Journal of Mechanics Engineering and Automation 9 (2019) 92-99
doi: 10.17265/2159-5275/2019.03.003

A System of Associated Intelligent Integration for Human State Estimation

Akihiro Matsufuji, Wei-Fen Hsieh, Eri Sato-Shimokawara and Toru Yamaguchi

Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University, Hino, Tokyo 191-0065, Japan

Abstract: We propose a learning architecture for integrating multi-modal information e.g., vision, audio information. In recent years, artificial intelligence (AI) is making major progress in key tasks like a language, vision, voice recognition tasks. Most studies focus on how AI could achieve human-like abilities. Especially, in human-robot interaction research field, some researchers attempt to make robots talk with human in daily life. The key challenges for making robots talk naturally in conversation are to need to consider multi-modal non-verbal information same as human, and to learn with small amount of labeled multi-modal data. Previous multi-modal learning needs a large amount of labeled data while labeled multi-modal data are shortage and difficult to be collected. In this research, we address these challenges by integrating single-modal classifiers which trained each modal information respectively. Our architecture utilized knowledge by using bi-directional associative memory. Furthermore, we conducted the conversation experiment for collecting multi-modal non-verbal information. We verify our approach by comparing accuracies between our system and conventional system which trained multi-modal information.

Key words: Multi-modal learning, bi-directional associative memory, non-verbal, human robot interaction.

1. Introduction

In recent years, the social robot becomes more popular for use at home. Especially, in human-robot interaction (HRI) research field, communication robots which use any techniques of image, audio and natural language processing are expected to talk with people in daily life. According to psychological study [1], human estimates the human state unconsciously by using multimodal non-verbal information during communication. The human state is the internal state like emotion or timid/hard to talk about the topic. Thus, the ability to estimate human state is necessary for communication robots to realize smooth communication like a human.

For estimating or classifying tasks by machines, the neural networks which utilize a large amount of labeled data for creating a classifier have been successfully applied to many pattern classification problems [2]. These classifiers solve human recognition problems [3], object detection [4] in image processing field, acoustic modeling for automatic speech recognition [5], and semantic analysis [6] in natural language processing field. These methods focus on realizing the end to end learning system by using a large amount of labeled data.

However, it is necessary to consider multi-modal information for estimating complex tasks like human state. It is difficult to accumulate the labeled data of multi-modal information at a large scale. Furthermore, neural networks for multi-modal learning model are quite complex and require labeled data more than general neural networks.

In this paper, we divide the multi-modal information into multi-simplified single modal information, and classifiers by learning simplified single modal information. Additionally, we integrate the results of classifiers by using rule-based architecture for estimating multi-modal information. This rule-based integration architecture is inspired by the methods which utilize structured knowledge to learn from small

Corresponding author: Akihiro Matsufuji, M.Sc. in computer science, research fields: deep learning and human robot interaction.

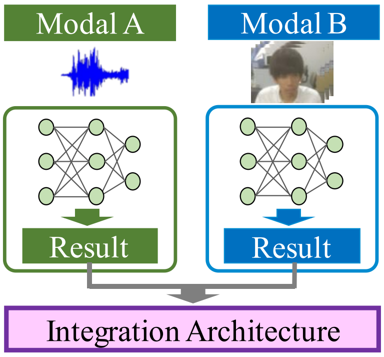

Fig. 1 Model overview. Our model uses multi-modal classifiers and integrates the results of each classifier which trained each modal information respectively.

amounts of experience for scene understanding [7]. Fig. 1 shows the concept of our proposed method. We utilized integrating multi-modal classifiers model to estimate the above-mentioned human internal state.

We conduct an experiment to obtain the non-verbal features for learning. In this research, we focus on the human state of "hard to talk" about topics. Furthermore, to verify the validity of the proposed method, we compare the accuracy between the conventional learning model which uses multi-modal features in a neural network model and proposed architecture.

This paper is organized as follows: Section 2 explains related works; our proposed method is described in Section 3; Section 4 introduces the experimental settings; Section 5 presents experimental results and discussions, and Section 6 gives conclusions.

