secretary: For Thursday next week, all the conference rooms on 17th floor have been scheduled already. How about the 11th floor conference room? Employee: Great! I feel relieved. Secretary: I will also update the reservation of the remote conference system. Employee: Thank you very much. Secretary: You're always welcome.

Table 3.6. Scenario of Entering the Restaurants as a Normal Customer.

Bar Lady: Welcome!
New Customer: Good evening
Lady: Where would you like to sit?
New Customer: There.
Lady: What would you like to order?
New Customer: Errmmm. Grilled fish, please.
Lady: What about the drink?
New Customer: Whisky please.
Lady: Thank you. Please wait for a moment.

Table 3.7. Scenario of Entering the Restaurants as a Regular Customer.

Employee: Good evening Bar Lady: Welcome! 2 people, right? Your usual seats, please. Oh. What's going on? Employee: Errrmmmmm, yep. Lady: Let's forget the unpleasantness with delicious dishes. Do you want the usual? Employee: Yes, thanks! Lady: Wait for a while, please! Is whisky OK? Employee: Thank you! Yes, please. Lady: Peanuts are also served for free. Employee: Really? That's wonderful. Lady: What would you like? Colleague B: Charge me up please. Lady: I'll bring your order. Please wait. Colleague B: Do not worry about today, it'll be OK. Employee: I've made too many mistakes. I am more stress these days. Colleague B: It takes time to work well. Employee: I will work hard. (……) Colleague B: Well, so much for today, anyway, Cheers! Employee: Cheers! (……)

Table 3.8. Scenario of Backing Home.

Employee: I'm home!!
PARO : [Moves because of lighting on]
Employee: Good evening PARO.
Wife: Welcome home. Did you drink?
Employee: Yup!
Wife: Do you want to take bath or have a cup of tea?
Employee: hmm...
Wife: What happened?
Employee: Well...
Wife: I guess, you made mistakes in the office.
Employee: Yep.
Wife: Everything will be OK if you work harder. Just believe in yourself.
Kid: Welcome home dad. I was waiting for you.
Employee: Really? For playing game together?
Kid: I want to play games with dad!
Employee: It's already late now. Let's play tomorrow.
Kid: Hmmm. Dad you're the best.
Wife: Ichiro-kun, get ready to go to the bed.
Kid: OK.
(Kid goes to sleep.)
Employee: I am tired today.

Wife: I will prepare the bath for you. Forget your stress. Enjoy and relax! Employee: Thank you, I will try my best tomorrow.

3.6 Experiment for Deep Level Situation Understanding

3.6.1 Experiment Setting

A simulation experiment is carried out to evaluate appropriate response of the proposed deep level situation understanding mechanism. A company scene is taken into consideration. There is a secretary robot who is supposed to do clerical works such as booking hotel, reserving meeting room, and TV conference system. One employee usually enters secretary room for reserving meeting rooms. Assume the secretary robot has the following knowledge about this employee:

This employee usually comes to the secretary room for some help. When he asks the secretary "Are you busy?", the probability of asking to reserve meeting room is 0.9; the probability of asking for other help is 0.1;

Meeting room and TV conference system should be reserved together for the purpose of holding remote meetings with the branch company.

In the scenario, the employee enters the secretary room for reserving meeting room. When he asks "Are you available now", the secretary is aware of that he is intent to reserve a meeting room. When the meeting room he wanted is not available at that time,

the secretary checks the other available time and recommends his favorite to him. When he tells the secretary, he wants to reserve for a remote conference, the secretary also reserves the TV conference system because it is necessary for holding a remote conference. Again, when he looks sad and wants to change the schedule of reserved rooms, the secretary responds immediately to calm down the employee. When the room he wanted is not available, the secretary also checks the other rooms instead of responding as "No available room" directly. Finally, the secretary also change the schedule of the remote system for the employee.

The experiment system is developed in C++ language. All of the knowledge are saved to a SQL database. The simulation experiment is carried out in Japanese language. Mecab [57] is used to detect part-of-speech for Japanese. The utterances of the employee are used as the input of the proposed inference system. The response of the secretary-robot is inferred as the output of the proposed inference system. The translated scripts of reserving a meeting-room is shown in Table 3.9. The translated scripts for changing the schedule of reserved meeting-room is shown in Table 3.10.

A questionnaire is prepared to evaluate the naturalness of the response inferred from the system comparing to the response from familiar people. The naturalness can be rated by 5 grades, i.e., natural, a little natural, normal, a little unnatural, and unnatural. The confidence of ratting the naturalness is queried by 3 grades, i.e., confident, normal, and diffident. Part of the questionnaire is shown in Table 3.11.

Table 3.9. Script of Reserving a Meeting Room.

