Multimodal Feature Fusion for Human Personality Traits Classification*

Zhihao Shen, Armagan Elibol, and Nak Young Chong

***Abstract*— Similar to human-human social interaction, the process of inferring a user's personality traits during human-robot interaction plays an important role. Robots need to be endowed with such capability in order to attract user engagement more. In this study, we present our on-going research on obtaining variable-length multimodal features and their fusion to enable social robots to infer human personality traits during face-to-face human-robot interaction. Multimodal nonverbal features, including head motion, face direction, body motion, voice pitch, voice energy, and Mel-frequency Cepstral Coefficient (MFCC), were extracted from videos and audios recorded during the interaction. The different combinations of multimodal features were verified, and their classification performance was compared.**

I. INTRODUCTION

Human social interaction is a complex process of understanding and responding to the counterpart's behaviors. Essential factors of interaction (*e.g.*, behavior, emotion, and thought) are also known as the reflections of an individual's personality traits [1]. With an increasing number of research on personality traits [2], their relationship to many important aspects of life has been revealed. If humans are willing to interact more, they adapt their behaviors based on the impression of the personality traits on each other to enrich the interaction.

Various studies have been proposed for inferring human personality traits by using many different resources including group meetings [3][4], YouTube vlogs [5], human-robot interactions [6], etc. Instead of analyzing a large number of words, which is an tedious work, the nonverbal behavioral cues are a better choice for inferring human personality traits. Looking into the research based on nonverbal behaviors, three main aspects will be addressed in this study:

1) Will the performance of inferring human personality traits be improved by fusing different multimodal features?
2) It would be difficult to collect all the samples with an equal time interval. Therefore, how to infer personality traits based on the features with variant lengths?
3) The robots will make some movements during an interaction with a human in order to make the human-robot interaction as natural as possible. Therefore, how to extract visual features from the videos that were recorded with a moving camera?

All the above mentioned questions were taken into consideration in this study.

*The authors are grateful for financial support from the Air Force Office of Scientific Research under AFOSRAOARD/ FA2386-19-1-4015.

The authors are with the School of Information Science, Japan Advanced Institute of Science and Technology {shenzhihao, aelibol, nakyoung}@jaist.ac.jp

II. EXPERIMENTAL SETUP

For addressing the research questions, we designed the human-robot interaction scenario in the laboratory setting. The human nonverbal behavior data and personality traits were collected during human-robot interaction to train a supervised learning model.

A. Human Personality Traits Annotations

The psychologists used personality traits to describe individual differences. Most of the existing studies on personality traits have referred to the Big-Five personality traits (Extroversion, Openness, Emotional Stability, Conscientiousness, Agreeableness) [7]. In this study, the International Personality Item Pool (IPIP) Big-Five Factor Markers [8] was used to assess the personality traits of each participant. We used the mean score of all participants as a cut-off point to binarized the personality traits of each participant. The binary personality traits were used to perform a classification task and indicate how high or low the participants rated their personality traits.

B. Experimental Setup

In a separate room, each participant sat in front of the Pepper robot[1] between 1.5 to 2 meters. In case of any robot failures, the operator was also present in the room during the entire of period of interaction. A camera and a microphone in the middle of the robot's head were used to record the video and audio during the whole interaction. The robot was also enabled to track the human's head, and some small movements to indicate that the robot is listening at the same time. The resolution of the camera was set to 640×480 pixels, and the frame rate was set to 5 frames per second. Simultaneously, the robot would record the audio with the sample rate of $16,000Hz$ by the microphone. We recruited 21 participants for the experiment.

C. Methodology

Fig. 1 illustrates the overview of the proposed framework explaining how to process the feature data for estimating human personality traits. The details of the feature extraction were omitted. Fig. 1 can be further explained in the following five steps:

Step 1: The visual and vocal features were extracted from video and audio, respectively, based on the previous research [6] (head motion, gaze, body motion, voice pitch, voice energy, and MFCC). Since the visual and vocal features were extracted with the different sampling rates, the length of the