2. Related Works

2.1 Multi-modal Non-verbal Features of Human Internal State

The sensors which obtain non-verbal features from human are divided into two types. One is contacted to human like heart rate sensor and polygraph [8]. The other is non-contact sensor like camera and microphone. In this research, we aim to apply for communication robots in daily life and non-contact sensors which are embedded in robot are appropriate for communication robot. In the next subsections, we discussed about non-contact sensors for capturing voice features and vision features respectively.

2.1.1 Voice Features

Kismet [9] used an indirect approach to analyze the emotional content of interlocutor voices. First, it classified voices into one of 5 categories (approval, prohibition, comfort, attention and neutral) using a Gaussian Mixture Model (GMM) [10] trained on 12 features. These categories were then mapped by hand onto affective dimensions of arousal, valence, and stance (defined as [A, V, S]). For instance, approval was mapped to medium high arousal, high positive valence, and an approaching stance. These [A, V, S] values in turn were fed into Kismet's emotional system to produce an appropriate social response. Neurobaby [11] was a simulated infant that responded to changes in voice, using a neural network to detect one of four emotional states, and a robotic hand interface that detected intensity. DePaulo et al. [12] used pitch and power of voice as features to estimate the participants lie. It shows that the non-verbal information of voice (pitch, power) is the important factor of estimate human internal state.

2.1.2 Vision Features

Thermal imaging is able to record the thermal patterns and measure the blood flow of the body [13]. However, thermal cameras are quite expensive, and RGB cameras and depth cameras get attention for capturing visual features. Early works [14], used blob analysis to track head and hand movements, which were used to classify human behavior in videos in three different behavioral states. However, these methods used images of specific sample person for training blob detector. Furthermore, since the database was small, the methods were prone to overfitting and did not generalize new subjects. Mancini et al. [15] created a real-time system for detecting emotions which were expressed through dancing video. However, the emotional behavior through dancing is largely different from the emotional behavior in daily life. The work of Ballihi et al. [16] detects positive/negative

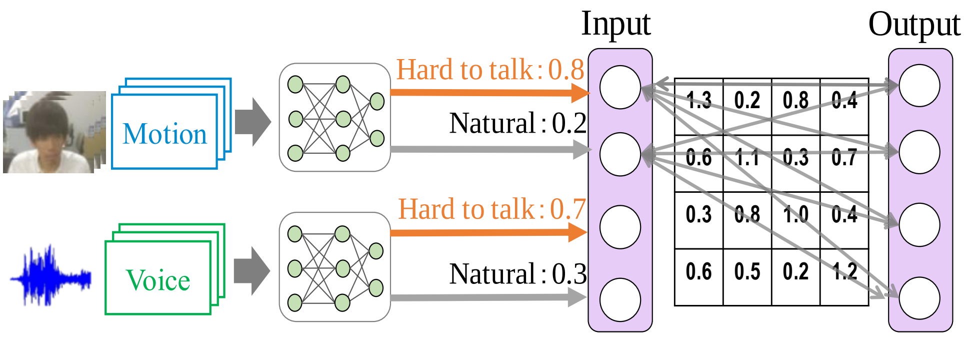

Fig. 2 The architecture of integrating multi-modal classifiers system.

emotion from RGB-D data. They classify the intensity of each expression using the upper body motion and face expressions. It shows that upper body movement is important factors of estimating human internal state.

2.2 Multi-modal Learning Methods

Recent successful neural networks methods are applied for tasks which utilize single model information. Furthermore, a lot of researchers attempt to solve complex tasks like the research of self-driving car and human state estimation using multi-sensor information. However, it is still difficult to estimate/understand the above-mentioned complex tasks by using simple modal model. Most researchers tend to solve such a complex task by using multi-modal cue information.

Afouras et al. [17] and Mroueh et al. [18] utilized the voice information and video information which captured rip movements for the robust voice recognition. This method realized to improve voice recognition accuracy by adding image information captured rip motion. Image captioning [19] is successful method of multi-modal learning. This method converted image to text. In contrast, text2image [20] converted text to image. As application, visual question answering [21] used natural language processing for conversation related to image information. But, to train multi-modal classifier needs a large amount of labeled data than to train single modal classifier. Furthermore, multi-modal labeled data are difficult to be collected.