Employee: Are you available now?
[O1]Secretary-robot: Are you going to reserve a meeting-room?
Employee: Is the meeting room available after 3 PM this Thursday?
[O2]Secretary-robot: It could be reserved after 3:30 PM.
Employee: Great! A quiet meeting room will be better.
[O3]Secretary-robot: Yes, it's the meeting room on the 17th floor as always.
Employee: Fine. Please reserve it till 5 pm, as we'll have a meeting with our branch office.
[O4]Secretary-robot: No problem. I'll also reserve the TV conference system for you.
Employee: Thank you so much!
Secretary-robot: You're welcome.

Twelve subjects are invited to rate the eight output utterance from the five options, where "natural", "a little natural", "normal", "a little unnatural", and "unnatural" are assigned as 1, 0.75, 0.5, 0.25, and 0 respectively. The confident are taken as weight of evaluated naturalness, where "confident", "normal", and "diffident" are mapped to 1, 0.5, and 0 respectively.

Table 3.10. Script of Changing the Schedule of Meeting.

Employee: [looks sad] Are you available now?
[O5] Secretary-robot: Yes, please.
Employee: Sorry, may I change the reservation of the meeting with branch office to next Thursday?
[O6] Secretary-robot: Got it! I'm checking the status of meeting room right now.
Employee: Great. Please help to check.
[O7] Secretary-robot: All of the meeting rooms on the 17th floor have been reserved on next Thursday. How about meeting room on the 11th floor?
Employee: Thanks a lot! You really saved me!
[O8] Secretary-robot: The reservation of the TV conference system also needs to be changed, right?
Employee: Sure. Thank you as always!
Robot: You're welcome!

Table 3.11. Part of the Questionnaire.

One day, the employee enters the secretary room. Employee: Are you available now? Secretary-robot: Are you going to reserve a meeting-room? Based on the background knowledge, The secretary-robot estimates that the employee is intend to reserve a meeting-room. How do you think of the response of this secretary-robot? (1)natural, (2) a little natural, (3) normal, (4) a little unnatural, (5) unnatural How confident do you answering previous question? (1)confident, (2) normal, (3) diffident

3.6.2 Results of Questionnaire Evaluation

Table 3.12. Average Rating of Each Output.

Output Utterance	Weighted Average of Naturalness
O1	0.84
O2	0.83
O3	0.74
O4	0.96
O5	0.65
O6	0.83
O7	0.92
O8	0.96
Average(AVG)	0.84

