Emotion Estimation Based on Facial Image, Voice Sound and Body Motion

Kosuke KAMIJO*, Ken TOMIYAMA**

** Chiba Institute of Technology, s0826035LJ@it-chiba.ac.jp*
*** Chiba Institute of Technology, tomiyama.ken@it-chiba.ac.jp*

Abstract: The main objective of this paper is to develop an emotion estimator as a part of Virtual KANSEI for robots. We had proposed the concept of Virtual KANSEI for robots as a mechanism to bestow robots with a capability of generating emotion-rich behavior. This study aims at developing an emotion estimation system based on the three raw inputs; facial images, voice sounds and body motions of human partners. We adopted five emotions out of six basic emotions defined by Ekman, namely, joy, anger, sadness, disgust and fear. We also added neutral as the state with no emotion. First, characteristic features from facial images, voice sounds, and body motions are extracted. FACS (Facial Action Coding System) parameters developed by Ekman, MFCC (Mel-Frequency Cepstrum Coefficients) for frequency analysis, and position and acceleration of the head and both hands for human posture detection from KINECT are used for facial features, voice features, and body motions, respectively. Those features are weighted and integrated to determine the partner's emotional state. We adopted Bayesian Network for this integration. The developed system is tested against detection systems with only one or two inputs to prove the favorable effects of integrating features from three inputs.

Keywords: KANSEI Robotics, Emotion Estimation, Virtual KANSEI

1. Introduction

In this study, an emotion estimation system based on the three raw inputs; facial images, voice sounds and body motions of human partners is developed. In recent years, the demand for robots working day-to-day sceneries such as care facilities and private homes has increased considerably due to the decrease of working age population. Those robots will have many opportunities to communicate with human partners and, as such, will have a capability of natural and smooth communication with humans. We believe those robots must have a capability to understand the feelings and emotions of human partners as well as a capability of behaving emotionally. The concept of Virtual KANSEI (VK) was proposed for this requirement. We have been developing modules of VK, including the emotion estimator, for some time. Here, we propose an emotion estimator that combines three basic measurements of facial images, voice sounds and body movements.

2. Virtual KANSEI

Conceptual structure of the Virtual KANSEI is shown in Fig. 1. The virtual KANSEI consists of three parts; (i) KANSEI Detector, (ii) KANSEI Engine and (iii) KANSEI Expressive Regulator. KANSEI Detector estimates the emotional state of the human partner [1, 2, 3, 4]. KANSEI Generator generates the virtual emotion of the robot based on the estimated partner emotion. KANSEI Expressive Regulator modifies motions of robot

actions in such a way that the generated virtual emotion of the robot is reflected in these motions. The robot actions are generated by the robot controller as tasks for the robot.

