https://doi.org/10.20965/jaciii.2015.p0091

Paper:

Deep Level Emotion Understanding Using Customized Knowledge for Human-Robot Communication

Jesus Adrian Garcia Sanchez*, Kazuhiro Ohnishi*, Atsushi Shibata*, Fangyan Dong, and Kaoru Hirota***

*Department of Computational Intelligence and Systems Science
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
G3-49, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan
E-mail: {garcia, ohnishi, shibata, hirota}@hrt.dis.titech.ac.jp
**Education Academy of Computational Life Sciences (ACLS), Tokyo Institute of Technology
J3-141, 4259 Nagatsuta-cho, Midori-ku, Yokohama 226-8501, Japan
E-mail: tou@hrt.dis.titech.ac.jp
[Received April 30, 2014; accepted September 2, 2014]

In this study, a method for acquiring deep level emotion understanding is proposed to facilitate better human-robot communication, where customized learning knowledge of an observed agent (human or robot) is used with the observed input information from a Kinect sensor device. It aims to obtain agent-dependent emotion understanding by utilizing special customized knowledge of the agent rather than ordinary surface level emotion understanding that uses visual/acoustic/distance information without any customized knowledge. In the experiment employing special demonstration scenarios where a company employee's emotion is understood by a secretary eye robot equipped with a Kinect sensor device, it is confirmed that the proposed method provides deep level emotion understanding that is different from ordinary surface level emotion understanding. The proposal is being planned to be applied to a part of the emotion understanding module in the demonstration experiments of an ongoing robotics research project titled "Multi-Agent Fuzzy Atmosfield."

Keywords: emotion understanding, human-robot communication, multi agent, kinect sensor

1. Introduction

Emotion recognition has been studied in human-robot communication using different types of devices [1]. The devices that are more close to a human way of understanding emotions are those that are based on voice, face, and body gesture information [2, 3]. Different research studies have been carried out to make this realizable; most of them have been based on the facial expressions from the eyes and mouth [4]. Others include the speech [5], gestures [6, 7], or a combination thereof [8, 9], but the missing part from these approaches is the experience [10]. Learning from interactions and creating knowledge is what gives humans the power to deeply understand the emotions of each other. Human emotions are complex, and in many situations, the emotion displayed in the face, voice, or body gesture may not always indicate the real or absolute emotion of the individual [3]. This raises the need to create an algorithm to model this human ability in order to improve human-robot communication [1].

To address and model this problem, a method for acquiring deep level emotion understanding is proposed for human-robot communication, where customized learning knowledge from communication history and a basic knowledge base about the observed agent are utilized with the observed visual/acoustic/depth information input. In this proposal, the voice, facial image, and body gestures are captured using a Kinect sensor device. Each input is fed into a corresponding neural network to obtain a six-dimensional $[0, 1]$ vector representing six basic emotions: anger, disgust, happiness, fear, sadness, and surprise. The three emotion vectors obtained are transformed into fuzzy memberships that are to be combined with the customized knowledge about the observed agent to create the 3D deep emotion vector in the affinity pleasure–arousal space [11]. After the final piece of emotion information is obtained, the knowledge about the observed agent is checked to determine whether a small modification is necessary or not.

By using visual, acoustic, and depth information about the observed agent, obtaining emotion understanding may be possible to some extent, but it is a surface level understanding because facial/voice expression conveys surface level emotion, which may sometimes be different from the real emotion. If, however, the observer agent has enough customized knowledge about the observed agent, then the real emotion may be perceived by taking the situation and the agent customized knowledge into consideration, which is called deep level emotion understanding.

To validate the proposal, two demonstration scenarios are created. The scenarios involve communication in a Japanese company setting where an interaction between a

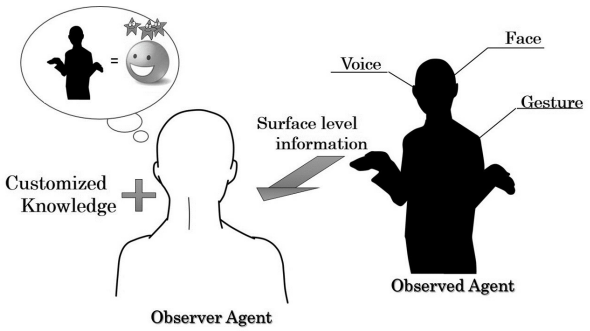


Fig. 1. Two humans meeting for the first time. They use general rules to learn the other's emotion, but they start creating custom knowledge for that specific person.

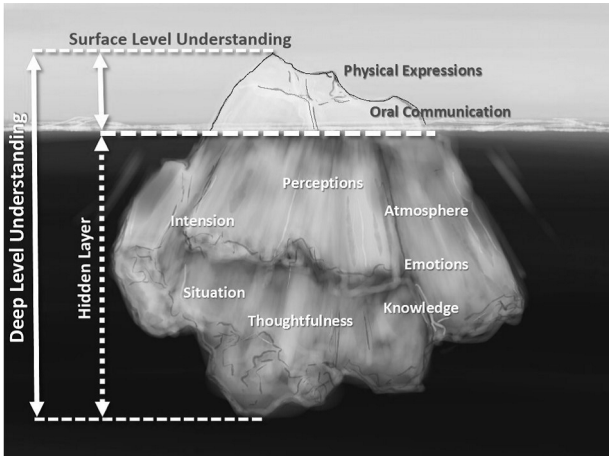


Fig. 2. The iceberg illustrating the complexity and levels of emotion in the human communication.

human employee (observed agent) and a robot secretary (observer agent), who is supposed to have enough customized knowledge about the employee. The employee's face, voice, and body gestures are captured using a Kinect sensor device attached to the robot secretary. The communication topic is a "meeting room reservation" request made by the employee to the secretary and is later modified because of a mistake made earlier by the employee.

