Multi-modal Feature Fusion for Better Understanding of Human Personality Traits in Social Human-Robot Interaction

ARTICLE INFO

Keywords:
human-robot interaction
human personality traits
multi-modal feature fusion
machine learning

ABSTRACT

Since the dynamic nature of human-robot interaction becomes increasingly prevalent in our daily life, there is a great demand for enabling the robot to better understand human personality traits and inspiring humans to be more engaged in the interaction with the robot. Therefore, in this work, as we design the paradigm of human-robot interaction as close to the real situation as possible, the following three main problems are addressed: (1) fusion of visual and audio features of human interaction modalities, (2) integration of variable length feature vectors, and (3) compensation of shaky camera motion caused by movements of the robot's communicative gesture. Specifically, the three most important visual features of humans including head motion, gaze, and body motion were extracted from a camera mounted on the robot performing verbal and body gestures during the interaction. Then, our system was geared to fuse the aforementioned visual features and different types of vocal features, such as voice pitch, voice energy, and Mel-Frequency Cepstral Coefficient, dealing with variable length multiple feature vectors. Lastly, considering unknown patterns and sequential characteristics of human communicative behavior, we proposed a multi-layer Hidden Markov Model that improved the classification accuracy of personality traits and offered notable advantages of fusing the multiple features. The results were thoroughly analyzed and supported by psychological studies. The proposed multi-modal fusion approach is expected to deepen the communicative competence of social robots interacting with humans from different cultures and backgrounds.

1. Introduction

The way we interact with other people is a complex process of understanding and responding to others' behavior. The essential factors that affect the social interaction (*e.g.*, behavior, emotion, and thought) are deemed to be 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, such as job performance [3] and health-related behaviors [4] have been revealed. Understanding personality traits is useful for predicting human behaviors [5, 6], and understanding the human's mind and how personality traits affect the attitude and behaviors towards other people [7]. If humans like their co-communicators in terms of personality traits and are willing to interact more, they will adapt their behaviors based on the impression of the personality traits on each other to enrich the interaction. It does not require much effort for humans to assess the personality traits of their counterparts. We are able to judge the personality traits of other people by looking at their face for 100 ms [8]. Although, the first impression may not be always correct [9], subsequently, a short interaction will increase the accuracy of personality recognition [10].

Over the last decade, social robots have been promoted for assisting people in their daily life in order to mitigate the problem associated with aging populations. It has been predicted that human-robot relationship may be more common than human-human connection by 2050 [11, 12]. Social robots will interact with humans in domestic environments and become a part of our life in the future [13]. Therefore, there is a huge demand for endowing machines with social intelligence and capabilities of interacting with humans in a natural manner [14]. For this purpose, some robots were designed with an appearance similar to humans [15], and capabilities such as synchronized verbal and nonverbal behavior [16], cultural competence [17], and emotional bodily expression [18]. More importantly, social robots will need to interact with humans with synchronized verbal and nonverbal behavior in alignment with their personality traits [19].

Recent research in human-robot interaction has been directed at investigating the relationship between human's attitudes towards the robot and human personality traits. One of the earliest works is [20], where humans made a judgment of the personality from the computer-synthesized sounds. Some robots or dialog machines were endowed with the capability of changing their personality traits (extroversion or introversion) to interact with humans [21, 22]. There are two contradictory aspects to the personality issue between humans and robots. Some people enjoy interacting with a robot with a similar personality [23], which is consistent with the similarity attraction principle [24]. On the contrary, some people prefer talking with others with the complementary personality traits to their own. The complementarity attraction was uncovered in human-robot or computer interaction [25, 26]. These two principles were considered in human-robot interaction for the purpose of analyzing the relationship between engagement and personality [27], and synchronizing verbal and nonverbal behavior based on personality traits [28]. Additionally, some studies [29, 30, 31] revealed that people who treat robots with more positive attitudes scored high on extroversion or openness to experience.

Not only the behavior, but also the profession affects people's impression on personality traits [32]. Generally, we tend to believe that doctors and teachers are introverted, while managers and salespersons are extroverted. However, during human-robot interaction, the robots that acted as extro-

ORCID(s):

verted teachers are perceived as more intelligent than introverted ones. Furthermore, the introverted manager robots are discerned more intelligent than the extroverted ones [28].