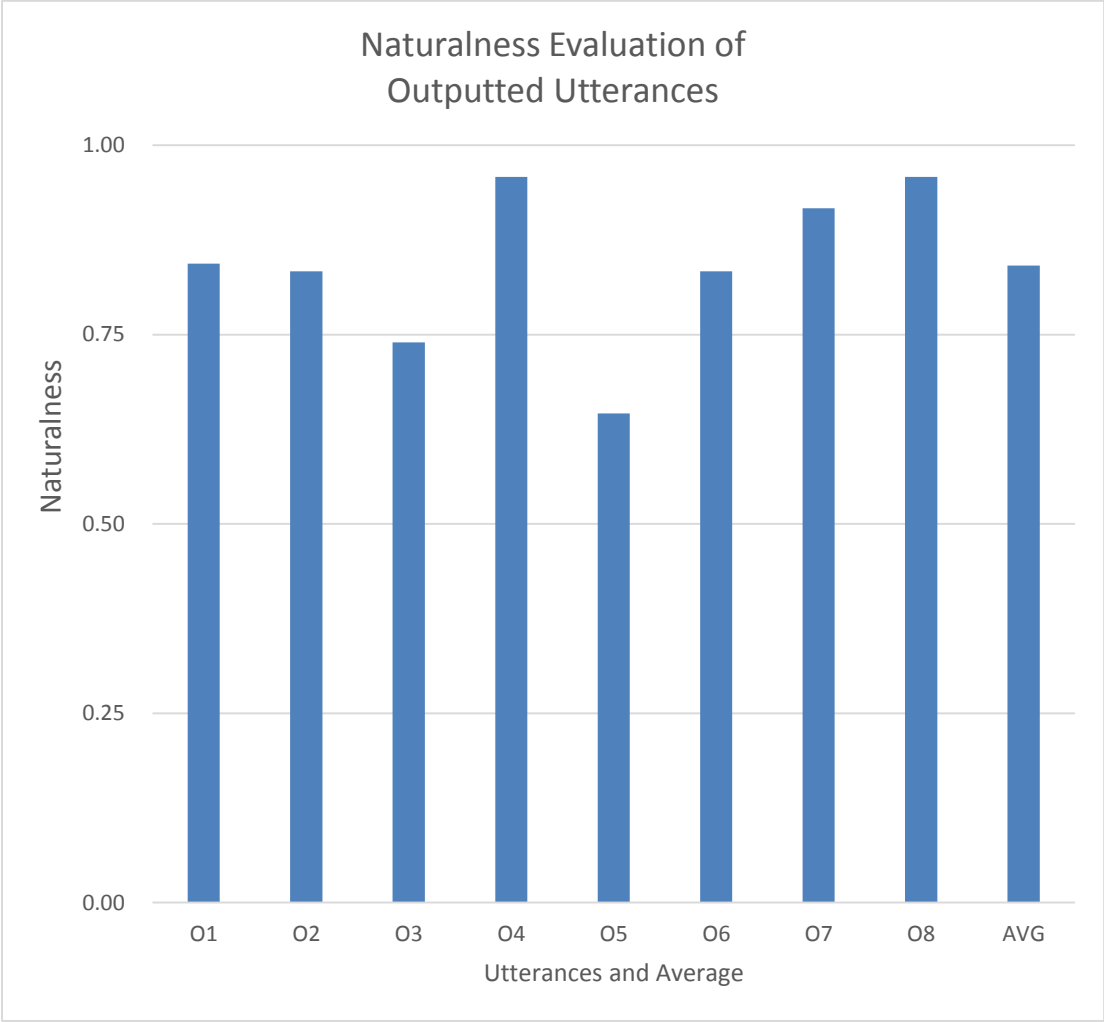


Figure 3.9. Result of Questionnaire Evaluation

The average result of each questions is shown in Table 3.6 and Figure 3.9 where O1, O2… O8 mean outputted utterances marked in Table 3 and Table 4. As shown in the Figure 7, most of the output are evaluated between “a little natural” and “natural”. Only two output (O3 and O5) are in between “a little natural” and “normal”. Finally the weighted average naturalness of the proposed deep level situation understanding in human-robot interaction achieves 0.84 compared to communication with familiar people

in casual communication. It is concluded that the proposed deep level situation understanding may help to accomplish human-level natural communication in casual human robot interaction. Because of personal differences, some output utterances are rated lower than others. Since this employee goes to the secretary room for reserving meeting-room by a probability of 0.9, it could almost definite that the employee is about to reserve a meeting room when he enters the secretary-room and asks "are you available". But some subjects still think that it will be natural to respond by "can I help you", instead of "Are you going to reserve a meeting-room". When the employee looks sad and abnormal, some think that it will be more natural to ask "what's up" than "please".

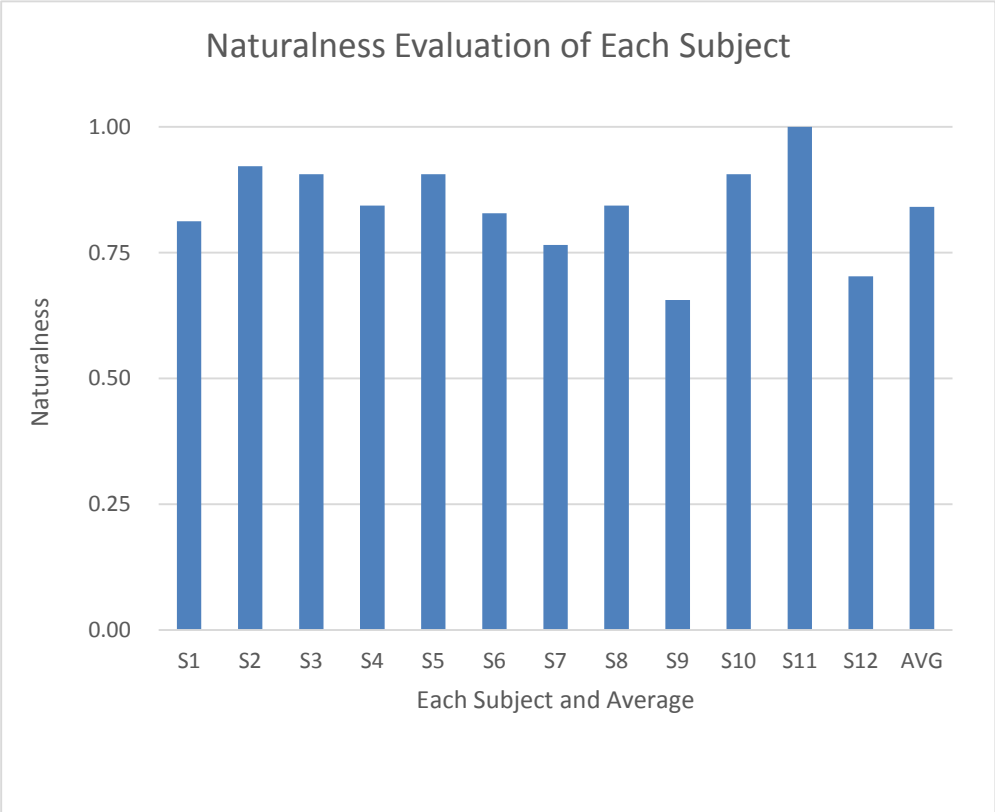


Figure 3.10. Average Rating of Each Subject.