The surface level emotion understanding and the possibility of deep level emotion understanding are investigated by utilizing knowledge of the observed agent, which is discussed in Section 2. A method for acquiring deep level emotion understanding is proposed in Section 3. Testing of demonstration scenarios to confirm the validity of the proposal is explained in Section 4.

2. Surface Level Emotion Understanding vs. Deep Level Emotion Understanding

In a multi-agent society consisting of many humans and many robots, studying human-robot communication is essential [2], and understanding of emotions plays an especially important role. Recognizing and understanding the emotions of human beings are easy tasks for human brains but not for the robots [3]. To comprehend the emotion-understanding functions of robots, consider human to human communication from a view point of emotion understanding. Normally, when two people are talking, they understand the emotional state of each other by using two senses, sight and hearing, to recognize the voice, facial expressions, and body gestures as shown in **Fig. 1**. When humans meet for the first time, they use general rules to understand the emotions of each other, thereby creating a first impression of one another. After knowing each other for a long time and becoming more familiar, the observing person starts to deeply understand the observed person's emotions by using the experience and acquired knowledge that, in this case, is called customized knowledge.

Because of these human behaviors, two types of emotion understanding are observed: the first is surface level emotion understanding and the second is deep level emotion understanding. **Fig. 2** makes a graphical visualization of this concept by comparing human behavior to an iceberg. On the surface level, only the physical expressions and oral communication are obtained, but under the water, the larger hidden part of the iceberg contains the ways of thinking and feeling like perceptions, intension, beliefs, knowledge, atmosphere, and emotions of each individual [1].

Available research on emotion recognition has been confined to how to realize and understand human emotions for a long time. Research studies have utilized different types of approaches of how to analyze the facial features [4], the voice [5], and the body gestures [6, 7] described in **Fig. 1**. This information is only for the surface level understanding; it just gives an output as the emotion. Another problem with using the surface level emotions is that the same method and parameters are used for every person who interacts with the system.

This is similar to what people do when meeting for the first time, but this approach is always general and will never improve. On the other hand, people are different, and each person has a different way of expressing themselves. In addition, after some interaction, humans start to understand and know the specific way of how the other person expresses themselves, even predicting what the reaction will be to a situation. This is deep level understanding. That is what the proposed method is trying to simulate: a human way of understanding the emotions and learning and knowing what characterizes a person and their unique way of expressing themselves.

3. Deep Level Emotion Understanding Utilizing Customized Knowledge of the Observed Agent

As described in Section 2, the idea of creating a deep level emotion understanding is based on the utilization of customized knowledge in the interaction between two agents, the observer and the observed. This information

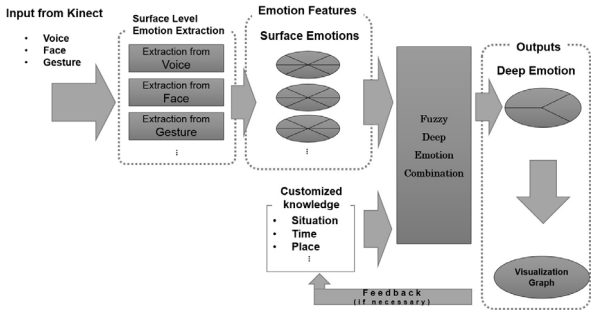


Fig. 3. Deep level emotion understanding method diagram.

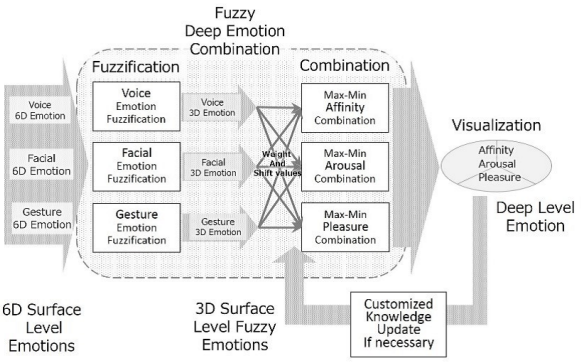


Fig. 4. Fuzzy deep emotion combination algorithm diagram.

is updated and tuned with each interaction until it reaches the level of a "well-known" observed agent, at which technically no changes to the customized knowledge will be needed anymore for that observed agent. The diagram of the proposed three-step method is presented in **Fig. 3**. The first step is the surface information acquisition consisting of three engines: the voice emotion extraction, the face emotion extraction, and the body gesture emotion extraction. The second step is the combination method, where the surface emotions and the customized knowledge are used to calculate the deep emotion. The third step is the visualization and graphics where the final results are recovered.