The aforementioned studies clearly explained the importance of the personality traits in human-robot interaction. In this study, we aim to enable the robot to better understand human personality traits. The robot can then adjust its behaviors such as voice volume, speech rate, and body movements to enhance the degree of user engagement.

1.1. Research problem

Various studies have been performed in different contexts to recognize human personality traits through different resources, including words used in blogs [33] or self-narratives [34], videos and audios in group meetings [35, 36], YouTube vlogs [37], and human-robot interaction [38]. Nonverbal behavioral cues are well-suited choices for inferring human personality traits instead of analyzing a large number of words, which is commonly applied. In our earlier works [38, 39], each nonverbal feature showed its advantage in a different aspect. However, different features can provide different personality traits classification results. This makes it hard to draw the conclusion in a standardized way for declaring the user's personality traits. Then, the problem arose as to how to unify the classification results, or how to fuse the multi-modal features. Furthermore, in our works, the robot posture (thus the camera located in the forehead) remained unchanged. This conflicts with the general idea of social robots designed to interact with humans freely changing their posture. Based on prior studies on nonverbal behaviors, the following three problems will be addressed in this paper:

(1) How can the accuracy of inferring personality traits be improved by combinations of multi-modal features?

multi-modal feature fusion has drawn increasing attention from researchers in analyzing various multimedia data [40, 41, 42]. Usually, the statistical features of audio and video were concatenated to generate a fusion vector, or used to analyze the co-occurrent event [36]. Most of the methods proposed for feature fusion rarely investigate how to selectively combine features. [43] mentioned some methods of combining the features of the target person and the other group members to recognize the personality traits of the target person. Therefore, we investigated whether it is necessary to use all the features available and what combination of features can achieve the best accuracy for inferring human personality traits.

(2) It is technically difficult to sample different features at equal intervals. How can the feature vectors of variable length be handled?

Dealing with the feature vectors of variable lengths is another important point to address. In [44], audio and video were input to the framework that combined Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs) to generate a fused vector handling variable length features. However, training a neural network required a large number of data. In the speech recognition, the dynamic Bayesian networks can process multi-stream features and features of variable lengths [45].

(3) The robot often makes communicative gestures during the interaction with humans. How can the video blurred by the robot's shaky camera be stabilized?

As previously stated, robots will need to interact with humans through synchronized verbal and nonverbal behaviors aligned with human personality traits. For this, robots need to analyze the video taken from the robot's first-person perspective with an on-board camera. However, to the best of our knowledge, previous studies [35, 36, 37, 38, 44] only used a fixed camera position. Some studies allowed the robot to move, however, features were extracted from an external RGB-D sensor placed above the robot's head [46] analyzing human motion and distance change to the robot. Likewise, a depth sensor was placed behind the robot to record human-robot interactions and to analyze the relationship between engagement and personality [27]. We believe that it is important to extract the nonverbal behaviors, such as eye contact, head movements, and body movements, from the robot's first-person perspective to better understand human characteristics using a self-contained system.

The Pepper robot equipped with cameras and microphones was used to interact with participants. The robot performs movements during the interaction, and records the audio and video data of each participant at the same time. The visual and vocal nonverbal features were extracted from the video and audio while the human was talking as shown in Fig.1. The utterances with different lengths were used to train a model for inferring human personality traits.

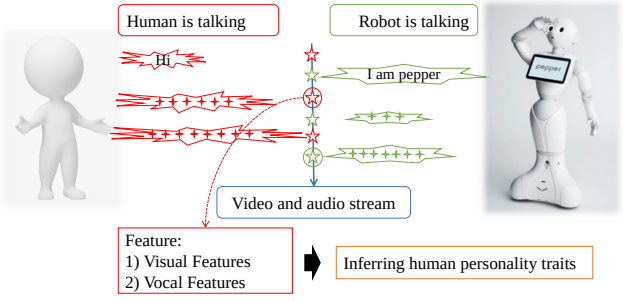

Figure 1: Diagram of human-robot interactions

The rest of this paper is organized as follows. Section 2 introduces how we annotated the participants' personality traits, the experimental setup, and overview of the system architecture. Section 3 explains related works on nonverbal feature representations and our methods to extract nonverbal features. Section 4 highlights the details of the feature fusion and classification models proposed. Section 5 presents and analyzes the experimental results with comparisons to the baseline. Section 6 draws conclusions and future work.