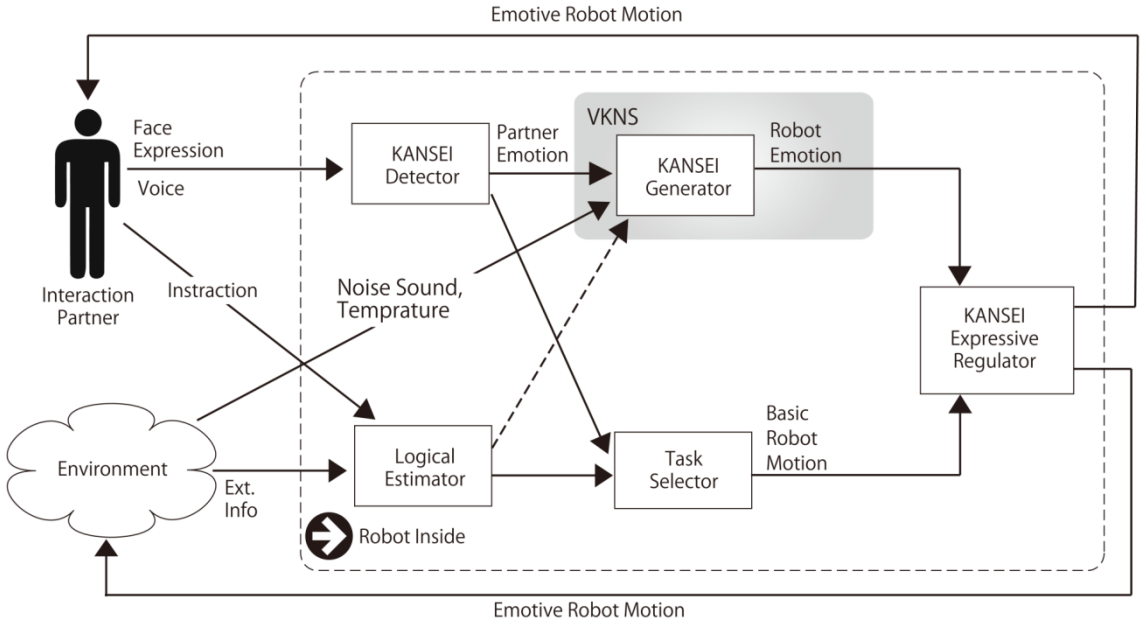

Figure 1. Conceptual structure of the Virtual KANSEI [1, 2, 3, 4]

3. Emotion Estimation System

Our developed system consist of four units; facial image processing unit, voice sound processing unit, body motion processing unit and emotion estimation unit that integrates the outputs from three processing units.

3.1 Facial Image

A collection of characteristic features of facial images, called Facial Action Coding System (FACS), devised by Ekman is adopted here as the representation of attributes of facial images [5]. FACS breaks down the movement in facial expressions anatomically into approximately 60 smallest motion unit called Action Units (AUs) [6, 7]. FACS describes the emotional expression of a face by a combination of AUs. The developed system extracts 21 AUs from the captured image. A USB camera is focused around the eyebrows, eyes and mouth of the human partner to obtain facial images. Those feature parts of the face are automatically recognized by prior learning with pattern matching.

3.2 Voice Sound

Voice sound is processed to obtain Mel-Frequency Cepstrum Coefficients (MFCC) [8, 9]. Voice sound is recorded using a USB monaural microphone as a wave data set (16 bit, 16 kHz). Then, an open source speech recognition engine Julius is used to extract 12 dimensions of MFCC parameters. Please refer to Figure 3 in Section 4.3 for the experimental setup.

3.3 Body Motion

The body motion could include much information such as the positions and accelerations of many body parts. We decided to concentrate on the positions and accelerations of the head and both hands and adopted Kinect for Xbox360 from Microsoft as the measuring device. We used the set of 18-dimensional characteristic features that

represents the position and acceleration information of the head and two hands in the three-dimensional coordinate frame [10].

3.4 Emotion Estimation

We adopted five emotions out of six basic emotions defined by Ekman; joy, anger, sadness, disgust, fear, and excluded the surprise emotion. This is due to the fact that the surprise emotion has a much shorter time constant compared with other emotions and that the surprise emotion tends to occur as a preamble to other emotions. We intend to treat the surprise emotion separately using a specifically designed module. We also added neutral as the state with no emotion.

Features from facial images, voice sounds and body motions are weighted and integrated to determine the partner's emotional state. The outputs of three modules are integrated by the emotion estimation unit. We use a Bayesian network for this purpose [11, 12]. Bayesian network is a graph of dependencies among random variables. Random variables are related each other through conditional probabilities [13]. Three outputs $P_{BN1}(EMOT)$, $P_{BN2}(EMOT)$, and $P_{BN3}(EMOT)$ from three Bayesian networks are the probability distributions of target attributes *EMOT* in those constructed Bayesian networks, where *EMOT* is a random variable that represents six emotions. Here, the Bayesian networks BN1, BN2 and BN3 are for facial images, body motions and voice sound, respectively. The probability distribution of the output of the emotion estimation unit $P_{mix}(EMOT)$ is computed by integrating the outputs of three Bayesian networks based on the integration technique of Kato [14] using mixing ratios as follows:

$$P_{mix}(EMOT) = \sum_{i=1}^{3} \beta_i P_{BNi}(EMOT)$$

Here, the mixing ratios β_i $(\sum_{i=1}^{3} \beta_i = 1, \forall \beta_i \geq 0)$ are adjusted to yield the highest detection result through preliminary experiments.