The steps to obtain the deep level understanding involve obtaining data from the Kinect inputs from the acoustic and visual sources. To extract the voice emotion energy entropy, the short time energy, spectral roll off, spectral centroid, and spectral flux features are used. To extract the face emotion, the lower eyebrow, raiser eyebrow, upper lip, lip corners depressor, lip stretcher, and lower jaw features are used. To extract the body gestures emotion, the head, hand, elbow, and shoulder features are used. Each input feature is fed into the corresponding neural network to obtain a six-dimensional $[0,1]$ vector representing six basic emotions: anger, disgust, happiness, fear, sadness, and surprise [8–10], which represent the surface emotion labels for the surface level emotions. Those values are fuzzyfied and combined with the customized knowledge using MAX-MIN, weight, and shift values. The output is an affinity pleasure–arousal 3D vector and represents the deep level emotion. The deep level emotion is displayed in the affinity pleasure–arousal space [11]. Afterwards, the method starts the feedback process to create/update the customized knowledge if it is necessary. The customized knowledge is a type of profile for a specific observed agent, and it is created at the first interaction.

The customized knowledge is constructed in an XML file with the identification information that consists of the collection of specific data for each different observed agent. The customized knowledge on the observed agent is stored in the observer agent and is based on the observer agent's experience and acquired knowledge. It may consist of the type of person, neutral state, key words, key faces, key gestures, situation, time, place, etc. Using the customized knowledge, the weight and shift values are calculated and implemented in the fuzzy deep combination algorithm. The shift values are determined by the neutral emotion state and are expressed by a triplet (sv_1, sv_2, sv_3) in an emotion space $[-1,1]^3$. The shift values are given by the average of the emotion outputs of the system of the observed agent, where the neutral emotion state is obtained after carrying out multiple measurements of the emotion of the observed agent in a neutral state (normal behavior). The average values are saved in the customized knowledge. For example, if a person had a high tone of voice, the voice emotion will show high arousal and less affinity. The weight values (w_1, w_2, w_3), where $w_1 + w_2 + w_3 = 1$, the adaptation of the algorithm to the specific agent's manner of emotion communication. A human usually uses one main channel (voice, face, or body gestures) to show their emotions. The channel used depends generally on the ethnicity or regions [12, 13]. For example, Asian/Japanese people tend to use their voices more to communicate their emotions, while European/Italian people use body gestures more. The weight values are given by the designer step by step based on the profile of the observed agent. For example, in the experiment of the Japanese observed agent, more priority is assumed and assigned to voice. **Fig. 4** shows the fuzzy deep emotion combination function where the deep knowledge is being used.

All emotions are represented in the affinity pleasure–arousal space as

$$E = \left(e_{affinity}, e_{pleasure}, e_{arousal}\right) \quad \forall e_{affinity}, e_{pleasure}, e_{arousal} \in [-1,1]^3, \quad . \quad (1)$$

where E is the emotion state, and $e_{affinity}$, $e_{pleasure}$, and $e_{arousal}$ are the attributes for "Affinity–No-affinity," "Pleasure–Displeasure," and "Arousal–Sleep" axes, respectively. Each surface level emotion is transformed from one of the six basic emotions to a point in the affinity pleasure–arousal space expressed by Eq. (1). In the case of the surface emotions, the *SEL* (surface emotion labels) from each three-layered feed forward neural network, i.e., 6 inputs and 6 outputs for the face, 6 inputs and 6 outputs for the voice, and 12 inputs and 6 outputs for the body

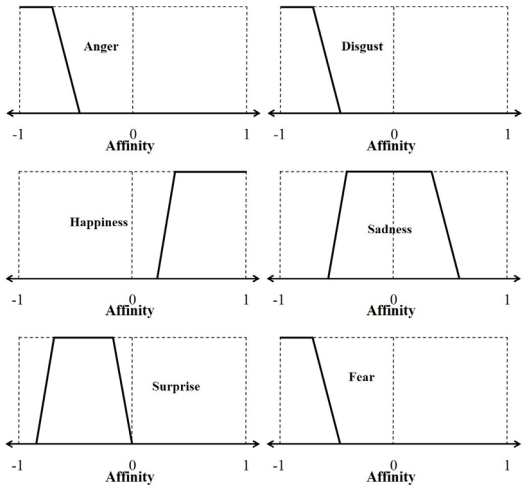


Fig. 5. Affinity axis membership functions for the surface emotion labels.

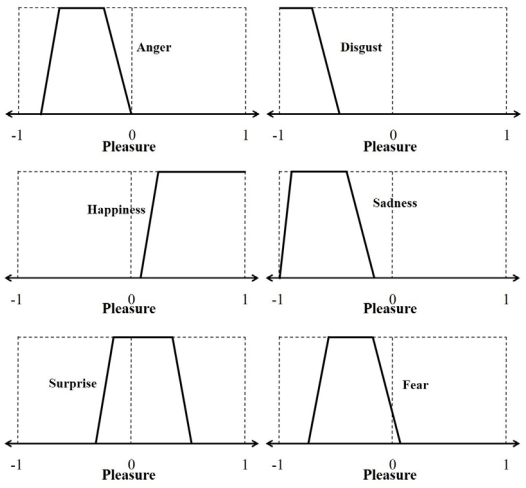


Fig. 6. Pleasure axis membership functions for the surface emotion labels.