2. Annotations and experimental setup

We designed human-robot interaction in the laboratory setting to address the research questions. The human nonverbal behavior data and personality traits were collected during human-robot interaction to train supervised learning models.

2.1. Human personality traits annotations

Personality traits strongly affect how humans behave through their lifetime: "*the pattern of collective character, behavioral, temperamental, emotional, and mental traits of an individual that has consistently over time and situations*" [22, 47]. Psychologists used personality traits to describe individual differences. Most existing research on personality traits discusses the Big-Five personality traits (Extroversion, Openness, Emotional Stability, Conscientiousness, Agreeableness) [48, 49]. An intuitive impression on this five-factor model was presented in our previous study [38] and also used in this study.

On the other hand, various questionnaires have been designed over the last few decades in order to assess human personality traits. Most of the popular questionnaires are formatted to the Likert scale such as the Ten Item Personality Inventory (TIPI), which includes 10 questions and each question is rated on a seven-point scale [50]; the Revised NEO Personality Inventory (NEO PI-R contains 240 items) [51]; the shortened version NEO Five-Factor Inventory (NEO-FFI contains 60 items) [52]; and the International Personality Item Pool (IPIP) Big-Five Factor Markers (50 items) [53]. Comparing all these questionnaires, we found that the questions in the IPIP Big-Five Factor Markers were designed to be easily understandable from the participant's perspective, such as "leave my belongings around", "feel comfortable around people", to name a few. In this paper, the IPIP Big-Five Factor Markers were used to assess the personality traits of each participant.

Each participant was asked to fill out an IPIP questionnaire. A total number of 50 questions are divided into 5 groups to describe 5 different personality traits. Each group contains 5 positive-scored questions that positively describe a personality trait and 5 reverse-scored questions that negatively describe a personality trait. Each question is rated on a five-point scale. For the positive question: Strongly Disagree equals 1 point, Neutral equals 3 points, and Strongly Agree equals 5 points. The rating for reverse-scored questions is just the opposite. The final score of each personality trait is the average score of 10 questions. Then, we used the mean score of all participants as a cut-off point to binarize the personality traits of each participant [35, 36]. The binary personality traits were used to perform a classification task and indicate how high or low the participants rated their personality traits.

The blue and red bars in Fig. 2 are the number of participants that were scored low and high, respectively, on the personality traits compared to the mean scores from the questionnaire survey. A total of 21 participants were recruited from the Japan Advanced Institute of Science and Technology. Each participant asked questions such as "how can I borrow a book from the library?", and "I am worried a lot about my research" to the robot. We synchronized the video and audio that were recorded separately. The noises of the robot's fan were also removed from the audio. The timestamps that indicate when the participant started talking and when the participant finished talking were not completely accurately recorded. Therefore, the timestamps were manually revised. Then, multi-modal features were extracted while participants were asking questions.

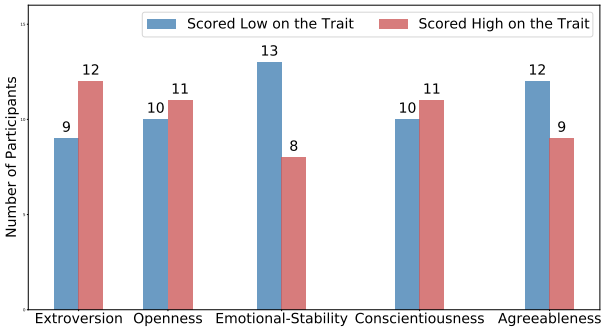

Figure 2: Number of participants that scored high or low on each personality trait compared to the mean scores

On the other hand, there is a public dataset of the big-five personality traits scores available online. Nearly twenty thousand people from more than one hundred and fifty countries answered the questionnaire (IPIP Big-Five Factor Markers), and their data were collected and placed in the category of "BIG5" [1]. The mean scores of five personality traits were presented in Table 1. The first row shows the mean scores of all participants in our study, which were used as the cut-off points. And the second row shows the mean scores of the people who participated in the IPIP Big-Five Factor Markers questionnaire.