The evaluating result of each subject is shown in Table 3.13 with its graph illustration in Figure 3.10 where S1, S2… S12 stand for the twelve subject. As shown in

Figure 8, ten of the twelve subjects rate the naturalness of utterance between levels of "natural" and "a little natural" while the rest of subjects just rate it as lower than the level of "a little natural".

Table 3.13. Average Rating of Each Subject.

Subjects	Weighted Average of Naturalness
S1	0.81
S2	0.92
S3	0.91
S4	0.84
S5	0.91
S6	0.83
S7	0.77
S8	0.84
S9	0.66
S10	0.91
S11	1.00
S12	0.70
Average(AVG)	0.84

3.7 Chapter Summary

A concept and an inference system of deep level situation understanding are proposed for casual communications among humans and robots/machines. Twelve subjects are asked by questionnaire to evaluate the naturalness of the response of the proposed inference system comparing to the responses from familiar people. The proposed system achieves a naturalness value of 0.84 which is between the ranks of "natural (=1.0)" and "a little natural (=0.75)" comparing to communicate with familiar people. It is concluded that the proposed deep level situation understanding may help to accomplish human-level natural communication in casual human robot interaction.

Not only surface level understanding (e.g., speech/voice recognition, gesture/posture recognition), emotion understanding, intention understanding, and atmosphere understanding but also customized agent-dependent and universal knowledge, and a thoughtfulness mechanism are considered for smoothing and naturalizing communication among humans and robots/machines. The proposal can be applied to the service robot systems to achieve casual communication when interacting with robots/machines. By considering the customized agent-dependent knowledge in human-robot communication, it will help robots/machines to understand the usual way in communication and avoid unnecessary troubles and misunderstandings. With the comprehensive consideration of speech/voice, gesture/posture, emotion, intention, atmosphere, and knowledge (including universal and customized knowledge), the proposal will smooth the communication among humans and robots/machines as well as create a peaceful, pleasant, and prosperous society consisting of humans and various specification robots.

Chapter 4

Conclusions

4.1 Summary of This Thesis

This thesis has presented the concept and inference system of deep level situation understanding for casual communication between human and robot.

In chapter 2, a surface level understanding method, i.e., Choquet integral multimodal gesture recognition method, is proposed for non-verbal casual communication between human and robot. In the Mascot Robot System, there are eight kinds of gestures such as "toast", "throw dart", "victory ", "banzai ", "squatting with hands over the head", "face covering", "guiding" ,and "bye-bye". The proposed multimodal gesture recognition system contains two gesture recognition units which are the camera based recognition unit and accelerometers based recognition unit. First, temporal sequence data in each unit are extracted, and the gesture similarities are computed by

AMSS algorithm. And then the similarities calculated form both unit are fused by Choquet integral. The gesture which has the biggest fused similarity are outputted as the final recognized gestures.

To demonstrate the validity, the proposed multimodal gesture recognition method is confirmed by using the gestures from the Mascot Robot System research project with a web cameras and two wearable 3D acceleration sensors. The gesture recognition system achieves the highest recognition rate in the 8 types of gestures. More than 96.0% of gestures are recognized as for all types of gestures, which show the proposed multimodal gesture recognition method advances to the recognition methods based on single sensor. To achieve near 100% accuracy, several improvement is necessary. Firstly, the sensor should be robust to the environment. Sensors like accelerometer and depth camera may be better. Secondly, the recognizer should be powerful enough (e.g., Weighted dynamic time warping [58]). Thirdly, as shown in the experiments, the multimodal system can help to improve the accurate greatly.

In chapter 3, a concept and an inference system of deep level situation understanding are proposed for casual communications among humans and robots/machines. Five parameter is necessary for situation inference. The first parameter is speech content of the interlocutor which may be recognized by speech recognition method (e.g., Julius for Japanese speech recognition). Special knowledge of the interlocutor, e.g., normal emotion state, favorite foods, habits, is considered as the second parameter. People may express their emotion in different ways. To estimate the real emotion of a people, the friend-level knowledge, i.e., customized knowledge of the people, is also necessary. The face features and voice features of normal state is used to train the classifier. The third parameter is the real emotion state of the interlocutor which may be

estimated by the trained classifier. Atmospheres estimated from the emotion changing of the agents is considered as the forth parameter. The fifth parameter is the gestures/postures of the interlocutor which may be recognized by gesture recognition algorithms from sensors like cameras and accelerometers. These fifth parameters, i.e., speech contents, universal and customized knowledge, emotions, atmospheres, and gestures/postures are important parameters for the situation inference system. The meaning of the speech content can be analyzed based on the part-of-speech dictionary and grammar dictionary. The intention of this utterance may be estimated based on the historical dialogs. Thoughtful inference may be done based on the intention of the interlocutor. Finally, the appropriate response in this situation is outputted to the interlocutor.