gesture, are expressed by 6 binary vectors as

$$SEL = \{\text{angry}, \text{disgust}, \text{happyness}, \text{fear}, \text{sadness}, \text{surprise}\} \in \{0,1\}^6, \quad \ldots\ldots\ldots\ldots (2)$$

where the elements of the *SEL* are binary emotion labels (anger, disgust, happiness, fear, sadness, and surprise). In the proposed method, there are three SEL labels from the face, voice, and gesture. In the fuzzification process, each *SEL* label is converted to *SE* (surface emotion) in the 3D space of affinity pleasure–arousal. This conversion is carried out using fuzzy sets to represent each emotion based on a cross-cultural circumplex [14, 15] by the membership functions $\mu(*)$ as

$$SE = \mu(SEL) = \left(se_{af}, se_{pl}, se_{ar}\right) \in [-1,1]^3, \quad \ldots\ldots (3)$$

where se_{af}, se_{pl}, and se_{ar} are the surface emotion affinity, surface emotion pleasure, and surface emotion arousal, respectively. The membership functions as shown in **Figs. 5–7** are used to calculate the surface emotion values.

In the fuzzy deep emotion combination process, each value of the *DE* (deep emotion) vector is calculated with the *SE* from each device *n* as

$$DE_{af} = \frac{\int af \max_{n\in N}\left(w_n se_{afn}\right) daf}{\int \max_{n\in N}\left(w_n se_{afn}\right) daf} + sv_{af}, \quad \ldots (4)$$

where *af* is a variable of the affinity axis, n is one device out of all the devices *N*, w_n is the weight value for the specific input based on the customized knowledge, se_{afn} is the membership in the affinity axis, and sv_{af} is the shift value for the affinity axis;

$$DE_{pl} = \frac{\int pl \max_{n\in N}\left(w_n se_{pln}\right) dpl}{\int \max_{n\in N}\left(w_n se_{pln}\right) dpl} + sv_{pl}, \quad \ldots (5)$$

where *pl* is the variable of the pleasure axis, n is one device out of all the devices N, w_n is the weight value for the specific input, se_{pln} is the membership in the pleasure axis, and sv_{pl} is the shift value for the pleasure axis; and

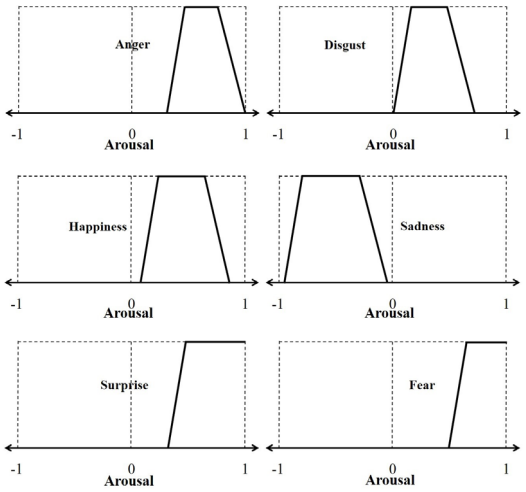


Fig. 7. Arousal axis membership functions for the surface emotion labels.

$$DE_{ar} = \frac{\int ar \max_{n\in N}\left(w_n se_{arn}\right) dar}{\int \max_{n\in N}\left(w_n se_{arn}\right) dar} + sv_{ar}, \quad \ldots (6)$$

where *ar* is the variable of the arousal axis, n is one device out of all the devices N, w_n is the weight value for the specific input, se_{arn} is the membership in the arousal axis, and sv_{ar} is the shift value for the arousal axis.

In Eqs. (4)–(6), the components from the surface emotion from each neural network are multiplied by the weight value (w_n) depending on the customized knowledge, the *max* value of components is detected, each centroid is calculated, and, finally, the corresponding shift value is applied based on customized knowledge.

The resulting values in Eqs. (4)–(6) are the components of the deep emotion affinity pleasure–arousal space axes.

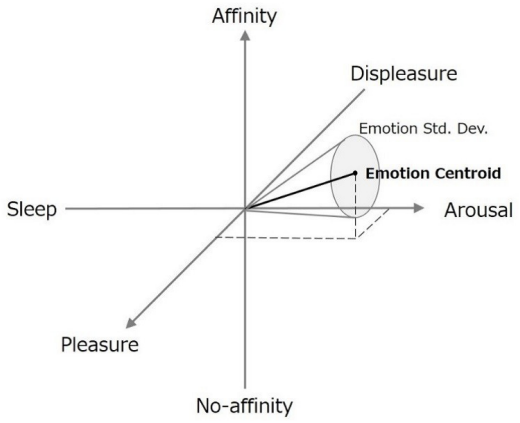


Fig. 8. Representation of deep emotion output where the distorted cone and the centroid of the emotion are shown.

Fig. 9. Scenario consisting of a secretary robot (observer agent) and human employee (observed agent).

Also, the standard deviation between the surface emotions is calculated to create a distorted cone to show where the deep emotion is located. The emotion centroid (deep emotion) and standard deviation are plotted in the affinity pleasure–arousal space, as shown in **Fig. 8**.

4. Demonstration Scenarios for Human-Robot Communication

Two demonstration scenes are created to validate the proposed method. Therein, communication between a human employee (observed agent) and a robot secretary (emotion observer) in a Japanese company setting is simulated as shown in **Fig. 9**.