We also presented Table 2 to show how many participants in our study score high on each trait depending on two different cutoff points (mean scores) given in Table 1. The first row of Table 2 shows the number of participants who score high on each trait using the cutoff points of our study (21 samples). The second row shows the number of participants who score high on each trait using the cutoff points of the IPIP Big-Five Factor Markers questionnaire participants (19,719 samples).

The differences in extroversion, emotional stability, and conscientiousness presented in the two tables are negligible, while the differences in openness and agreeableness are seemingly notable. In our experiments, almost all participants (20 out of 21) were international postgraduate students. In the literature, a study [54] reported that *most of the international postgraduate students rate high in agreeableness, openness, and conscientiousness, while extroversion and neuroticism are subsequently at medium levels.* The findings in the above-mentioned study are consistent with the data of our participants that showed considerably high cutoff points on the openness and agreeableness scales.

[1] Big Five Personality Test: https://openpsychometrics.org/_rawdata/

Table 1
The mean scores of five personality traits of our study participants and IPIP Big-Five Factor Markers questionnaire participants.

Personality Trait	Extroversion	Openness	Emotional Stability	Conscientiousness	Agreeableness
Our study	**3.0286**	3.8048	**2.9571**	**3.5381**	3.9048
Public dataset	**3.1499**	3.0357	**2.8290**	**3.2072**	2.9011

Table 2
The number of participants who score high on each trait based on different cutoff points.

Personality Trait	Extroversion	Openness	Emotional Stability	Conscientiousness	Agreeableness
Our study	**9**	10	**13**	**10**	12
Public dataset	**9**	20	**15**	**14**	21

Furthermore, Hypothesis Tests including the T-test and Kolmogorov-Smirnov-test (KS-test) were also performed and presented in Table 3. The results of T-test were presented on the first row of Table 3. The second row of Table 3 showed the results of Kolmogorov-Smirnov-test. The null hypotheses of T-test and Kolmogorov-Smirnov-test are given below:

T-test : the data in vectors b_1 and b_2, which represent the personality trait scores of the participants in our study and IPIP dataset, come from independent random samples from normal distributions with equal means and equal but unknown variances at the 5% significance level.

KS-test : the data in vectors b_1 and b_2, which represent the personality trait scores of the participants in our study and IPIP dataset, are from the same continuous distribution at the 5% significance level.

The results of hypothesis tests are in line with our previous analysis associated with Table 2. There are comparatively small number of participants in our study. However, their personality traits distribution is representative.

2.2. Experimental setup

The semi-humanoid robot Pepper[2] was used to interact with and create the video and audio recordings of behaviors and interactions of each participant. Pepper was equipped with more than 300 applications including speech recognition engine, speech engines, and interaction engines.

Fig. 3 shows how we enabled the Pepper robot to communicate with each participant. It consists of two parts: the built-in NAOqi applications [3] and the natural language understanding platform (Dialogflow [4]). A similar method also was used in [55] as a smart home user interface [56]. The built-in speech recognition engine provided by NUANCE converts speech to text. The text is then sent to Dialogflow for acquiring a proper response. As soon as the robot received the response, the NUANCE speech engine synthesizes the speech to communicate with each participant. To avoid spending too much time designing a conversational interface that covers multiple topics, we proactively narrowed down the topics, mainly related to our campus life. The robot played as an advisory staff providing such information as research laboratories and facilities on campus as well as students' welfare services.