Twelve subjects are asked by questionnaire to evaluate the naturalness of the response of the proposed inference system comparing to the responses from familiar people. The proposed system achieves a naturalness value of 0.84 which is between the ranks of “natural (=1.0)” and “a little natural (=0.75)” comparing to communicate with familiar people. It is concluded that the proposed deep level situation understanding may help to accomplish human-level natural communication in casual human robot interaction.

Not only surface level understanding (e.g., speech/voice recognition, gesture/posture recognition), emotion understanding, intention understanding, and atmosphere understanding but also customized agent-dependent and universal knowledge, and a thoughtfulness mechanism are considered for smoothing and naturalizing communication among humans and robots/machines. The proposal can be applied to the service robot systems to achieve casual communication when interacting with robots/machines. By considering the customized agent-dependent knowledge in human-

robot communication, it will help robots/machines to understand the usual way in communication and avoid unnecessary troubles and misunderstandings. With the comprehensive consideration of speech/voice, gesture/posture, emotion, intention, atmosphere, and knowledge (including universal and customized knowledge), the proposal will smooth the communication among humans and robots/machines as well as create a peaceful, pleasant, and prosperous society consisting of humans and various specification robots.

4.2 Potential Applications

With the development of hardware technologies (e.g. ARM chips) and software technologies (e.g., Android system [59]), more and more equipment are capable of processing the intelligent information. Understanding the real emotion, intention becomes important for equipment such as service robot, car robot, and domestic robot.

4.2.1 Service Robot System

In the restaurant, regular customers may have special habits such as ordering the same food and drinks, sitting on the same tables. Remembering these kinds of special knowledge of this regular customer is important for service robots to improve the service qualities.

Understanding customer's real emotion state is also important for service robot. Because customers may order different food in different way when they are in different emotion state (e.g., angry or sad). Different people may express their emotions in different ways. Most people usually keep calms. Then keep silent may be the normal state of these

people. When they become smiling, it means they are in the emotion state of happy. While some other peoples may keep smiling all-day. For these people, smiling is their normal state. When they become calm, it may mean they get angry. Understanding the real emotion of customers may avoid communication troubles and smooth communication between the human and robot.

4.2.2 In-Car System

The in-car system has been rapid progressed in recent years. More and more sensors has been embedded in the car system. Embedded system based on Android system are also used as in-car system [60] [61], The Windows Embedded Automotive system [62] enable speech communication and eye tracking via microphone and camera. Some key function of the in-car system is addressed in [63] as follows:

- Handling phones

Driving is a hands-busy activities. The in-car speech recognition system enable voice dialing by name or numbers. The incoming calls can also be announced by telephone number or names. Most of the system also detect the incoming mobile messages and read it for the drivers.

- Media Players

The in-car speech recognition system is able to recognize the title, artist and so on. Besides, playing media from the prepared play list to comfort the drivers.

- Navigation

Global Positioning System (GPS) sensor has been integrated in the car system for many years. Recognizing the destination in voice and guiding the drivers to the right road via the electric map and GPS sensors may facilitate the drivers.

- Inform information

Collecting the information about traffic, stocks, news, and weather has become very easy via internet. Inform information about weather and traffic may insure the safety of drivers and comfort the drivers.

- Business Information

The in-car system may also embedded functions like recording the schedule of meetings, place of the meetings. When the driver forget the meeting, reminding of the meeting is also important.

The in-car system has been more and more multifunction. Drivers may have various needs depending on the emotion state and intentions. Understanding the real emotion and real intention of drivers become necessary for the in-car system. For example if the driver gets angry when driving, playing the favorite music of the driver may calm the driver and insure the safety of the driver. When the driver is going to attend some meeting, checking the weather and traffic information may let the driver feel comfortable. When the driver looks abnormal, talking some interesting things or changing to some interesting topic may heal the driver.

The real emotion of the driver may be estimated based on the customized knowledge of this driver. The real intention may be inference by the historical dialogue data. The thoughtfulness knowledge should be established on ahead. By comprehensively considering the utterance, the real emotion state, real intention and special knowledge of

the driver, the proposed system may inference appreciate response to the driver so that the proposal could smooth communications between driver and in-car system. This will help to achieve human-human level casual communication between driver and in-car system and finally contribute to realize a peaceful, harmonious, and prosperous society for robots and humans.

Bibliography

[1] Y. Yamazaki, H. A. Vu et al., "Mascot Robot System by Integrating Eye Robot and Speech Recognition using RT Middleware and its Casual Information Recommendation", 3rd International Symposium on Computational Intelligence and Industrial Applications (ISCIIA2008), pp. 375-384, (2008).