The employee's face, voice, and body gesture are captured by a Kinect sensor device attached to the robot secretary. The topic is a "meeting room reservation," where a reservation made by the employee is subsequently modified because of a mistake made in the earlier reservation. A script is created and recorded in Japanese language, and the translated version of the script in English is presented in **Table 1**.

In this scenario, it is assumed that the secretary robot and the employee have already known each other for a long time, meaning that the customized knowledge about the employee has already been checked, adjusted, and updated. In Scene 1, the observer agent senses the normal behavior of the employee and understands the importance of the meeting and assigns the necessary priority. In Scene 2, the secretary robot is able to deeply understand the emotions of the employee and then uses more adequate words to try to calm the observed agent.

To capture the inputs for the proposed method, a Kinect motion sensor device is used. The Kinect sensor has different libraries that perform the detection and extraction with enough precision/speed [16]. To capture and record the audio, the Audio library is used, for the facial features, the Face Tracking library is used, and the Skeletal Tracking library is used to obtain the information of the head and the upper extremities [17]. **Fig. 10** shows the output image of the Kinect sensor of the Skeleton and Face Tracking in the center.

Each of the three emotion extraction engines is cre-

Table 1. Script of the simulated interaction titled "a routine of beloved employee," where a meeting room reservation conversation between an employee and the secretary (robot) took place.

Scene 1:
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 are welcome. Good luck.*
Change of Scene:
The employee later realizes his mistake after a conversation with his manager. The employee goes back to the secretary room to change the date of the earlier reservation.
Scene 2:
Employee: *Are you busy now?* ***Secretary:*** *Go ahead!* ***Employee:*** *Excuse me, I want to change the meeting with the branch office to next Thursday.* ***Secretary:*** *Sure! I will check it now.* ***Employee:*** *Yes, please.* ***Secretary:*** *For next Thursday, 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.*

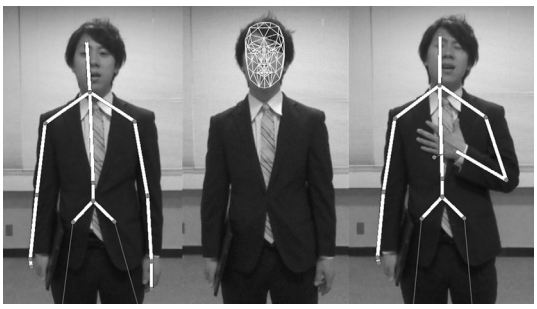


Fig. 10. Output from Kinect sensor device to be analyzed.

Fig. 11. Some of the participants of the Kinect emotion dataset created for the project.

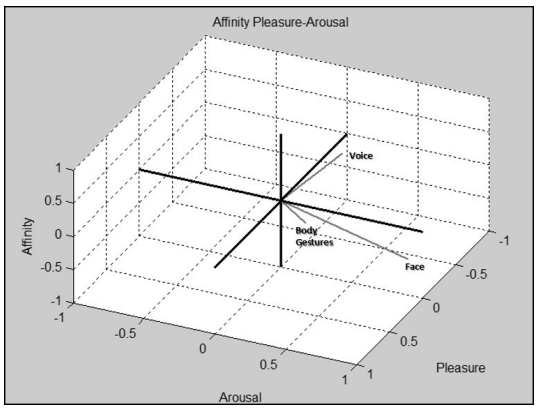


Fig. 12. Affinity pleasure–arousal space showing the surface outputs.

ated and trained separately; also, the programming experiments are divided into two parts. The coding to process the video and depth information is generated in Visual Studio 2012 using C# language, and the audio input is processed using the Audio Analysis Tool of MATLAB.

To train the three neural networks, two different datasets are used. One dataset containing the depth and video information (created originally for this project) consists of 10 people showing the six different basic emotions in front of the Kinect sensor twice. Samples of the records created using Kinect Studio v1.7.0 to make a total of 120 inputs are shown in **Fig. 11**. The second dataset used for the voice train is the Berlin Database of Emotion Speech [18], which consists of 10 different people showing six basic emotions plus a neutral voice multiple times, making a total of 535 audio voice files.

Three different feed forward neural networks are used to obtain the emotion for the face, body gesture and voice. The configuration for the face neural network consists of 6 nodes as inputs, 18 nodes in the hidden layer, and 6 nodes as outputs, which achieves 75.5% accuracy. Using the Kinect sensor makes it possible to recognize and process the skeleton of the people using a support vector machine, thus detecting the position of the human body extremities and head. The body skeleton values outputs are directly input for the gesture neural network. The gesture neural network consists of 12 inputs nodes, 22 nodes in the hidden layer, and 6 nodes in the outputs layer with an accuracy of 85.0%. The emotion from the voice is processed in the voice neural network, a configuration that consisted of 6 nodes as inputs, 20 nodes in the hidden layer, and 6 nodes as the outputs, achieving 76.9% accuracy.

After the surface emotion is calculated, the three 6D emotion vectors are transformed to membership functions for each of the axis values (affinity, pleasure, and arousal) and combined based on the properties of the observed agent and the situation information. This combination is formed firstly by using the type of person who gives more weight to the stronger channel of emotion communication; the observed situation information is then taken into consideration. These values are combined in a fuzzy-based algorithm with the surface emotion vectors to obtain the deep level emotion 3D vector. The deep emotion vector is displayed in the affinity pleasure–arousal space [11].