Therefore, it is necessary to build a simplified architecture for integrating multi-classifiers which train single modal information.

3. Proposed Method

The architecture of our integrating multi-modal classifiers system is described in Fig. 2. Our model combined the neural networks classifiers by using associative architecture. To verify the validity of the proposed system, we set the learning task by using two features: motion and voice features during communication (details of features are described in Section 3.1).

3.1 Multi-modal Features for Learning

In this research, we aim to estimate the one of human internal states: “hard to talk” situation during communication. As the above-mentioned previous insights, we utilized the motion of upper body and acoustic of voice information. Furthermore, we analyzed these non-verbal information, and decided the learning features which have significant difference between “hard to talk” and “natural” situations. The analysis is described in Section 4.

We set the three neural networks. One is a model which trained motion features only. Second model trained voice features. These two classifier’s models are used to be integrated by our system. The other is the model which is trained by multi-modal features (motion and voice features). It is used for comparing the accuracy to our model.

3.2 Integration the Outputs of Classifiers

Our integrating learning system consisted of bi-directional associative memory (BAM) [22] construction. The BAM is the rule base architecture which defines rules by using if-then rule model. Our system has two layers (input and output). Input layer stored input rules which is the if part of if-then rule model. Output layer describes the inference results by calculation from integrating inputs, which is the then part of if-then rule modal. Input rules and output rules are described as matrixes and each rule is represented as one hot vector.

For learning our model, we defined the if-then rule and calculated a correlation matrix which connected the input layer and output layer according to defined if-then rule model. In this paper, each neural network learned voice/motion features respectively. These neural networks' inference results are utilized as input of our system and inference of the output by calculating with correlation matrix. The rule base method could be able to change easily according to the well-defined weight of NNs relationships or specific environment.

4. Experiment

4.1 The Procedure of Collecting Multi-modal Non-verbal Features

We conducted the conversation experiment that experimenter asks three questions in each topic to a participant. The three questions consist of two types. One is a binary question to answer yes or no. The other is a question that required participants to answer the reason following the previous yes/no binary questions. We asked two yes/no binary questions and a reason question in a topic, and we asked four topics during the conversation experiment.

For setting the "hard to talk" situation, we made participants answer opposite opinions from what they think in yes/no binary questions in two topics. Similarly, for setting the "natural" situation, we made participants answer own opinions naturally to questions in yes/no binary questions. The experimental process is illustrated in Fig. 3.

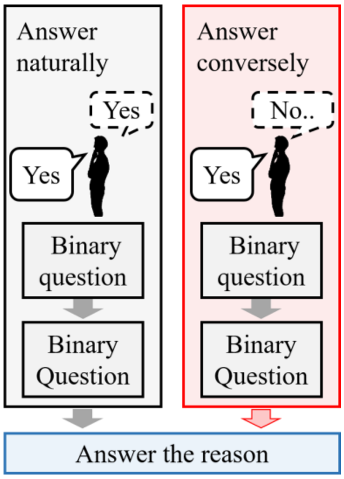

Fig. 3 The process of communication experiment.

When participants answered opposite opinions from what they think in yes/no questions, participants needed to consider a reason at question that required participants to answer the reason following the previous yes/no binary questions. Thus, we defined the duration that participants answer a reason following the previous yes/no binary questions as "hard to talk" situation in two topics. In contrast, participants answered own opinions naturally in yes/no binary questions for collecting "natural" features in other two topics. We conducted the conversation experiment for nine participants.

4.2 Experimental Settings

The experimental settings are shown in Fig. 4. A motion sensing device, Kinect (Microsoft, Kinect v2) is placed on the desk for capturing participant's motion during a conversation experiment. Our system is considered to capture behavior of a participant who is seated on a chair. It means that we recorded participant's upper body motion, head motion and voice simultaneously. The illumination setting of experimental is similar to general home's. In order to prevent erroneous other person recognition, the experiment settings allow only one participant in

Table 1 Noticeable difference of the “hard to talk” and “natural”.