4. Experiment

In this section, we explain preliminary and verification experiments on the developed system.

4.1 Preliminary Experiment

The purposes of the preliminary experiments are to construct three Bayesian networks for three raw inputs and to adjust the mixing ratio. We collected data sets of facial images, voice sound, and body motions of six university students in 20's (three male and three female students). We collected two samples per one emotion from six subjects. First, we constructed Bayesian networks for three raw inputs. Then, we computed optimal values for the mixing ratios β_i so that the probability of the correct emotion is maximized. Constructed emotion estimation unit including three Bayesian networks is shown in Fig. 2.

4.2 Experimental Procedure of Preliminary Experiment

We instructed the subjects the following: (i) to stand at a specified position from Kinect (see Fig. 3) and (ii) to perform an act with the specified emotion using all of facial image, voice sound and body motion. We asked the subjects to repeat (ii) six times for all emotions.

4.3 Verification Experiment

We compare the accuracy of the developed system with emotion estimation using one or two raw inputs only. A new data set consisting of 12 samples each from the same six subjects as the preliminary experiment is collected. This set is called Same Subject. We also collected another data set consisting of 12 samples each from six university students in their 20s (five men, one woman) who did not participate in the preliminary experiment. This set is called Unknown. This resulted in 24 samples for a single emotion.