The surface emotion outputs from the voice, face, and body gesture are different, and **Fig. 12** shows an example of them after transformation along the affinity pleasure–arousal space axes.

The resultant emotions in the experiment still do not closely match the results described in a survey administered to Japanese people based on similar scenarios about observed agent emotion. However, the results give a good approximation of the emotion described by the observed agent. The results obtained by the system are better than those given by people who are not familiar with the observed agent. An example of the resulting deep emotion is shown in **Fig. 13**, where all the surface emotions and customized knowledge are combined. **Figs. 14–16** show the comparison of the centroid of the deep level emotion (dot line), centroid of the surface level emotion (double line), and the emotion reported from the observed agent (solid line). The emotion reported from the observed agent is closer to the deep emotion centroid than to the surface emotions.

The mean square errors are calculated for the three different axes to compare the deep level emotion with sur-

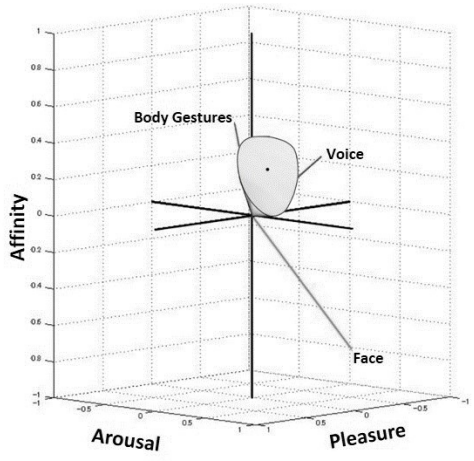


Fig. 13. Affinity pleasure–arousal space showing the deep level emotion centroid and the standard deviation cone.

face level emotion. The values (0.0805, 0.086, 0.0396) are the mean square error values for the deep level emotion, and (0.111, 0.257, 0.056) are for the surface emotion. The error for the deep level emotion is smaller than that for the surface level emotion.

The experimental environment consists of a computer with a 32-bit Intel® Core ™i7-2600 CPU 3.40 GHz processor, 4 GB RAM, and a Kinect sensor device for Windows model 1517. For the software requirements, a computing device equipped with Microsoft Operating system Windows 7 Enterprise, a MATLAB environment with Neural Network Toolbox and Audio Analysis Library installation, Microsoft Visual Studio Express 2012 for Windows Desktop Version, and 11.0.60315.01 Update 2 for coding for the Kinect sensor were used.

5. Conclusions

A deep level emotion understanding method is proposed for human-robot communication to handle the complex and individual way that every human expresses their emotions. In the experiment, the inputs are captured by a Kinect sensor and processed in three neural networks to obtain the surface level emotions. Then the surface emotions and customized knowledge are combined in a "fuzzy deep emotion" algorithm to obtain the deep level emotion.

In the experiment, the customized knowledge is introduced as the feature necessary to achieve a deep level emotion understanding for realizing smooth communication between an observed agent (human employee) and an observer agent (robot secretary). In the demonstration scenario, there exist two interactions, i.e., in the first interaction, using the customized knowledge about the employee and the employee's situation information, the secretary (robot) makes a meeting room reservation for the human employee based on an understanding of the employee's emotions in a normal state. Subsequently, in the second interaction, at the request of the employee, the secretary (robot) changes the earlier meeting room reservation schedule by understanding the employee's mistake and giving sympathized thoughtfulness to the employee. The secretary (robot) understands the change in the emotional state of the employee between the first and second interactions by applying the knowledge about the employee to the observed surface level information from the Kinect sensor device. This confirms that the proposed method can assist in deriving the deep emotion from the surface emotion in combination with the customized knowledge. The mean square errors of (0.0805, 0.086, 0.0396) show a closer relation between the real reported emotion and the deep level emotion.

The proposed method is being planned to be applied

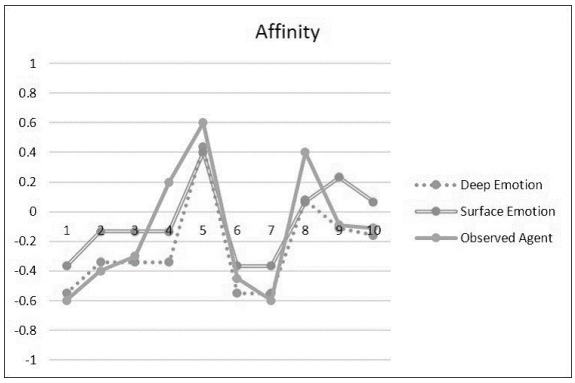


Fig. 14. Comparison between the deep emotion, surface emotion, and the observed reported emotion in affinity axis.

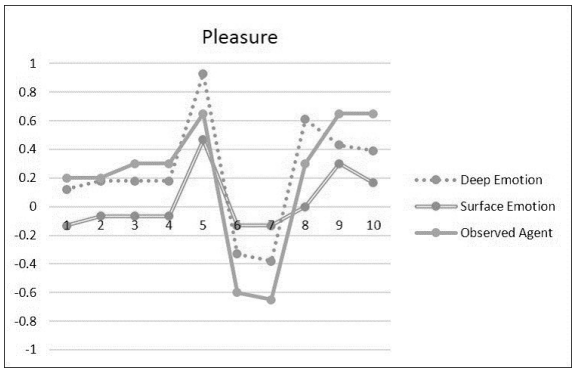


Fig. 15. Comparison between the deep emotion, surface emotion, and the observed reported emotion in pleasure axis.

