

# Emotion Recognition based on Human Gesture and Speech Information using RT Middleware

H. A. Vu<sup>1</sup>, Y. Yamazaki<sup>2</sup>, F. Dong<sup>1</sup>, K. Hirota<sup>1</sup>

<sup>1</sup>Dept. of Computational Intelligence & Systems Science,  
Tokyo Institute of Technology  
Yokohama, Japan  
{vuanhai, tou, hirota }@hrt.dis.titech.ac.jp

<sup>2</sup>Dept. of Electrical, Electronic, and Information Eng.  
Kanto Gakuin University  
Yokohama, Japan  
yamazaki@kanto-gakuin.ac.jp

**Abstract**— A bi-modal emotion recognition approach is proposed for recognition of four emotions that integrate information from gestures and speech. The outputs from two uni-modal emotion recognition systems based on affective speech and expressive gesture are fused on a decision level fusion by using weight criterion fusion and best probability plus majority vote fusion methods, and the performance of classifier which performs better than each uni-modal and is helpful in recognizing suitable emotions for communication situations. To validate the proposal, fifty Japanese words (or phrases) and 8 types of gestures that are recorded from five participants are used, and the emotion recognition rate increases up to 85.39%. The proposal is able to extent to using more than other modalities and useful in automatic emotion recognition system for human-robot communication.

**Keywords**- *Emotion Recognition; Decision-level Fusion; Expressive Gesture; Affective Speech*

## I. INTRODUCTION

In the last years, emerging research in human-robot communication is daily life support robot system such as entertainment robots system, service robot system, or household robot system. In such systems, robots communicate with humans friendly by talking to humans and expressing the emotion [1,2] corresponding to the humans' speeches. The casual communication between humans and multiple robots via both of humans' speech and gestures has proposed in mascot robot system [3, 4]. Recognition users' mood or emotional state is main key to determining the corresponding emotion of robots. Most of the work analyzes the human emotion by verbal and non-verbal information. For instance, the emotion recognition using only speech by combining acoustic and semantic analyses [5], and the bi-modal emotion recognition using visual information from expressive face and body gestures [6]. Moreover, the most approach is fusion of both speech information and gesture information [7], also including the expressive faces [8,9].

In this work, the bi-modal emotion recognition is proposed by combining the information from speech, hands and simple body gestures. The motivation of the proposal is improve social communication between humans and robots in assist

robot systems. First, the two uni-modal emotion recognition system for human's speech and gesture are performed separately. Then, at the decision-level fusion, the output of the two uni-modalities are fused by using weight criterion fusion and majority vote fusion methods. The 50 Japanese word or phrases and 8 type of gestures that are using in mascot robot system are used to perform the emotion recognition accuracy.

In II, the overview of mascot robot system is presented. In III, two uni-modal emotion recognitions based on speeches and gestures multimodal are presented; bi-modal emotion recognition on decision-level fusion is proposed. The experiments to confirm the proposal is presented in IV.

## II. MASCOT ROBOT SYSTEM

The mascot robot system (MRS) is an information retrieval and recommendation system for household environments which is able to use casual communication between humans and robots through humans' speeches and gestures. In this system, the robots need to recognize human's speeches and give back the actions corresponding to semantic of requirements. Also, the emotional reactions are appeared to do friendly casual communication. The MRS is developed as a national project with the purpose of developing the Common Basis of Next-Generation Robots led by the New Energy and Industrial Technology Development Organization (NEDO 2005-2008), and now under the support from Japan Society for the Promotion of Science (JSPS 2009-2012). The construction of MRS is shown in Fig.1.

In the home environment such as living room, the fixed type robots are setup on furniture and appliances such as the TV, a dart, information retrieval monitor, drink bar, and a self-propelled type mobile robot. Each robot is controlled by a PC. The common server which controls whole system and the four fixed robots are connected using local area network, LAN cables, while the connection with the mobile robot is achieved via a wireless LAN. The overall management server controls and supervises the whole system, allowing it to coordinate the action of the whole group of robots.

In order to connect the system, the Robot Technology

middleware (RTM) [10] is used to construct the network system which is called RTM-Network as shown in Fig.2. In RTM-Network, each robot can be viewed as a network component and each function unit of every robot is called a robot technology component (RTC).