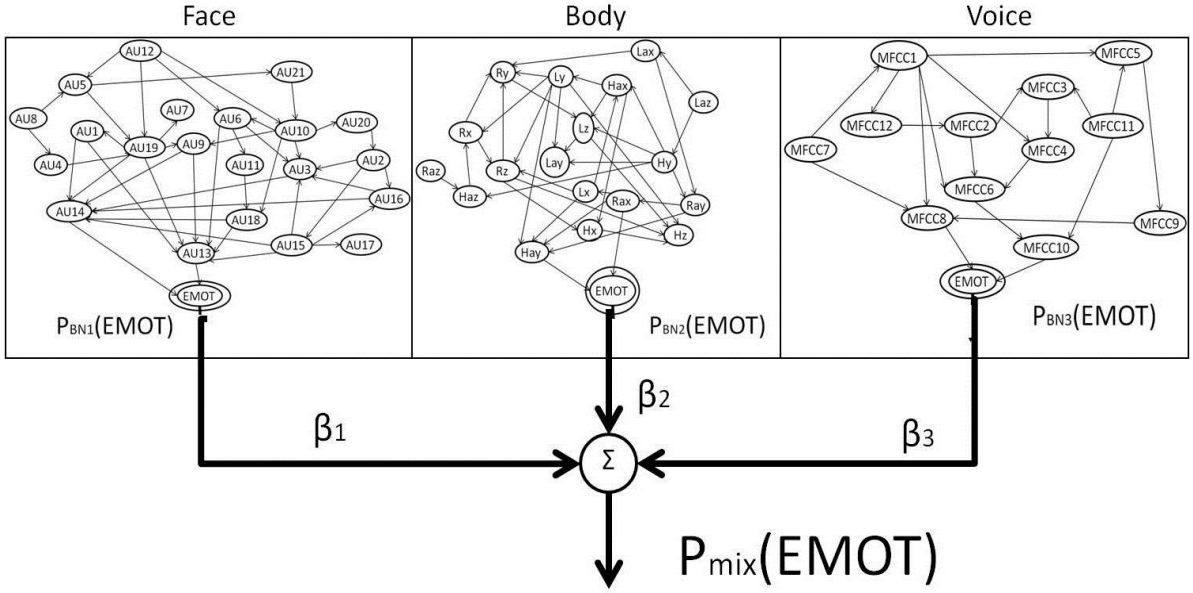

Figure 2. Emotion estimation unit with Bayesian networks

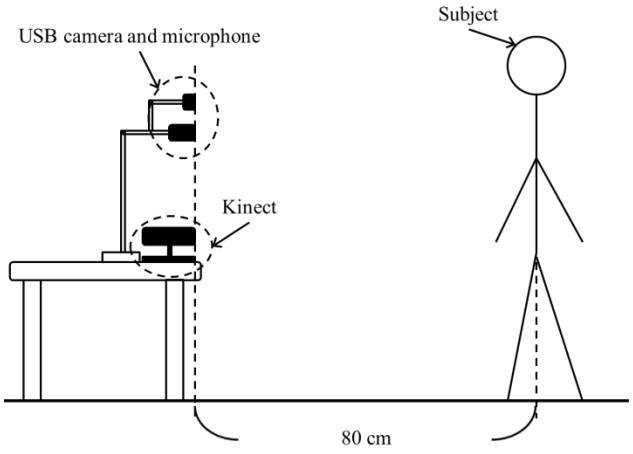

Figure 3. Schematic diagram of both experiments.

4.4 Results of Verification Experiment

The results of the experiment are shown in Tables 1 through 7. Unit of the numbers in the table are percentages. Tables 1 - 3 show the accuracy of estimation using a single raw input only. Tables 4 - 6 show the accuracy of estimation using two raw inputs. Table 7 shows the estimation accuracy of the developed system. In general, estimation accuracy was higher in joy and anger, which are active emotions compared to fear, sadness and disgust. Overall tendency is that more the raw data, higher the accuracy, as we expected. Between the Same

Subject and Unknown sets, the accuracy was understandably lower in the Unknown set. We can see, by comparing Table 7 to Tables 1 - 6, that high overall accuracy was observed except for the fear in both Same Subject and Unknown data sets. It is also noted that the accuracy does not change significantly in Table 7 between the Same Subject set and the Unknown set. This implies that the proposed system can be applied to subjects whose data are not used in developing the system. Thus, the effectiveness of the proposed system was confirmed.

In order to compare the developed system with all other cases, we computed the average of the accuracies in Tables 1 - 6 and listed in Table 8. The comparison of Table 8 with Table 7 shows that the integration of three raw inputs can noticeably increase the estimation accuracy.

Table 1. Emotion estimation from facial image

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	69.2	62.6	76.3	59.8	45.3	71.0	64.0
Unknown	63.5	51.1	61.2	49.7	44.6	68.7	56.5

Table 2. Emotion estimation from voice sound

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	68.0	58.4	69.8	55.6	42.7	51.7	57.7
Unknown	61.7	50.5	61.0	46.3	40.5	59.3	53.2

Table 3. Emotion estimation from body motion

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	64.9	58.4	67.2	32.5	32.9	39.4	49.2
Unknown	61.7	41.2	56.3	31.6	33.8	24.0	40.3

Table 4. Emotion estimation from facial image and voice sound $(\beta_1 = 0.68, \beta_3 = 0.32)$

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	70.3	62.3	72.3	61.1	68.5	72.5	67.8
Unknown	68.2	57.6	69.8	55.1	52.3	69.8	62.1

Table 5. Emotion estimation from voice sound and body motion $(\beta_2 = 0.27, \beta_3 = 0.63)$

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	67.9	47.6	60.2	50.5	40.2	59.1	54.3
Unknown	62.3	48.3	58.7	46.5	49.7	56.4	53.7

Table 6. Emotion estimation from facial image and body motion $(\beta_1 = 0.68, \beta_2 = 0.32)$

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	69.7	50.1	69.5	56.8	39.9	67.8	59.0
Unknown	68.0	42.5	65.8	48.6	31.0	68.2	54.0

Table 7. Emotion estimation from all three inputs, namely, the proposed system $(\beta_1 = 0.48, \beta_2 = 0.20, \beta_3 = 0.32)$

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	78.3	67.1	78.5	69.8	66.2	74.8	72.5
Unknown	74.5	60.9	73.3	65.7	52.4	72.5	66.5

Table 8. Average accuracies of emotion estimation based on one or two raw inputs in Tables 1 - 6

Examinee	Joy	Sadness	Anger	Disgust	Fear	Neutral	Average
Same Subject	68.3	56.6	69.2	52.7	44.9	60.3	58.7
Unknown	64.2	48.5	62.1	46.3	42.0	57.7	53.5

5. Conclusions

In this paper, we proposed the emotion estimation system for Virtual KANSEI that utilizes three characteristic features as facial images, voice sound and body motions. We also proposed to adopt Bayesian network for integrating the characteristic features of three raw inputs. We constructed a system based on our proposal and evaluated the performance of the developed system. First, Bayesian networks for three raw inputs were generated from a set of human data and the set of mixing ratios of outputs from the three networks are optimally determined. Then, two other data sets were used to verify the accuracy of the developed system. Thus, we demonstrated that the integrating three raw inputs is significant in estimating the emotional states of humans.