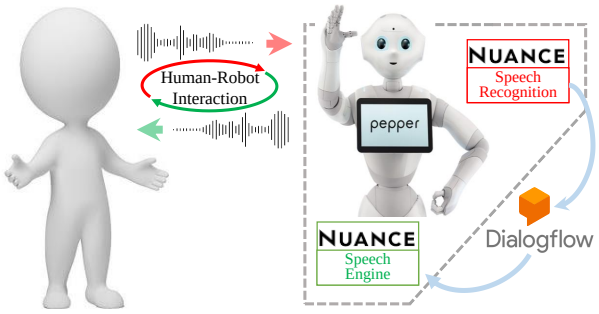

Figure 3: Spoken dialog system using NUANCE and Dialogflow

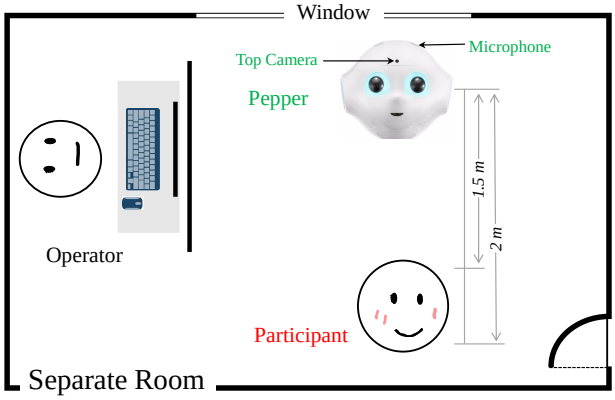

Figure 4: Floor plan of the experimental room

In a separate room, each participant sat in front of the robot 1.5 to 2 meters away as shown in Fig. 4. In order to respond to robot failures, an operator was present in the room during the interaction. Sometimes the robot abruptly looked up at the ceiling due to air conditioner noises. Then, the operator would tell the participant and terminate the inter-

[2] Softbank Pepper robot: https://www.softbankrobotics.com/emea/en/pepper

[3] NaoQi documentation: http://doc.aldebaran.com/2-5/index_dev_guide.html

[4] Dialogflow: https://dialogflow.com/

Table 3
The results of hypothesis tests.

Personality Trait	Extroversion	Openness	Emotional Stability	Conscientiousness	Agreeableness
T-test	**Accept**	Reject	**Accept**	Reject	Reject
KS-test	**Accept**	Reject	**Accept**	Reject	Reject

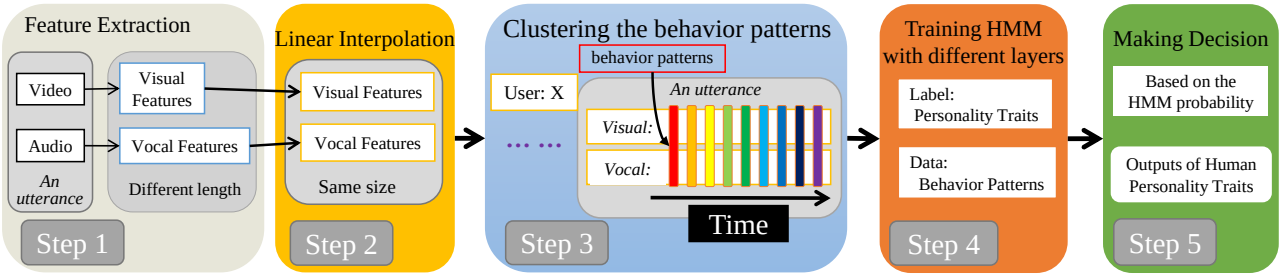

Figure 5: Overview of the proposed framework

action. After resolving such problem, the participant was asked to keep on interacting with the robot. In our experiments, all the participants were initially asked to interact with the robot for a while before starting an experiment (training of the participants). Once they felt that they became familiar with interacting with the robot, they could start the experiment.

All the participants (from China, Italy, Vietnam, Thailand, and Turkey) were asked to interact with the robot by using English. Due to the accent, sometimes, the speech recognition engine was not able to accurately translate the participant's speech to the texts. Human personality traits affect the way that people use their language. The participants basically tended to substitute their mother tongue's sounds for those of English. Their accent and speech patterns do not change much regardless of whether they speak their native language or English. We think this is also partly because the participants were speaking with the robot, not the native English speakers. The voice pitch and energy might be somewhat different when the participants use different languages, but their personality traits are the main factors that affect the way they generate their voice. On the other hand, we also considered that humans' vocal and visual (bodily) behaviors tend to synchronize, even when they were using a foreign language, which motivated a data-fusion approach in this study. Therefore, we did not design any conversational interfaces for different languages.