[2] H. A. Vu, Y. Yamazaki et al., "The Interrupt of Mascot Robot System Embedded in RT Middleware Based on Fuzzy Logic", FACTA UNIVERSITATIS Series: Mechanics, Automatic Control and Robotics, 7-1, pp. 11-28, (2009).

[3] C. Shan, T. Tan et al., "Real-time Hand Tracking Using a Mean Shift Embedded Particle Filter", Journal of the Pattern Recognition Society, pp. 1958-1970, (2007).

[4] E. Huber and D. Kortenkamp, "A Behavior based Approach to Active Stereo Vision for Mobile Robots", Engineering Applications of Artificial Intelligence Journal, pp. 229-243, (1998).

[5] R. Douglas-Cowie, E. Tsapatsoulis et al., "Emotion recognition in human-computer interaction", Signal Process Magazine, IEEE 18-1, pp. 32-80, (2001).

[6] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: state-of-the-art and future perspectives of an emerging domain". In Proceedings of the 16th ACM international conference on Multimedia, pp. 1061-1070, (2008).

[7] Z. Othman, A.R. Yaakub et al., "Virtual Environment Navigation Using an Image based Approach", Student Conference on Research and Development, pp. 364-367, (2002).

[8] C. Keskin and L. Akarun, "STARS: Sign Tracking and Recognition System Using Input-Output HMMs", Patter Recognition Letters 30, 1086-1095, (2009).

[9] H. Kang, C.W. Lee et al., "Recognition based gesture spotting in video games", Pattern Recognition Letters 25, pp. 1701–1714 , (2004).

[10] G. Caridakis, K. Karpouzis et al., "SOMM: Self organizing Markov map for gesture recognition", Pattern Recognition Letters 31, pp. 52–59, (2010).

[11] V. Paquin and P. Cohen, "A Vision based Gestural Guidance Interface for Mobile Robotic Platforms", HCI/ECCV, LNCS 3058, pp. 39–47, (2004).

[12] C. Amma, D. Gehrig et al., "Airwriting Recognition Using Wearable Motion Sensors", Augmented Human Conference, (2010).

[13] Y. Yamazaki, H. A. Vu et al., "Gesture Recognition Using Combination of Acceleration Sensor and Images for Casual Communication between Robots and Humans", IEEE Word Congress on Computational Intelligence, pp. 18-23, (2010).

[14] T. Murofushi and M. Sugeno, "An Interpretation of Fuzzy Measures and the Choquet Integral as an Integral with Respect to a Fuzzy Measure", Fuzzy Sets and Systems

29, pp. 201-227, (1989).

[15] T. Nakamura, K. Taki et al., "AMSS: A Similarity Measure for Time Series Data", IEICE D, 91-D-1, pp. 2579-2588, (2008).

[16] Micro Stone, MVP-RF8, http://www.microstone.co.jp/

[17] M. Sasajima, T. Yano, Kono, Y.. EUROPA, "A generic framework for developing spoken dialogue systems". In Proc. of EUROSPEECH'99, pp. 1163-1166, (1999).

[18] Jason D. Williams, Iker Arizmendi, and Alistair Conkie. "Demonstration of AT&T "Let's Go": A production-grade statistical spoken dialog system". Spoken Language Technology Workshop (SLT), 2010 IEEE, pp. 157-158, (2010).

[19] Asanterabi Malima, Erol Ozgur, and Müjdat Çetin. "A fast algorithm for vision-based hand gesture recognition for robot control." Signal Processing and Communications Applications, 2006 IEEE 14th. IEEE,pp. 1-4, (2006).

[20] X. Zhang, X. Chen, W. H. Wang, J. H. Yang, V. Lantz, and K. Q. Wang, "Hand gesture recognition and virtual game control based on 3D accelerometer and EMG sensors". In Proceedings of the 14th international conference on Intelligent user interfaces, ACM, pp. 401-406, (2009).

[21] Just, Agnès, and Sébastien Marcel. "A comparative study of two state-of-the-art sequence processing techniques for hand gesture recognition." Computer Vision and Image Understanding, 113-4, pp. 532-543, (2009).

[22] Y. Tang, V. Hai, et al.. "Multimodal Gesture Recognition for Mascot Robot System Based on Choquet Integral Using Camera and 3D Accelerometers Fusion", Journal of Advanced Computational Intelligence and Intelligent Informatics. (15), pp. 563-572, (2011).

[23] Kaoru Hirota, and Fangyan Dong. "Concept of Fuzzy Atmosfield and Its Visualization." On Fuzziness. Springer Berlin Heidelberg, pp. 257-263, (2013).

[24] M. H. Ko, G. West et al., "Using dynamic time warping for online temporal fusion in multisensory systems", Information Fusion 9, pp. 370-388, (2008).

[25] Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection", In IJCAI, vol. 14, no. 2, pp. 1137-1145, (1995).

