

Multimodal Gesture Recognition Based on Choquet Integral

K. Hirota¹⁾, H. A. Vu¹⁾, P. Q. Le¹⁾, C. Fatichah¹⁾, Z. Liu¹⁾, Y. Tang¹⁾, M. L. Tangel¹⁾, Z. Mu¹⁾, B. Sun¹⁾, F. Yan¹⁾, D. Masano¹⁾,
O. Thet¹⁾, M. Yamaguchi¹⁾, F. Dong¹⁾, and Y. Yamazaki²⁾

¹⁾Dept. of Computational Intelligence & Systems Science,
Tokyo Institute of Technology
Yokohama, Japan
{ hirota, vuanhai, phuclq, chastine, liuzhentao, tang, martin, mu, sun,
yan, masano, thetoohan, yamaguchi, tou }@hrt.dis.titech.ac.jp

²⁾Dept. of Electrical, Electronic, and Information Eng.
Kanto Gakuin University
Yokohama, Japan
yamazaki@kanto-gakuin.acwq43.jp

Abstract— A multimodal gesture recognition method is proposed based on Choquet integral by fusing information from camera and 3D accelerometer data. By calculating the optimal fuzzy measures for the camera recognition module and the accelerometer recognition module, the proposal obtains enough recognition rate 92.7% in average for 8 types of gestures by improving the recognition rate approximate 20% compared to that of each module. The proposed method aims to realize the casual communication from humans to robots by integrating nonverbal gesture messages and verbal messages.

Keywords- *Gesture Recognition; Sensor Fusion; Choquet Integral; Human-Robot Interaction; 3D Accelerometer*

I. INTRODUCTION

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, that 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 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 object that has similar color of the skin [11]. On the other hand, application of accelerometer data to 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 image are combined based on fuzzy logic. In the study, when the similarity calculated from the acceleration recognition modules is greater than given threshold value, the recognition result

from the accelerometer module is taken as the final result. In reverse, image recognition module processes the candidate gestures that come from acceleration module to get final result. They, however, have not discussed about how to decide the given threshold. So, it is difficult to use the method for other cases of gestures in general.

To improve the robustness of [13], 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 modules 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 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.

In II, the multimodal gesture recognition system based on camera and 3D accelerometers is proposed. The experiments to confirm the proposal and discussion are presented in III.

II. MULTIMODAL GESTURE RECOGNITION SYSTEM

The architecture of the proposed gesture recognition system is shown Fig. 1.

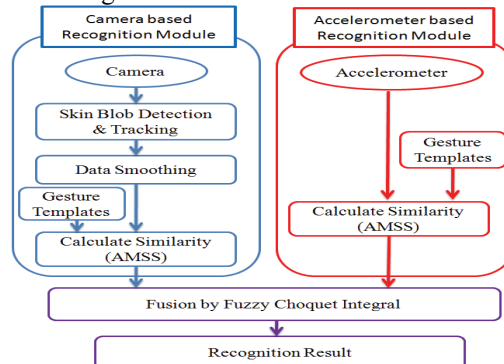


Fig. 1: System Architecture of Multimodal Gesture Recognition System

The system consists of two modules, i.e., camera based gesture recognition module and accelerometers based gesture recognition module.

A. Camera based Gesture Recognition Module

1) Skin Blob Detection and Tracking

The skin blob detection and tracking module 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 $2 \leq H \leq 14$, $65 \leq S \leq 180$, and $20 \leq V \leq 250$ by using OpenCV Library. It is confirmed that the three regions of the head and two hands parts are detected in the skin blob detection processing module.

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.

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:

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

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

where x_i stands for the raw locations of blobs.

b) Quantization: The DTW (Dynamic Time Warping) algorithm mentioned in 3.1.4 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.

3) 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. 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.

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 [16]. After checking from the cost performance viewpoint, minimum selection is used to select the appropriate template for the gesture recognition.

B. Accelerometer based Gesture Recognition Module

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 module. The AMSS 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.

C. Proposed Fusion Method based on Choquet integral

The information obtained from two accelerometers and a camera is different in accuracies and frequencies. This information is 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 modules.