Features	Participant A		Participant B		Participant C	
	Natural	Hard to talk	Natural	Hard to talk	Natural	Hard to talk
Head movement						
Pitch [m/frame]	0.08377	0.12509	0.05120	0.12946	0.06728	0.18781
Yaw [m/frame]	0.12230	0.18125	0.11140	0.14080	0.07185	0.25163
X-axis [m/frame]	0.02582	0.04527	0.01674	0.02544	0.21020	0.74050
Z-axis [m/frame]	0.03453	0.04737	0.02326	0.02715	0.01973	0.07992
Voice						
Maximum values of pitch [Hz]	594.34	372.73	427.81	305.79	459.87	129.97
Duration [s]	9.750	2.081	4.577	3.623	8.089	2.117

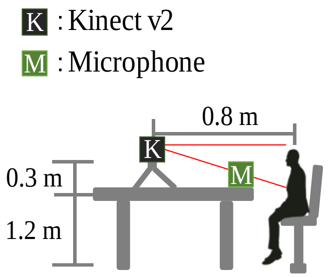

Fig. 4 The setting of experiments.

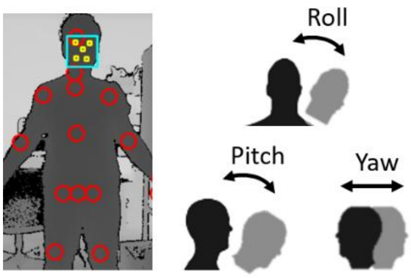

Fig. 5 Motion features extracted by depth sensor.

Kinect’s field of view. Kinect is installed over 0.8 m from a participant and 0.3 m in height from desk for obtaining skeleton robustly. The height of desk is 1.2 m. We use another microphone (SONY, ICD-SX1000) for recording the voice clearly in quiet environment. Each feature was described in the next section.

4.3 Multi-modal Features Extraction

In this paper, we utilized human user’s motion (gesture or posture) obtained by depth camera, and voice obtained by the microphone as multi-modal non-verbal features. Fig. 5 shows the motion features extracted by Kinect. Furthermore, we select features that make a noticeable difference as learning features. Table 1 shows the features which have the noticeable differences of the two state: “hard to talk” and “natural” from three participants. The bold numbers described the numbers which are larger than opposite human state. Motion features are head movements and audio features are pitch and duration. We describe the detail of these features in next section.

4.3.1 Motion Features

Kinect is a marker-less motion capture RGB-D camera. Kinect has an infrared camera to recognize human users as skeleton data and follow their actions in the Kinect’s field of view. Skeleton data (human user ’s body) can be obtained by locating joints of tracked participants and track their movement. Moreover, Kinect can capture face feature points and head rotation (pitch roll yaw) data. The depth information from body motion is recorded by Kinect to create a velocity vector. The resolution of depth images which are obtained by Kinect is 512 × 424 pixels and frame rates are 15 frame per second. We used Kinect to capture human posture and record the voice at the same time. The features which have the noticeable difference are head’s rotation (pitch, yaw) and head’s movement (x-axis, z-axis).

Head’s rotations: these parameters (pitch and yaw) are obtained by tracking face features (eyes, mouse and noses). We used the velocity of these parameters.

Head’s movement: these parameters are obtained by tracking joint’s 3-dimentional axis of head.

4.3.2 Voice Features

Most implementation of the speech is converted into text form, and then processed by using natural language processing technologies. However, it is not only text words but also non-verbal information during human-human communication. Non-verbal information of voice is usually called voice signals. Voice signal is not verbal message, and consists of the tone and pitch of the voice, the speed, volume, range at which a message is delivered, pauses and hesitates between words. We obtained above-mentioned features from captured voice by using PRAAT software [23]. The features which have a noticeable difference are Pitch and Duration.

Pitch (F0): It is fundamental frequency of speech signal. We estimated max, minimum and average values of pitch.