[1]https://www.softbankrobotics.com/emea/en/pepper

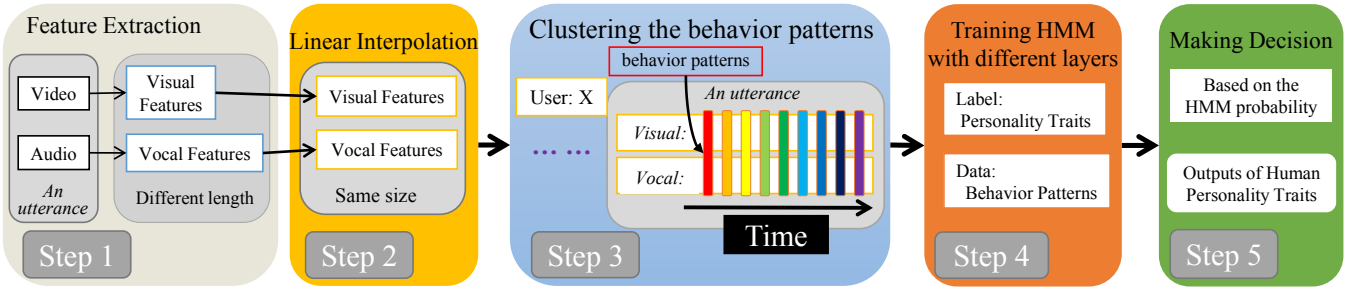

Fig. 1. Overview of the Proposed Framework

visual feature is different from the length of the vocal feature, even they were extracted from the same sentence.

Step 2: The linear interpolation was applied to the visual features to generate new features whose length equals the length of vocal features.

Step 3: All the features from the training data were gathered to generate a matrix, where each row is an independent feature. The column vector represents a behavior pattern at a specific time point, *e.g.*, the person was facing to a robot or not, was there a significant movement comparing to the last time point, was the person using high or low voice pitch to talk, etc. The behavior patterns were clustered into several categories.

Step 4: The feature matrix of each sentence from the training data was represented by a consecutive series of category labels representing the different behavior patterns that happened at a specific time point. The time-based arrays were used to calculate the initial probabilities and state transition probabilities based on the concept of HMM. Since the duration of representing each behavior could vary, we combined every two or more behavior patterns as one pattern to generate the second-layer or more in order to compute initial and state transition probabilities.

Step 5: Based on the results of combination of multiple layers HMM probability, we used the SVM with different kernel functions, and voting method to reach a final decision on the user's personality trait.

III. Experimental Results

Based on the proposed framework outlined above, we tested a part of feature combinations and presented the results in Table. I, where ***Lin*** is the SVM classifier with the linear kernel, ***RBF*** is the SVM classifier with the RBF kernel, and ***SIG*** is the SVM classifier with the Sigmoid kernel, respectively. Note that ***Voting*** is majority voting, which means that the final output is the label that received more than half of the votes.

In the table, ***C*** is the number of clusters used in Step 3, ***F*** is the index of the feature combination, and ***L*** is index of the layer combination (maximum layer is 6). In detail, ***F46*** is the combination of Head Motion, Gaze, Energy, and MFCC1 which is the first feature vector of 13 MFCC vectors. ***F39*** is the combination of Body Motion, Energy, and MFCC1. ***F4*** is Energy. **F15** is the combination of Body Motion and Pitch.

TABLE I

Classification Results of Big Five Personality Traits

Trait	Lin	RBF	SIG	Voting
Extroversion	C6F53L12	C8F54L52	C3F16L0	C7F46L45
	0.6534	0.6629	0.6645	**0.6805**
Agreeableness	C4F42L39	C7F23L11	C4F39L2	C7F23L26
	0.6837	0.6869	**0.7589**	0.6901
Conscient-iousness	C8F58L47	C8F58L34	C3F4L50	C8F46L59
	0.6629	0.6629	**0.6997**	0.6837
Emotional Stability	C5F23L41	C6F16L56	C3F13L34	C7F15L57
	0.7476	0.7540	0.7508	**0.7652**
Openness	C6F32L6	C8F42L2	C8F42L11	C8F23L57
	0.7732	0.7716	0.7604	**0.7843**

F23 is the combination of Head Motion, Gaze, and Energy. ***L45*** is the combination of 1st, 2nd, 4th, and 6th layer. ***L3*** is the 3rd layer. ***L50*** is the combination of 1st, 4th, 5th, and 6th layer. ***L57*** is the combination of 1st, 2nd, 3rd, 4th, and 6th layer.

References

[1] G. W. Allport. Personality: a psychological interpretation.Holt. 1937.

[2] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing, jul 2014.

[3] Shogo Okada, Oya Aran, and Daniel Gatica-Perez. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction*, pages 15–22, New York, New York, USA, nov 2015. Association for Computing Machinery, Inc.

[4] Oya Aran and Daniel Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction*, pages 11–18, 2013.

[5] Joan Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 15(1):41–55, 2013.

[6] Zhihao Shen, Armagan Elibol, and Nak Young Chong. Nonverbal behavior cue for recognizing human personality traits in human-robot social interaction. In *2019 4th IEEE International Conference on Advanced Robotics and Mechatronics, ICARM 2019*, pages 402–407. Institute of Electrical and Electronics Engineers Inc., jul 2019.

[7] Lewis R. Goldberg. An Alternative "Description of Personality": The Big-Five Factor Structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229, 1990.

[8] Lewis R. Goldberg. The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment*, 4(1):26–42, 1992.