In the future, we will include surprise in the developed system. We also need to integrate the developed emotion estimator into our Virtual KANSEI.

6. References

[1] Y. Miyaji and K. Tomiyama, (2003) *Construction and Evaluation of a Virtual KANSEI by Petri-net with GA and Method of Constructing Personality*, Proceedings of the 2003 IEEE International Workshop on Robot and Human Interactive Communication, Millbrae, California, USA, pp. 391-396

[2] K. Tomiyama and Y. Miyaji. (2001) *Towards Realization of Care Worker Support Robot* , in Shin Ohara and Isao Kaminaga eds., NIHON NO FUKUSHI [Current Status of Welfare in Japan], Chapter 13, pp. 301-329, IBUNSYA, (ISBN4-7531-0217-3: In Japanese).

[3] Y. Miyaji and K. Tomiyama. (2007) *Virtual KANSEI for Robots in Welfare*, IEEE/ICME International Conference on Complex Medical Engineering, pp. 1323-1326.

[4] J. Kogami, Y. Miyaji and K. Tomiyama. (2008) *Construction and Evaluation of a Virtual KANSEI System for Robots*, KANSAEI Engineering International, Vol. 8, No. 1, pp. 83-90.

[5] P. Ekaman and W. V. Friesen,(1975) *UNMASKING THE FACE*, Prentice-Hall, Inc., Englewood Cliffs.

[6] Y. Ando, R. Yonekura. (1997) *A Study of Emotional Interaction System Using Facial Expression*, The Institute of Electronics Information and Communication Engineers, Technical Report of IEICE MVE96-72, pp. 29-36.

[7] Y. Tian, T. kanade, J. F. Cohn. (2001) *Recognizing Action Units for Facial Expression Analysis*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 23, NO. 2, pp. 97-114.

[8] A. Ando. (2006) *Real-time Speech Recognition*, CORONA PUBLISHING CO., LTD.

[9] N. Wada, N. Hayasaka, Y. Miyanaga, N. Hataoka. (2003) *A Noise Robust Speech Detection System using MFCC analysis*, The Institute of Electronics Information and Communication Engineers, Technical Report of IEICE CAS2003-5, VLD2003-15, DSP2003-35, pp. 25-30.

[10] T. Kagaya, J. Hakura, H. Fujita. (2005) Emotion Extraction Method based on Unconscious Human Gesture, 22th Annual Meeting of Japan Society for Software Science and Technology, pp. 1-4.

[11] Y. Motomura. (2004) *Bayesian networks*, The Institute of Electronics Information and Communication Engineers, Technical Report of IEICE.

[12] K. Tanaka. (2004) *Bayesian Network and Probabilistic Inference Iterated Algorithm for Probabilistic Inference Based on Variational Approach*, The Institute of Electronics Information and Communication Engineers, Technical Report of IEICE NC2004-63, pp. 35-42.

[13] H. Ohnishi. (1997) Interactive Bayesian Structural Modeling: An Integrated Approach for Constructing Causal Structure from Data, The Institute of Electronics Information and Communication Engineers, Technical Report of IEICE ET97-6, pp. 41-48.

[14] Shohei Kato, Yoshiki Sugino and Hidenori Itoh. (2006) *A Bayesian Approach to Emotion Detection in Dialogist's Voice for Human Robot Interaction*, Lecture Notes in Artificial Intelligence (The 10th International Conference on Knowledge-Based Intelligent Information and Engineering Systems (KES 2006)), Vol. 4252, pp. 961-968.