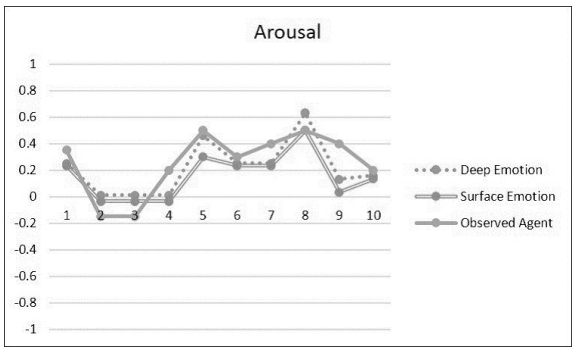


Fig. 16. Comparison between the deep emotion, surface emotion, and the observed reported emotion in arousal axis.

to an emotion understanding module in the demonstration experiments of authors' ongoing robotics research project titled "Multi-Agent Fuzzy Atmosfield" [19]. For the project, the script titled "a routine of beloved employee" that will be performed by five humans (employee, manager, colleague A, bar new guest, and employee's wife) and five robots (secretary, colleague B, bar lady, PARO, and a kid) is being pursued. It is important to mention the perceived necessity of systems that can learn from normal interactions with humans and improve their reaction based on that acquired knowledge to achieve the maximum adaption and personalization.

In future direction, the need to 1) add more features to the customized knowledge and 2) improve an identification system to implement the right customized knowledge to the right person will require attention. Similarly, utilizing pattern recognition to create a new input for the system to handle the quick and fast gestures during conversations could improve emotion understanding in human-computer communication.

Acknowledgements

This study was supported by the Japanese Government via the Monbukagakusho Scholarship Program (25540107). The authors would like to thank Takahiro Kawabuchi for his participation in the demonstration scenarios described in this study. Similarly, the support of M. L. Tangel, J. Pomares B., H. Chang, B. C. Sprague, and A. M. Iliyasu during the revision of the manuscript and all of the participants who helped in the Kinect emotion dataset creation is immensely appreciated.

References:

[1] L. Cañamero, "Emotion understanding from the perspective of autonomous robots research," Neural Networks, Emotion and Brain, Vol.18, Issue 4, pp. 445-455, 2005.

[2] C. L. Bethel, "Survery of Psychophysiology Measurements Applied to Human-Robot Interaction," The 16th IEEE Int. Symposium on Robot and Human interactive Communication, pp. 732-737, 2007.

[3] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions is No.Two-Dimensional," Psychological Science, Vol.18, No.12, pp. 1050-1057, 2007.

[4] M. Ilbeygia and H. Shah-Hosseinib, "A novel fuzzy facial expression recognition system based on facial feature extraction from color face images," Engineering Applications of Artificial Intelligence, Vol.25, Issue 1, pp. 30-146, 2012.

[5] B. I. Ashish and D. S. Chaudhari, "Speech Emotion Recognition," Int. Journal of Soft Computing and Engineering (IJSCE), Vol.2, Issue 1, pp. 235-238, 2012.

[6] Y. Zhao, X. Wang, M. Goubran, T. Whalen, and E. M. Petriu, "Human emotion and cognition recognition from body language of the head using soft computing techniques," Journal of Ambient Intelligence and Humanized Computing, Vol.4, Issue 1, pp. 121-140, 2013.

[7] L. Miranda, T. Vieira, D. Martinez, T. Lewiner, and A. W. Vieira, F. M. Campos, "Real-time gesture recognition from depth data through key poses learning and decision forests," Brazilian Symposium of Computer Graphic and Image Processing, 25th SIBGRAPI: Conf. on Graphics, Patterns and Images, pp. 268-275, 2012.

[8] G. Castellano, L. Kessous, and G. Caridakis, "Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech," Affect and Emotion in Human-Computer Interaction, Lecture Notes in Computer Science, Vol.4868, pp. 92-103, Springer Berlin Heidelberg, 2008.

[9] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," J. on Multimodal User Interfaces, Vol.3, Issue 1-2, pp. 33-48, 2010.

[10] M. Kazemifard, N. Ghasem-Aghaee, adn B. L. Koenig, T. I. Ören, "An emotion understanding framework for intelligent agents based on episodic and semantic memories," Autonomous Agents and Multi-Agent Systems, Springer US, 2013.

[11] Y. Yamazaki, Y. Hatakeyama, F. Dong, K. Nomoto, and K. Hirota, "Fuzzy Inference based Mentality Expression for Eye Robot in Affinity Pleasure-Arousal Space," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.12, No.3, pp. 304-313, 2008.

[12] D. Matsumoto, "Cultural Similities and Differences in Display Rules," Motivation and Emotion, Vol.14, No.3, pp. 195-214, 1990.

[13] D. Matsumoto, "Culture and Emotional Expression. Understanding Culture: Theory, Research, and Application," Psychology Press, pp. 263-279, 2009.

[14] J. Russel, "A Circumplex model of affect," J. of Psychology and Social Psychology, Vol.39, No.6, pp. 1161-1178, 1980.