Duration: It is the features model temporal aspects having the basis unit milliseconds, such as position in time or length intervals. We defined that the duration feature is the time of participant's speaking duration.

4.4 Integrating Multi-modal Classifiers System

4.4.1 The Learning of Neural Networks

Our model utilized neural network's inference outputs. These neural networks are trained in advance. Thus, we set three neural network models for comparing the accuracy. One is a model which trained motion features only. Second model trained voice features. These two classifier's models are used to be integrated by our system. The other is the model which is trained by multi-modal features (motion and voice features). It is used for comparing the accuracy to our model. We train each neural network by using Weka software [24]. The motion feature and voice feature that were collected during a reason question in four topics from nine participants are used for training. In the reason question of two topics which participants answer opposite opinions what they think at yes/no binary question, features are labeled as "hard to talk" while, in other two topics, features are labeled as "natural".

4.4.2 If-Then Rule Model

We defined the If-Then rule model ("natural" and "hard to talk") as input rule matrix and output inference matrix. We defined the 4 rules for BAM architecture. The rules were organized as follows:

Rule 1: IF motion is "hard to talk" and voice is "hard to talk" THEN "hard to talk".

Rule 2: IF motion is "hard to talk" and voice is "natural" THEN "hard to talk".

Rule 3: IF motion is "natural" and voice is "hard to talk" THEN "hard to talk".

Rule 4: IF motion is "natural" and voice is "natural" THEN "natural".

4.4.3 The Learning the Integrating Associative Model

The above-mentioned rules are represented as one hot vector like a [0.1] values. In this experiment, we defined "hard to talk" state as 1, and "natural" state as 0. In the multi-modal features, integrating patterns of motion and audio are described to [1,0,1,0] as a matrix. Moreover, for energy minimization problem, we employ the bi-polar translation to 0 for -1. Eq.(1) is used to calculate the correlation matrix M_{IO} which connected input and output layers.

$$M_{IR} = \sum_{k=1}^{n} I_k R_k^T \qquad (1)$$

where, the I, R are input and output rule matrixes, respectively. The n is the number of if-then rule. In the inference phase, we can calculate the result by multiplying the input of neural network's inference results and correlation matrix M_{IR}.

5. Results

In this section, we describe the result of comparison of the accuracy between neural networks which trained multi-modal features as conventional model and our integrating multi-modal classifiers system which combined the single-modal neural networks. We evaluate the classification accuracy by the F-measures.

Table 2 The result of classification accuracy.

Methods	Precision	Recall	F-measure
Conventional model	0.625	0.625	0.625
Proposed method	0.750	0.667	0.706

F-measures are calculated by precision and recall. F-measure can be interpreted as a weighted average of the precision and recall. We employed the 10-fold cross validation to evaluate the precision, recall and F-measures. Table 2 shows the result of each classification accuracy. The result shows our proposed system's classification accuracy exceeds conventional model which is a neural network trained multi-modal information. From the result, our system is effective to integrate neural networks which trained single-modal information.

6. Conclusions

In this paper, we presented the integrating classifier's model to combine the pre-trained single-modal neural networks by using rule base architecture. Our proposed system estimated the human internal state during the communication by using non-verbal information. We utilized the motion and voice information of human as multi-modal non-verbal information. Furthermore, we conducted the conversation experiment and obtained above-mentioned features for estimating human internal states ("hard to talk" or "natural") during communication. To verify the validity of the proposed system, we compared the accuracies of our proposed method which combined single-modal neural networks and conventional method which trained multi-modal feature. The result shows that our proposed method is more accurate than conventional method. Our proposed method is able to be applied to complex tasks by using multi-modal cue information.

References

[1] Mehrabian, A. 1996. "Pleasure-Arousal-Dominance: A General Framework for Describing and Measuring Individual Differences in Temperament." *Current Psychology* 14 (4): 261-92.

[2] Christian, S., Alexander, T., and Dumitru, E. 2013. "Deep Neural Networks for Object Detection." In *Advances in Neural Information Processing Systems*, pp. 2553-61.