A camera and a microphone embedded into the robot head (as shown in Fig. 4) were used to record the video and audio during the interaction. The robot can track the human head movements to indicate that the robot pays attention to the person. The camera resolution was set to 640×480 pixels, and the frame rate was set to 5 frames per second. Simultaneously, the robot recorded the audio with the sample rate of $16,000 Hz$ by the microphone.

2.3. System architecture

Fig. 5 illustrates the overview of our framework for estimating human personality traits that will be detailed in the following five steps:

Step 1: The visual and vocal features, namely, head motion, gaze, body motion, voice pitch, voice energy, and Mel-Frequency Cepstral Coefficient (MFCC) were extracted from video and audio, respectively, following our prior research [38]. Since the visual and vocal features were extracted at different sampling rates, although they were extracted from the same sentence, the length of the visual feature is different from that of the vocal feature.

Step 2: The linear interpolation was applied to the visual features to make their length equal to 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?, the person was using high or low voice pitch while talking, 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 and later layers to compute initial and state transition probabilities.

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

3. Feature representation

Humans are surprisingly good at understanding others' nonverbal behavior [57]. We conducted a literature review with special attention to inferences of human characteristics like leadership [58] and personality traits from nonverbal behavior cues detected by microphones and cameras. And the process of extracting nonverbal features is detailed in our experimental setting.

3.1. Related work on nonverbal features

The study of proxemics [59], or the interpersonal distance, enabled the robot to change the distance adaptively to its users depending on social factors [60]. The extroverted person accepts people to come closer than the introverted person [61]. Distance changes during human-robot interaction have already been applied to predict if the participant is an extrovert or introvert [46]. In this study, however, proxemics is not taken into account due to the fact that if the distance between the camera and the object is changed, the size of the object in the image will change accordingly. This will greatly affect the visual features extracted through pixel-wise operations.

In [62], the head pose, facial expressions, body movement, body postures, and proximity information were used to assess human personality traits. Some studies exclusively focused on analyzing the correlation between vocal nonverbal features and personality traits [63, 64]. With the huge amount of training data, CNNs [65, 66] were applied to infer personality traits from audio, video, and text information. The nonverbal features such as prosodic feature (pitch and energy), visual features (head activity and body activity), and motion template-based features were used to identify the emergent leaders in a winter survival task scenario [58, 67]. The commonly used nonverbal features for predicting human personality traits or the emergent leaders were summarized in [68, 69]. Notably, the activity length features and statistical features were frequently used to represent the participants' behaviors, since the personality traits have the long-term effect on people's behaviors. However, how the behavior transits from one state to another is also intriguing. Therefore, in this paper, we investigated the time-series state transition of the human behavior from the visual and vocal nonverbal features to train the machine learning models.

3.2. Nonverbal feature extraction

We extract the above-mentioned nonverbal visual and vocal features. The brief descriptions of each feature were presented in Table 4. Under our human-robot interaction scenario, we performed the image stabilization compensating for shaky camera motion while extracting the visual features.

3.2.1. Head motion

In [58], the average of the optical flow vectors calculated from two successive frames within the face area represented the head activity. Different from [58], we measured the participant's head motion from the rotation of the head. Specifically, the Manhattan distance of the 3D head angles (roll, pitch, and yaw) of two adjacent frames was used to represent the head motion. A part of early studies on head pose estimation was summarized in [70]. How to distinguish the participant's head motion from the camera's rotation, however, was not mentioned in these studies. An interesting and straightforward geometric method was proposed in [71]. Hence, our head angle calculation method built upon the idea of [71] by minimizing the effects of camera movement, as will be detailed in the following content.