The Choquet integral is proposed by Murofushi and Sugeno[14]. Let $S_{i,g,t}^a$ denotes the similarity between the i-th training data of g-th gesture and the t-th template sequence by the accelerometers based module, and $S_{i,g,t}^c$ the similarity

calculated from the camera based module. M_g^c and M_g^a stand for the fuzzy measure of camera module and the accelerometer module of g-th gesture template separately. Then, the fused similarity calculated by Choquet integral [14] is shown as

$$C_{i,g,t}(S_{i,g,t}^c, S_{i,g,t}^a, M_g^c, M_g^a) = \begin{cases} S_{i,g,t}^a + (S_{i,g,t}^c - S_{i,g,t}^a) \times M_g^c, & \text{if } (S_{i,g,t}^c \geq S_{i,g,t}^a) \\ S_{i,g,t}^c + (S_{i,g,t}^a - S_{i,g,t}^c) \times M_g^a, & \text{if } (S_{i,g,t}^c < S_{i,g,t}^a) \end{cases} \quad (2)$$

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

$$f(i, g, M_1^a, M_1^c, \dots, M_G^a, M_G^c) = \max_{k \in \{1, \dots, T\}} \{C_{i, g, k}\} - C_{i, g, g}, \quad (3)$$

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

Then, the problem of finding the optimal fuzzy measure $(^*M_1^a, ^*M_1^c, \dots, ^*M_G^a, ^*M_G^c)$ from all training data can be modeled as

$$\min_{(M_1^a, M_1^c, \dots, M_G^a, M_G^c)} \sum_{g \in \{1, \dots, G\}} \sum_{i \in \{1, \dots, TN\}} f(i, g, M_1^a, M_1^c, \dots, M_G^a, M_G^c), \quad (4)$$

where TN is the number of training data for each gesture. The optimal fuzzy measure value of both modules 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 $(M_1^a, M_1^c, M_2^a, M_2^c, \dots, M_G^a, M_G^c)$ randomly in the range of between 0 and 1 by uniform distribution.

Step-1: Calculate the fused similarity from the two modules by equation (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 $\{1, 2, \dots, G\}$, similarity error of all training data is calculated by changing the parameters and one step forward and backward as in equation (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 it is less than the current error, and go back to step 1.

III. EXPERIMENT ON MULTIMODAL GESTURE RECOGNITION

A. Experimental Environment

To demonstrate the validity and applicability, the proposed method is confirmed by using the gestures from 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 on eight types of gestures, which are used in the mascot robot system.

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 parts of a body in experiments. The wearable accelerometers (W=45mm,

H=20mm), manufactured by Microstone Inc.[17], are used to get three dimensional acceleration data of human body movements in (x, y, z) 3D space, and the accelerometer transmits these 3D movement data via a Bluetooth connection. Each human participant wears accelerometer on his/her wrists as shown in Fig. 5. The video sequences are captured by a Logitech Qcam(R) Connect camera mounted on a notebook PC with 30 fps rate and 320x240 pixels resolution.

Eight gestures are employed in the home party scenario 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 pose: 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 pose: 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.)

(G5) Squatting with hands over the head: From the relaxed position besides legs, both hands are put on the head and human participant squats. (This expresses disappointment when the dart loses the target.)(Fig. 2)

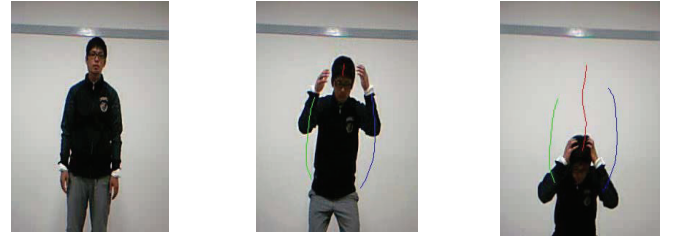


Fig. 2: G5 (Squatting with hands over the head). The blue, red, green lines are the trajectory of the three body parts.

(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.)(Fig. 3)



Fig. 3: G6 (Face covering).

legs and then right hand is held at face. (This expresses sadness when the dart loses the target.)(Fig. 3)

(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.

The eight types of gestures are performed by 6 experimenters for five times separately (240 gestures in total).

B. Training the value of fuzzy measure

The main task in the training phase is to select the proper template and to calculate the optimal fuzzy measure values for similarities fusion.