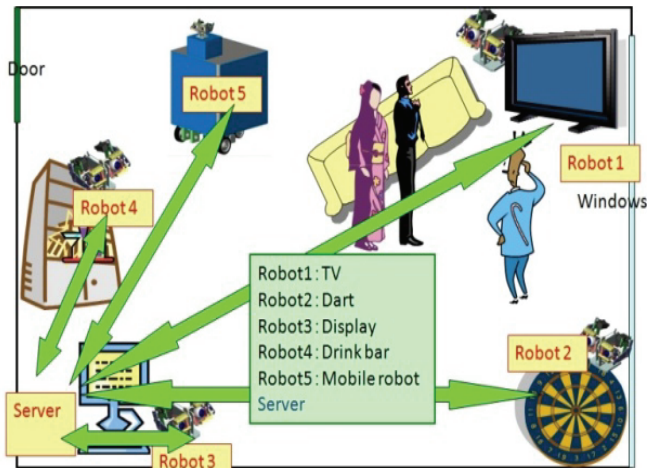


Figure 1. Mascot robot system in a home environment

Every robot is composed of a controlled PC, a microphone,a web camera, a Speech Recognition Module (SRM), a Gesture Recognition Module (GRM), and an Eye robot. The input speeches and gestures that are gotten from the microphone and web camera are processed and then indicated the recognition result on the PC display or written down the information in a log file through the RT Middleware network system. Depending on the recognition results, the eye robot expresses their corresponding emotions through designated actions, and the other actions or talking with people from five robots.

The eye robot on its part has a pair of eyeballs and eyelids. The eyeball movement is based on fuzzy control (up and down, left and right) and serves to express different emotions such as happiness, sadness, surprise, and anger. The eye robots are shown in Fig.2, the appearance of mobile eye robot is shown in Fig.3.

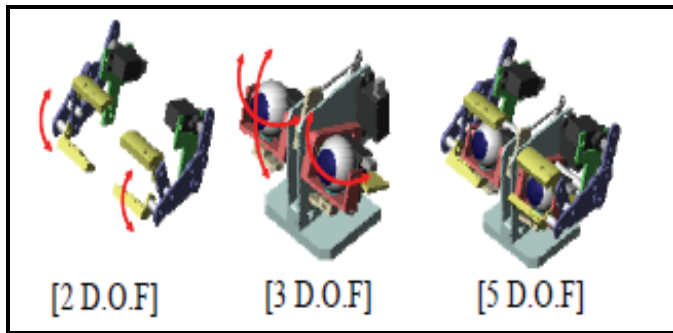


Figure 2. Structure of eye robot



Figure 3. Mobile type eye robot

A host and three guests are considered as users of the MRS. The MRS is ascertained the validity based on a typical situation scenario where the MRS performs representative function of the system. The specification of the scenario is as follows. When the users mention something to the robots that requires an affinity response (for example, "this is interesting", "we won the game", or doing banzai pose, etc...), the system searches for and subsequently provides information based on the user's speech and gestures. Communication between the user and the system is established by the user's speech and gestures to the robots and the robot's response as an emotional expression, as well as displaying information on a screen. For example, the MRS can propose cocktail recipes or provide companionship during a game of darts and reacting accordingly to the performance of the players. The MRS could also take cues from baseball games on TV and offer the user information about the players participating in the game. Futhermore, MRS presents the atmosphere of communications from emotion expresstions of humans and robots.

### III. BI-MODAL EMOTION RECOGNITION

#### A. Uni-modal Emotion recognition

##### 1) Uni-modal Emotion Recognition based on Speeches

To achieve the speakers' emotion based on their speeches, both acoustic analysis and semantic analysis are used in recently research. Here it is assumed that the speakers' expresion of their emotion consists in meaningful talking words or phrases. Therefore, an approach in spotting for emotional information is needed. In order to reduce the complexity and reduce the cost of speech recognition module (SRM), in this work, free software Julius is used to recognize the human's speech under the RT middleware [10]. The RT-Middleware aims at establishing a common platform based on the distributed object technology and which would support the construction of various networked robotic systems by the integration of various network enabled robotic elements called RT-Components. The concept of the Julius based SRM is shown in Fig.4. There are two RT components to recognize humans' speech. The first component- PulseAudioInput is using a microphone to get the speech signal. The second component-pyJulius is using Python language to recognize the input speech signal. The speech based uni-modal emotion recognion is using the data that collect from 4 people using in

home party scenario in mascot robot system [3,4]. There are five emotion extract from the communicaiton in mascot robot system: happiness, sadness, supriseness, and neutral.