[26] Sugeno, Michio. "Theory of fuzzy integrals and its applications", Tokyo Institute of Technology, (1974).

[27] Imaoka, H. "A proposal of opposite-Sugeno integral and a uniform expression of fuzzy integrals." In Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium, Proceedings of 1995 IEEE International Conference on, vol. 2, pp. 583-590. IEEE, (1995).

[28] Raux, Antoine, Brian Langner, Dan Bohus, Alan W. Black, and Maxine Eskenazi. "Let's go public! taking a spoken dialog system to the real world." In in Proc. of Interspeech 2005, (2005).

[29] Zue, Victor, Stephanie Seneff, James R. Glass, Joseph Polifroni, Christine Pao, Timothy J. Hazen, and Lee Hetherington. "JUPlTER: a telephone-based conversational interface for weather information." Speech and Audio Processing, IEEE Transactions on 8, no. 1, pp. 85-96, (2000).

[30] Walker, Marilyn A., John S. Aberdeen, Julie E. Boland, Elizabeth Owen Bratt, John S. Garofolo, Lynette Hirschman, Audrey N. Le et al. "DARPA communicator dialog travel planning systems: the june 2000 data collection." In INTERSPEECH, pp.

1371-1374, (2001).

[31] Minker, Wolfgang, Udo Haiber, Paul Heisterkamp, and Sven Scheible. "The SENECA spoken language dialogue system." Speech Communication 43-1, pp. 89-102, (2004).

[32] Weng, Fuliang, Sebastian Varges, Badri Raghunathan, Florin Ratiu, Heather Pon-Barry, Brian Lathrop, Qi Zhang et al. "CHAT: a conversational helper for automotive tasks", In INTERSPEECH, (2006).

[33] J. Allen, D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent, . "An architecture for a generic dialogue shell". Natural Language Engineering, 6-3&4, pp. 213-228, (2000).

[34] Larsson, Staffan; Ericsson, Stina. "GoDiS–issue-based dialogue management in a multi-domain", multi-language dialogue system. In: Demonstration Abstracts, ACL-02, (2002).

[35] K. Komatani, N. Kanda, M. Nakano, K. Nakadai, H. Tsujino, T. Ogata, , and H. G. Okuno, "Multi-domain spoken dialogue system with extensibility and robustness against speech recognition errors", In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue. Association for Computational Linguistics, pp. 9-17,(2009).

[36] O. Lemon, A. Gruenstein, A. Battle, and S. Peters, "Multi-tasking and collaborative activities in dialogue systems". In: Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-Volume 2. Association for Computational Linguistics, pp. 113-124,(2002).

[37] Pakucs, Botond, "Towards dynamic multi-domain dialogue processing". In: INTERSPEECH, (2003).

[38] Mctear, Michael F. "Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit". development, 5, 7, (1998).

[39] Lamel, Lori, et al. "The LIMSI ARISE system for train travel information. In:Acoustics", Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on. IEEE, pp. 501-504, (1999).

[40] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies", Speech and Audio Processing, IEEE Transactions, 8(1), pp. 11-23, (2000).

[41] J. D. Williams, P. Poupart, and S. Young, "Partially observable Markov decision processes with continuous observations for dialogue management". In Recent Trends in Discourse and Dialogue, pp. 191-217, (2008).

[42] Lemon, Oliver; Georgila, Kallirroi; Henderson, James. "Evaluating effectiveness and portability of reinforcement learned dialogue strategies with real users: the TALK TownInfo evaluation". In: Spoken Language Technology Workshop. IEEE. IEEE, pp. 178-181,(2006).

[43] Williams, Jason D. "The best of both worlds: unifying conventional dialog systems and POMDPs". In: INTERSPEECH. pp. 1173-1176, (2008).

[44] Pietquin, Olivier; Dutoit, Thierry. "A probabilistic framework for dialog simulation and optimal strategy learning". Audio, Speech, and Language Processing, IEEE Transactions on, 14.2: 589-599, (2006).

[45] Jost Schatzmann, Blaise Thomson, and Steve Young, "Error simulation for training statistical dialogue systems", In: Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on. IEEE, pp. 526-531, (2007).

[46] James Henderson, Oliver Lemon, and Kallirroi Georgila, "Hybrid reinforcement/supervised learning of dialogue policies from fixed data sets". Computational Linguistics, 34.4, pp. 487-511, (2008).

[47] O. Lemon, K. Georgila, , J. Henderson, and M. Stuttle, , "An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the TALK in-car system", In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations . Association for Computational Linguistics, pp. 119-122, (2006).

[48] J. Mäntyjärvi, J. Kela, P. Korpipää, , and S. Kallio, "Enabling fast and effortless customisation in accelerometer based gesture interaction." Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia. ACM, pp.25-31, (2004).