The dataset includes 240 data which is divided into a training set and a testing set for each gesture with the ratio of 3:2. As mentioned in 3.1.3, the minimum selection is used to choose the suitable template. To find the optimal fuzzy measure values, the following three steps are done;

(1) Similarity calculation: Calculate the similarities of camera and accelerometers modules 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., dividing by the maximum of the eight similarities.

(3) Finding optimal measure value: The algorithm illustrated in 3.3 is implemented for calculating the optimal fuzzy measures for each gesture.

Because the algorithm mentioned in 3.3 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.

In this experiment, the algorithm is applied for 100 times with step value of 0.05, which achieves the best recognition rate of 97.9% by the training data. Table 1 demonstrates the initial and calculated final fuzzy measures of the two modules in the training stage.

Table 1: The initial and final values of fuzzy measure

Gesture	Camera Module		Accelerometers module	
	Initial	Final	Initial	Final
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

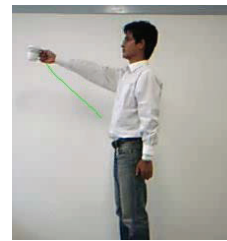
C. Result of Multimodal Gesture Recognition and Discussion

The recognition result of the camera based and the accelerometers based modules are given in Table 2 and Table 3. According to Table 2, G1 (toast gesture) shows the worst result of the camera based recognition module. Because G7 (guide gesture) has the similar trajectory to G1 (Fig. 4), 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

module, while some gestures (i.e. G7) got better recognition result in the accelerometers module.

Table 2: Gesture Recognition Results of camera based recognition module

Gesture	Gesture category								Total
	G1	G2	G3	G4	G5	G6	G7	G8	
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)

Fig. 4: The trajectories (shown in green) of G1 and G7

Table 3: Gesture Recognition Results of accelerometers based recognition module

Gestures	Gesture category								Total
	G1	G2	G3	G4	G5	G6	G7	G8	
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

The recognition rate of training data and test data is shown in Fig. 5. By calculating the optimal fuzzy measure of fuzzy

Choquet integral, the proposed fusion method achieves an average recognition rate of 92.7%, which is 19.4% higher than the average recognition rate of both modules. Seven of all eight gestures achieve higher recognition rate than the one of each modules. For the large difference of the both modules' recognition rate, the recognition result of G7 is affected by the recognition result of camera module so that it gets lower recognition rate than the rate of accelerometers modules. The proposed multi-modal recognition method obtains the average evaluation value of 92.7%, i.e., the highest recognition rate in the averages of all gestures.