[3] Dodge, S., and Karam, L. 2017. "A Study and Comparison of Human and Deep Learning Recognition Performance under Visual Distortions." In *Proceedings of the International Conference on Computer Communications and Networks*, pp. 1-7.

[4] Borji, A., and Itti, L. 2014. "Human vs. Computer in Scene and Object Recognition." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113-20.

[5] Ying, Z., Pezeshki, Z., Brakel, M., Zhang, P., Bengio, S., and Aaron, C. 2017. "Towards End-to-End Speech Recognition with Deep Convolutional Neural Networks." arXiv preprint arXiv:1701.02720.

[6] Young, T., Hazarika, D., Poria, S., and Cambria, E. 2017. "Recent Trends in Deep Learning Based Natural Language Processing." arXiv preprint arXiv:1708.02709.

[7] Gay, P., Stuart, J., and Bue, A. D. 2018. "Visual Graphs from Motion (VGfM): Scene Understanding with Object Geometry Reasoning." arXiv preprint arXiv:1807.05933.

[8] Andrew, F., Kevin, K., Johnson, A., Emily, M., Grenesko, L., Laken, J., and George, M. 2009. "Detecting Deception Using Fuctional Magnetic Resonance Imaging." *Biological Psychology* 27 (1): 46-7.

[9] Breazeal, C. L. 2004. *Designing Sociable Robots*. Cambridge: The MIT Press.

[10] Manuel, A., and Bond, S. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58 (2): 277-97.

[11] Yamada, T., Hashimoto, H., and Tosa, N. 1995. "Pattern Recognition of Emotion with Neural Network." In *Proceedings of IECON*, pp. 183-7.

[12] DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. 2003. "Cues to Deception." *Psychological Bulletin* 129: 74-118.

[13] Buddharajum, P., Dowdall, J., Tsiamytzis, P., Shastis, D., Pavlidis, I., and Frank, M. G. 2005. "Automatic Thermal Monitoring System (Athemos) for Deception Detection." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2: 1179.

[14] Lu, S., Tsechpenakis, G., Metaxas, D. N., Jensen, M. L., and Kruse, J. 2005. "Blob Analysis of the Head and Hands: A Method for Deception Detection." In *Proceedings of 38th Annual Hawaii International Conference on System Sciences*, p. 20.

[15] Mancini, M., and Castellano, G. 2007. "Real-Time Analysis and Synthesis of Emotional Gesture Expressivity." In *Proceedings of the Doctoral Consortium of 2nd International Conference on Affective*

Computing and Intelligent Interaction.

[16] Ballihi, L., Lablack, A., Amor, B. B., Biasco, I. M., and Daoudi, M. 2015. *Positive Negative Emotion Detection from RGB-D Upper Body Images, Face and Facial Expression Recognition from Real World Videos*. Springer International Publishing, vol. 8912, pp. 109-20.

[17] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. 2018. "Deep Audio-Visual Speech Recognition." arXiv preprint arXiv:1809.02108.

[18] Mroueh, Y., Marcheret, E., and Goel, V. 2015. "Deep Multimodal Learning for Audio-Visual Speech Recognition." In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2130-4.

[19] Lu, J., Xiong, C., Parikh, D., and Socher, R. 2017. "Knowing When To Look: Adaptive Attention via a Visual Sentinel for Image Captioning." Presented at IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[20] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. 2016. "Generative Adversarial Text to Image Synthesis." arXiv preprint arXiv:1605.05396.

[21] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. 2015. "Vqa: Visual Question Answering." In *Proceedings of the IEEE International Conference on Computer Vision*.

[22] Kosko, B. 1987. "Adaptive Bidirectional Associative Memories." *Applied Optict* 26 (23): 4947-60.

[23] Boersma, P. 2001. "A System for Doing Phonetics by Computer." *Glot International* 5 (9/10): 341-5.

[24] Frank, E., Hall, M. A., and Witten, I. H. 2009. "The WEKA Data Mining Software: An Update." *SIGKDD Explorer* 11 (1): 10-8.