[15] J. Russel, T. Niit, and M. Lewicka, "A Cross-Cultural Study of a Circumplex Model of Affect," J. of Personality and Social Psycology, Vol.57, No.5, pp. 848-856, 1989.

[16] M. A. Livingston, J. Sebastian, Z. Ai, and J. Decker, "Performance Measurements for the Microsoft Kinect Skeleton," Conf. Proc., Virtual Reality Short Papers and Posters IEEE, pp. 119-120, 2012.

[17] Kinect libraries used in the coding: Audio library http://msdn.microsoft.com/en-us/us-en/library/jj131025.aspx, Face Tracking library http://msdn.microsoft.com/en-us/library/jj130970.aspx, Skeletal Tracking library http://msdn.microsoft.com/en-us/us-en/library/jj131025.aspx [Accessed April, 2013].

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," Proc. Interspeech Lissabon, Portugal, pp. 1517-1520, 2005.

[19] Z. Liu, M. Wu, D. Li, L. Chen, F. Dong, Y. Yamazaki, and K. Hirota, "Concept of Fuzzy Atmosfield for Representing Communication Atmosphere and its Application to Humans-Robots Interaction," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.17, No.1, pp. 3-17, 2013.

Name:
Jesus Adrian Garcia Sanchez

Affiliation:
Doctor degree candidate, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology

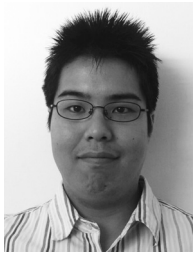
Address:
G3-49, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

Brief Biographical History:

2000- Computer Engineer, Universidad Autonoma de Baja California, Mexico
2006- Senior Programmer, Panasonic AVC Network
2010- Research Student in Computational Intelligence and Intelligent Systems Science, Tokyo Institute of Technology
2011- Master of Engineering in Computational Intelligence and Intelligent Systems Science, Tokyo Institute of Technology
2013- Doctor of Engineering in Computational Intelligence and Intelligent Systems Science, Tokyo Institute of Technology

Membership in Academic Societies:

• Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)



Name:
Kazuhiro Ohnishi

Affiliation:
Doctor student, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology

Address:
G3-49, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

Brief Biographical History:
2012- Doctor student, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology

Main Works:
• K. Ohnishi, F. Dong, and K. Hirota, "Atmosphere Understanding for Humans Robots Interaction based on SVR and Fuzzy Set," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.18, No.1, pp. 62-70, 2014.

Membership in Academic Societies:
• Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)

Name:
Atsushi Shibata

Affiliation:
Ph.D. student, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology

Address:
G3-49, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

Brief Biographical History:
2005-2009 B.E., Aoyama Gakuin University
2010-2012 M.E., Department of Computational Intelligence and Systems Sciences, Tokyo Institute of Technology
2012- Ph.D. student, Department of Computational Intelligence and Systems Sciences, Tokyo Institute of Technology

Main Works:
• A. Shibata, J. Lu, F. Dong, and K. Hirota, "A Neural Structure Decomposition Based on Pruning and its Visualization Method," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.17, No.3, pp. 443-449, 2013.

Membership in Academic Societies:
• Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)

Name:
Fangyan Dong

Affiliation:
Associate Professor, Education Academy of Computational Life Sciences (ACLS), Tokyo Institute of Technology

Address:
J3-141, 4259 Nagatsuta, Midori-ku, Yokohama 226-8501, Japan

Brief Biographical History:
2006-2014 Assistant Professor, Tokyo Institute of Technology
2014- Associate Professor, Tokyo Institute of Technology

Main Works:
• F. Dong, K. Chen, E. M. Iyoda, H. Nobuhara, and K. Hirota, "Solving Truck Delivery Problems Using Integrated Evaluation Criteria Based on Neighborhood Degree and Evolutionary Algorithm," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.8, No.3, pp. 336-345, 2004.
• F. Dong, K. Chen, and K. Hirota, "Computational Intelligence Approach to Read-world Cooperative Vehicle Dispatching Problem," Int. J. of Intelligent Systems, Vol.23, pp. 619-634, 2008.

Membership in Academic Societies:
• Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)
• The Japanese Society for Artificial Intelligence (JSAI)

Name:
Kaoru Hirota

Affiliation:
Professor, Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology

Address:
G3-49, 4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

Brief Biographical History:
1982-1995 Professor, College of Engineering, Hosei University
1995- Professor, Tokyo Institute of Technology

Main Works:
• "Fuzzy Configuration Space for Moving Obstacle Avoidance of Autonomous Mobile Robots," J. of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.10, No.1, pp. 26-34, 2006.
• "Fuzzy Controlled Stepping Motors and Fuzzy Relation based Image Compression/Reconstruction (Banquet Keynote Speech)," 40th of Fuzzy Pioneers (1965-2005) BISC Special Event in Honor of Prof. Lotfi A. Zadeh (BISCSE2005), (U.C.Berkeley, U.S.A.), pp. 223-224, 2005.

Membership in Academic Societies:
• The Institute of Electrical and Electronics Engineers (IEEE)
• International Fuzzy Systems Association (IFSA)
• Japan Society for Fuzzy Theory and Intelligent Informatics (SOFT)