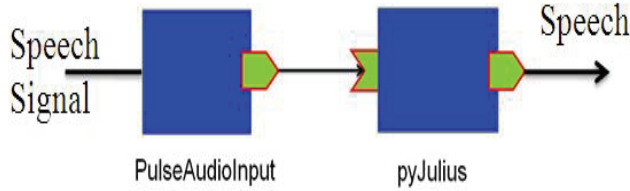


FIGURE 4. RTM based speech recognition module

## 2) Uni-modal Emotion Recognition based on Gestures

Recognition of human emotion from body and hands movements is unresolved in non-verbal communication. In gesture based emotion recognition uni-modal, 8 types of gestures in mascot robot system [3, 4] are used to detect the corresponding emotions. Table 1 provides the list of gesture emotion categories recognized by the system.

The fuzzy Choquet integral [12] based multimodal gesture recognition method is used, where human gestures are recognized based on the fusion of video images and 3D acceleration sensors. The Angular Metrics for Shape Similarity (AMSS) [11] algorithm is employed to evaluate the similarity between gesture templates and input data from the two modules above separately. Next, the optimal fuzzy measures of the Choquet integral are calculated from the training data by a hill climbing algorithm. In the fusion processing, the fused similarities are calculated from similarities of both modules by Choquet integral. Finally, the gesture with maximum fused similarity is chosen as the recognition result. The architecture of the gesture recognition system is shown Fig.5.

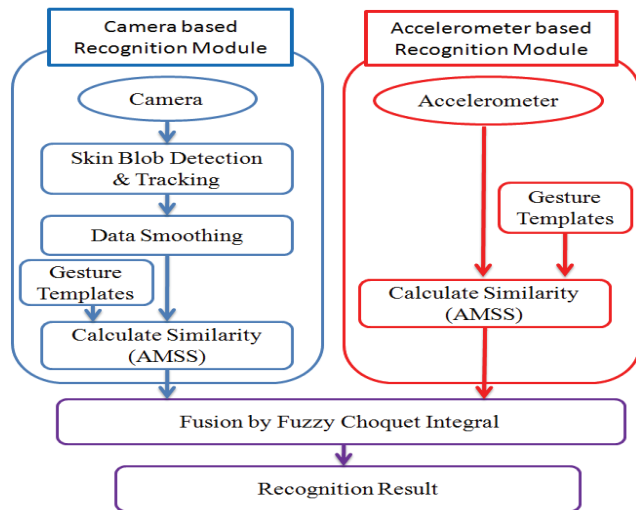


Figure 5. System Architecture of Multimodal Gesture Recognition System

## a) Feature Extraction from Acceleration Data

With each gesture time series data, the maximum and minimum are calculated at first. Then the parameters of mean, variance, and mean crossing rate are calculated by:

$$mean = \sum_{k=1}^N sample_k / N \quad (1)$$

$$variance = \sum_{k=1}^N (sample_k - mean)^2 / N \quad (2)$$

$$mcr = \sum_{k=1}^{N-1} \text{sgn}(sample_k - mean) - \text{sgn}(sample_{k+1} - mean) \quad (3)$$

where  $N$  is the number of total samples of a gesture in the time series data. All of the parameters have be calculated on axis  $x$ ,  $y$ , and  $z$ .

In order to indentify the gesture from the eight types of gestures, the Angular Metrics for Shape Similarity (AMSS) [11] is used for measuring the similarity of input sequence and template sequences. The average similarity of the 3-axes is taken as the similarity of this module.

## b) Image Recognition Module

Gestures are also recognized by video image processing along with acceleration sensor recognition to help make decisions when the system cannot recognize the gestures from acceleration sensor which is the main part of algorithm. Video image sequences are captured while a web camera is put in front of the human within home environment background. The gesture is extracted from each frame in the stream by three main steps: skin-color detection, head and hands tracking and rules matching.

## c) Skin Color Detection and Tracking

The skin detection algorithm is based on the HSV (Hue-Saturation-Value) color space with its robustness to the way that humans tend to perceive colors. Generally, a pixel's color can be considered as skin color if  $h_{min} < H < h_{max}$  and  $s_{min} < S < s_{max}$ , where  $H$ ,  $S$ , and  $V$  are the pixel's color components in the HSV color space and,  $h_{min}$ ,  $h_{max}$ ,  $s_{min}$ ,  $s_{max}$  are the threshold values. From this operation, three regions including the head and two hands can be found as groups of skin color areas.