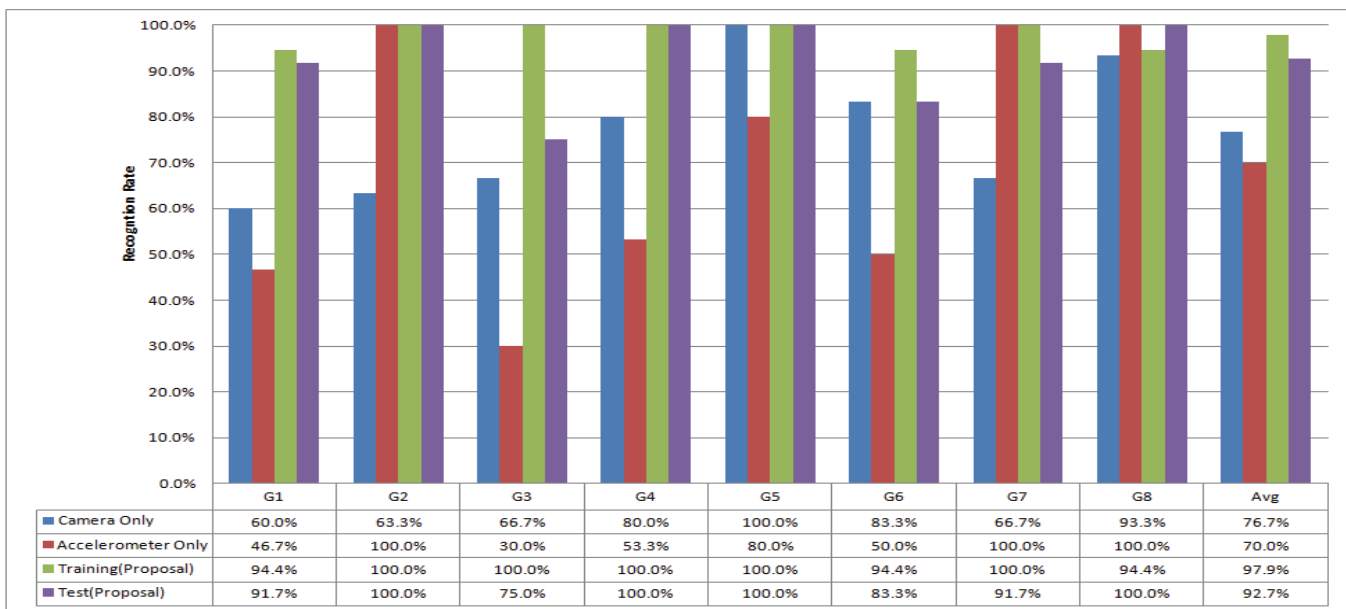


Fig. 5: The recognition rate of training data and test data in proposed multimodal gesture recognition system

This result confirms the practical applicability of the proposed multi-modal recognition method compared with other the recognition methods which are based on a single sensor.

IV. CONCLUSION

A Choquet integral based multimodal gesture recognition method is proposed. The proposal achieves the highest recognition rate in the 8 types of gestures. More than 92.7% 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.

The proposed multimodal gesture recognition provides casual communication abilities for the Mascot Robot System, where robots assist humans in view of the home party atmosphere based on recognitions of verbal and nonverbal messages from humans. Multimodal sensing based Emotion Spaces and Fuzzy Atmosfield [18] will enable robots to help humans according to the atmosphere in all social situations by integrating verbal and nonverbal communication cues from humans and robots, such as speech, gesture, and facial expression.

References

- [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), 2008, pp. 375-384.
- [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, 2009, pp. 11-28.
- [3] C. Shan, T. Tan et al., "Real-time Hand Tracking Using a Mean Shift Embedded Particle Filter", Journal of the Pattern Recognition Society, 2007, pp. 1958-1970.

- [4] E. Huber and D. Kortenkamp, "A Behavior based Approach to Active Stereo Vision for Mobile Robots", Engineering Applications of Artificial Intelligence Journal, 1998, pp. 229-243.
- [5] R. Douglas-Cowie, E. Tsapatsoulis et al., "Emotion recognition in human-computer interaction", Signal Process Magazine, IEEE 18-1, 2001, pp. 32-80.
- [6] A. Vinciarelli, M. Pantic et al., "Social Signal Processing: State-of-the-Art and Future Perspectives of an Emerging Domain", ACM Multimedia, 2008, pp. 1061-1070.
- [7] Z. Othman, A.R. Yaakub et al., "Virtual Environment Navigation Using an Image based Approach", Student Conference on Research and Development, 2002, pp. 364-367.
- [8] C. Keskin and L. Akarun, "STARS: Sign Tracking and Recognition System Using Input-Output HMMs", Patter Recognition Letters 30, 2009, pp. 1086-1095.
- [9] H. Kang, C.W. Lee et al., "Recognition based gesture spotting in video games", Pattern Recognition Letters 25, 2004, pp. 1701-1714 .
- [10] G. Caridakis, K. Karpouzis et al., "SOMM: Self organizing Markov map for gesture recognition", Pattern Recognition Letters 31, 2010, pp. 52-59.
- [11] V. Paquin and P. Cohen, "A Vision based Gestural Guidance Interface for Mobile Robotic Platforms", HCI/ECCV, LNCS 3058, 2004, pp. 39-47.
- [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, 2010, pp. 18-23,.
- [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, 1989, pp. 201-227.
- [15] T. Nakamura, K. Taki et al., "AMSS: A Similarity Measure for Time Series Data", IEICE D, 91-D-1, 2008, pp. 2579-2588.
- [16] M. H. Ko, G. West et al., "Using dynamic time warping for online temporal fusion in multisensory systems", Information Fusion 9, 2008, pp. 370-388
- [17] Micro Stone, MVP-RF8, <http://www.microstone.co.jp/>
- [18] Z. Liu, F. Dong et al., "Proposal of Fuzzy Atmosfield for Mood Expression of Human-Robot Communication", International Symposium on Intelligent Systems(iFAN), 2010