[49] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uWave: Accelerometer-based personalized gesture recognition and its applications", Pervasive and Mobile Computing, 5-6, pp. 657-675, (2009).

[50] S. Hommel, and U. Handmann, "AAM based continuous facial expression recognition for face image sequences", In Computational Intelligence and Informatics (CINTI), 2011 IEEE 12th International Symposium, pp. 189-194,(2011) .

[51] M. Paleari, B. Huet, and R. Chellali, "Towards multimodal emotion recognition: a new approach", In Proceedings of the ACM International Conference on Image and Video Retrieval, pp.174-181,(2010) .

[52] T. Takagi, T.Nishi, and D. Yasuda, "Computer assisted driving support based on intention reasoning", In Industrial Electronics Society, 2000. IECON 2000. 26th

Annual Confjerence of the IEEE, (1), pp.505-508,(2000).

[53] K. Shimada, K. Iwashita, and T. Endo,. "A case study of comparison of several methods for corpus-based speech intention identification", In Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pp.255-262,(2007).

[54] Z. T. Liu, F. Y. Dong, K. Hirota, M. Wu, D. Y. Li, , and Y. Yamazaki,. "Emotional states based 3-D Fuzzy Atmosfield for casual communication between humans and robots", In Fuzzy Systems (FUZZ), 2011 IEEE International Conference, pp. 777-782,(2011) .

[55] L. Chen, Z. Liu, et al.. "Multi-Robot Behavior Adaptation to Communication Atmosphere in Humans-Robots Interaction Using Fuzzy Production Rule Based Friend-Q learning", International Symposium on Soft Computing,(2012).

[56] Lee, Akinobu, and Tatsuya Kawahara. "Recent development of open-source speech recognition engine julius." Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee, (2009).

[57] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto, "Applying Conditional Random Fields to Japanese Morphological Analysis". In: EMNLP. pp. 230-237,(2004).

[58] Jeong, Young-Seon, Myong K. Jeong, and Olufemi A. Omitaomu. "Weighted dynamic time warping for time series classification", Pattern Recognition 44-9, pp. 2231-2240, (2011).

[59] http://www.google.co.jp/mobile/android/

[60] Whipple, John, William Arensman, and Marian Starr Boler. "A public safety application of GPS-enabled smartphones and the android operating system.", Systems, Man and Cybernetics, SMC 2009. IEEE International Conference on. IEEE, (2009).

[61] S. Diewald, A. Möller, L. Roalter, and M. Kranz, (2012, September). "DriveAssist-A V2X-Based Driver Assistance System for Android", In Mensch & Computer Workshopband, pp. 373-380, (2012).

[62] http://www.microsoft.com/windowsembedded/en-us/auto.aspx

[63] Tashev, Ivan, Michael Seltzer, Yun-Cheng Ju, Ye-Yi Wang, and Alex Acero. "Commute UX: Voice enabled in-car infotainment system.", In Mobile HCI, vol. 9. (2009).

Related Publications

Journal Papers

[J1] **Y. Tang**, H. A. Vu, P. Q. Le, D. Masano, O. Thet, C. Fatichah, Z. Liu, M. Yamaguchi, M. L. Tangel, F. Dong, Y. Yamazaki, and K. Hirota, "Multimodal Gesture Recognition for Mascot Robot System Based on Choquet Integral Using Camera and 3D Accelerometers Fusion" , Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.15, No.5, pp. 563-572, 2011

[J2] **Y. Tang**, F. Dong, Y. Yamazaki, T. Shibata and K. Hirota, " Deep Level Situation Understanding for Casual Communication in Humans-Robots Interaction", Journal of Automation, Mobile Robotics & intenlligent Systems (**Submitted**).

International Conference Papers

[C1] **Y. Tang**, F. Dong, M. Yuhki, Y. Yamazaki, T. Shibata and K. Hirota, " Deep Level Situation Understanding and its Application to Casual Communication between Robots and Humans " , 10th Int. Conf. on Informatics in Control, Automation and Robotics (ICINCO2013), Vol.2, pp.292-299, 2013

[C2] K. Hirota, H. Vu, A., P. Q. Le, C. Fatichah, Z. Liu, **Y. Tang**, and Y. Yamazaki, "Multimodal gesture recognition based on Choquet integral". In Fuzzy Systems (FUZZ) IEEE International Conference, pp. 772-776, IEEE , 2011

[C3] Y. Yamazaki, H. A. Vu, P. Q. Le, Z. Liu, C. Fatichah, M. Dai, H. Oikawa, D. Masano, O. Thet, **Y. Tang**, N. Nagashima, M. L. Tangel, F. Dong, and K. Hirota, “Gesture recognition using combination of acceleration sensor and images for casual communication between robots and humans”. In Evolutionary Computation (CEC), 2010 IEEE Congress, pp. 1-7, IEEE, 2010