## B. Bi-modal Emotion Recognition based on Speeches and Gestures

After describing the two uni-modal emotion recognition systems based on semantic information of speech and gestures separately, the effective bi-modal emotion recognition system based on both human's speech and gestures is proposed to do the casual communication between humans and robots. By fusing speech and gesture modalities at the decision level, the combined emotion is classified. The structure of the bi-modal emotion recognition is shown in Fig.6. In the decision level, the output is computed combining the probabilities of the two uni-modal modules. Various researches have been proposed such as sum rule, product rule, median rule, majority vote, etc.

In this work, two different approaches are used to get output emotions by combining the posterior probabilities of the two uni-modal systems. The first approach selects the emotion that received the best probabilities in the two uni-modal system. When each uni-modal system gives the output of different emotion with same probabilities, majority vote is used (in this case, the highest voting for speech based uni-modal system) to get suitable emotions. The second approach selects the emotion using the weight criterion rule. Suppose that,  $\alpha$  is assigned as the weight of speech based uni-modal system,  $\beta$  is assigned as the weight of gesture based uni-modal system.

Then, the output emotion is defined by using the weight criterion rule is shown as

$$\text{Emotion}(i) = \arg \max_{i=1,2,\dots,M} \{ \alpha p(e_i|S) + \beta p(e_i|G) \}, \quad (4)$$

where  $M$  denotes the maximum of possible emotions,  $p(e_i|S)$  and  $p(e_i|G)$  denote probabilities of  $i^{\text{th}}$  emotion that gotten from each speech based emotion recognition system and gesture based emotion recognition system. Cause of speech is still main language in daily communication,  $\alpha$  and  $\beta$  are assigned 0.7 and 0.3 respectively.

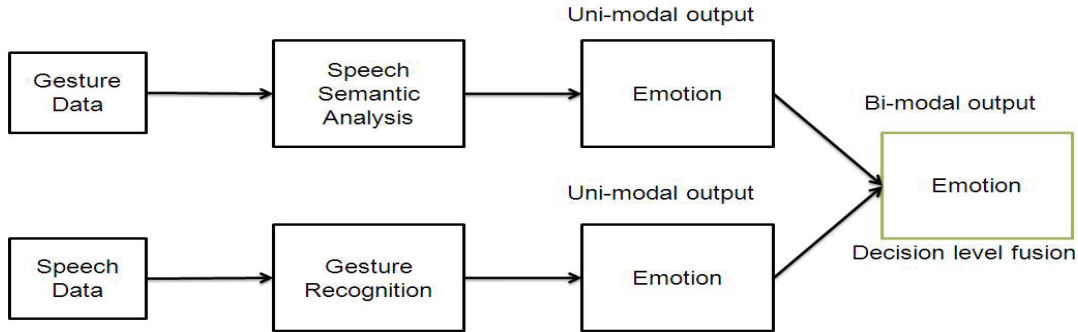


Figure 6. Bi-modal emotion recognition based on speech and gesture

Table 1. 8 types of emotional gestures recognized by the system

#### Happiness

(G1) **Victory pose**: From the relaxed position besides the legs, the right hand is raised up to a level between the chest and the head.

(G2) **Banzai pose**: Both hands are first put down besides legs and then both are held right up.

#### Sadness

(G3) **Face covering**: Both hands are first put down besides the legs and then the right hand is held at face.

(G4) **Bye-bye**: Both hands are first put down besides the legs and then the right hand is waved iteratively at right side.

#### Disappointment

(G5) **Squatting with hands over the head**: From the relaxed position hands besides the legs, both hands are put on the head and human participant squats.

#### Neutral

(G6) **Toast**: Raising a hand with a glass upwards from the waist level.

(G7) **Dart throwing**: Moving a hand with a dart forward speedily from the head level.

(G8) **Guiding**: From their relaxed position, the right hand beside legs is swung up to the point towards the right direction.

## IV. EXPERIMENTS ON BI-MODAL EMOTION RECOGNITION

### A. Data Collection

Five participants (4 male and 1 female) within the age range 22 - 30 took part in the recording. The participants are three different nationalities: Japanese, Chinese, and Vietnamese. Fifty (50) Japanese word and phrases, and 8 type of gestures that are used in home party scenario of mascot robot system are recorded. The sample frames of banzai pose gesture are shown in Fig.7.



Figure 7. Banzai pose gesture

## B. Emotion Recognition Results

### 1) Speech based emotion recognition

Table 2 shows the performance of the uni-modal emotion recognition based on speech. The overall performance of this classifier is 70%. Happiness is recognized with high rate (87.67%). Sadness and Disappointment have lower rates (60.33% and 6.33% respectively).

Table 2. Confusion matrix of emotion recognition based on speeches

|                | Happiness | Sadness | Disappointment | Neutral |
|----------------|-----------|---------|----------------|---------|
| Happiness      | 87.67     | 3.33    | 3.33           | 5.67    |
| Sadness        | 3.67      | 60.33   | 26.33          | 9.67    |
| Disappointment | 0         | 23.33   | 62.33          | 14.34   |
| Neutral        | 7.33      | 13.33   | 11             | 68.34   |

### 2) Gesture based emotion recognition

Table 3 shows the performance of the uni-modal emotion recognition based on humans' gestures. The overall performance of this classifier is 88.5%. Disappointment is classified with the highest rate (96.67). Happiness has the lowest recognition rate of 83.33%.

Table 3. Confusion matrix of emotion recognition based on gestures

|                | Happiness | Sadness | Disappointment | Neutral |
|----------------|-----------|---------|----------------|---------|
| Happiness      | 83.33     | 0       | 0              | 16.67   |
| Sadness        | 1.67      | 85      | 0              | 13.33   |
| Disappointment | 0         | 3.33    | 96.67          | 0       |
| Neutral        | 5.55      | 5.55    | 0              | 88.89   |

### 3) Decision-level fusion

In daily communication, speech is still the main language to express meaningful moods and emotions. Then here the weight criterions are set 0.7 and 0.3 for speeches based uni-modal and gestures based uni-modal respectively. Table 4 shows the performance of the emotion recognition bi-modal based on both speech and gestures. The accuracy of majority vote approach is higher than that of weight criterion.

Table 4. Confusion matrix of emotion recognition based on decision-level fusion

|                | Weight Criterion | Majority Vote |
|----------------|------------------|---------------|
| Happiness      | 86.67            | 86.33         |
| Sadness        | 63.34            | 80.22         |
| Disappointment | 76.63            | 90.67         |
| Neutral        | 78.48            | 84.34         |
| Overall        | 76.28            | 85.39         |

## V. CONCLUSIONS

A bi-modal emotion recognition approach is proposed for recognition of four emotions (happiness, sadness, disappointment, and neutral) that integrates information from humans' speech and gestures. The outputs from two uni-modal emotion recognitions based on speech and gestures are fused on the decision level fusion by using weight criterion fusion and best probability plus majority vote fusion methods. The performance of the classifier was found to be better than each of the uni-modal recognition system and is more helpful in recognizing suitable emotions for communication situations. Firstly, two uni-modal emotion recognition are performed under the 50 Japanese word or phrases and 8 types of gestures that are recorded from five participants. Then, two approaches on the decision-level fusion are performed and obtained 76.28% and 85.39% respectively. The proposal is able to extent to using more than other modalities and useful in automatic emotion recognition system for human-robot communication.

## References

- [1]. H. Gunes, M. Piccardi, "Bi-modal emotion recognition from expressive face and body gesture", Journal of Network and Computer Applications, pp. 1334-134, 2007.
- [2]. P. Bernardis, M. Gentilucci, "Speech and gesture share the same communication system", Neuropsychologia, Vol.44, pp.178-190, 2006.
- [3]. 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), 2008, pp. 375-384.
- [4]. 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, 2010, pp. 18-23.
- [5]. 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, 2009, pp. 11-28.
- [6]. R. Douglas-Cowie, E. Tsapatsoulis et al., "Emotion recognition in human-computer interaction", Signal Process Magazine, IEEE 18-1, 2001, pp. 32-80.
- [7]. Z. Liu, F. Dong et al., "Proposal of Fuzzy Atmosfield for Mood Expression of Human-Robot Communication", International Symposium on Intelligent Systems(iFAN), 2010.
- [8]. A. Suzuki, T. Hoshino, and K. Shigemasu, "Measuring individual differences in sensitivities to basic emotions in faces", Cognition, Vol. 99, Issue 3, 2006, pp. 327-353.
- [9]. A. I. Goldman and C. S. Sripada "Simulationist models of face-based emotion recognition", Cognition, Vol. 94, Issue 3, 2005, pp.193-213.
- [10]. <http://www.openrtm.org/OpenRTM-aist/html/FrontPage.html>
- [11]. T. Nakamura, K. Taki et al., "AMSS: A Similarity Measure for Time Series Data", IEICE D, 91-D-1, 2008, pp. 2579-2588.
- [12]. 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, 1989, pp. 201-227.