The robot moved its head while interacting with each participant. For calculating the 3D head angle from images, first of all, we have to minimize the effect of the camera's movements shown in Fig. 6 where two successive frames ($Image_i$ and $Image_{i+1}$) were used. The frame ($Image_{i+1}$) was warped based on the previous frame ($Image_i$) using a feature-based image registration pipeline by extracting distinctive points and matching them through descriptor vectors. If key points detected from the body of the participant were matched while he/she was moving, this would generate large errors in motion estimation thus warping the image. Therefore, the human was detected by a deep learning-based object detection model (*e.g.*, MobileNets [72] and SSD [73]), and removed from both images. The SIFT [74] was used to detect key points. Then, the RANSAC [75] algorithm was applied to uncover a set of optimal inliers of two images. Based on the matched point pairs, the 2D planar motion between the coordinate frames of the images can be easily calculated. The target image could be warped by using this motion matrix [76].

Once the image was warped, an open-source library dlib[5] [77] was used to detect the key points of the human face from the warped image. There are 68 facial landmarks that can be localized from the images as mentioned in [78]. Fig. 7 shows the facial key points that were localized using dlib and default 3D key points. We used six facial key points which include left corner of the left eye, right corner of the right eye, nose tip, left mouth corner, right mouth corner, and chin to calculate the 3D head angles (roll, pitch, and yaw).

The following equation[6] shows how the participants moved their head from the default pose to other poses which were projected to the images:

$$F_{2D} = K * [R|T] * P_{3D}, \tag{1}$$

where F_{2D} is the facial key points that were detected from the image, P_{3D} is the corresponding default 3D key points, K is the camera matrix, R is the 3×3 rotation matrix which indicates how participants rotated their head, T is a translation vector. The robot's camera was calibrated using the method proposed in [79]. Therefore, the rotation matrix R can be easily calculated by Eq. 1.

Eq. 2 shows how to calculate 3D head angles (roll, pitch, and yaw) in radians. Each element of the rotation matrix R

[5] dlib: http://dlib.net/

[6] OpenCV: https://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html

Table 4
Nonverbal feature representation

Activity	Abbr.	Description
Visual Nonverbal Features		
Head Motion	HM	A score describes the scale of the participants' head motion while they are talking to the robot
Gaze Score	GS	A score describes the confidence in the fact that the participant is looking at the robot
Body Motion	ME	A score describes the scale of the participants' body motion while they are talking to the robot
Vocal Nonverbal Features		
Pitch	Pt	The voice pitch of the participants
Energy	En	The voice energy of the participants
MFCC	$MFCC_s$	One of the 13 MFCC vectors, s is from 1 to 13

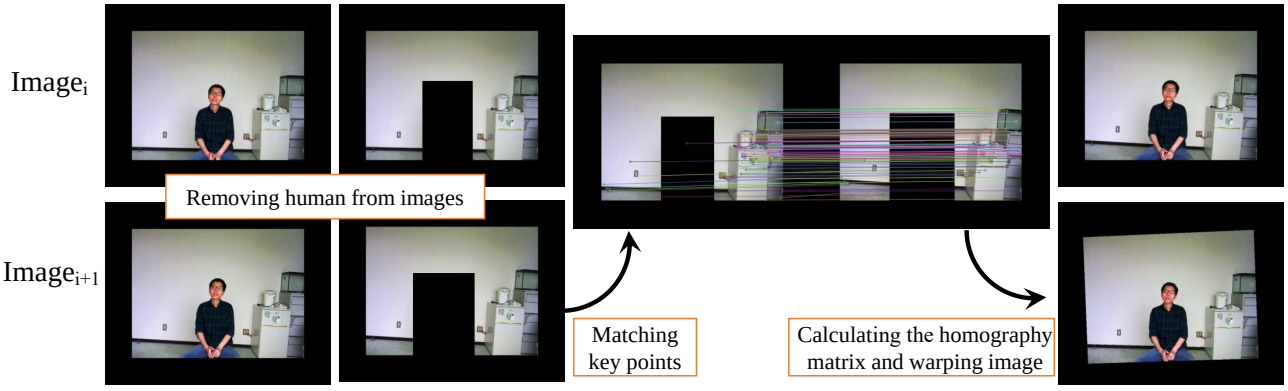

Figure 6: Warping the target image

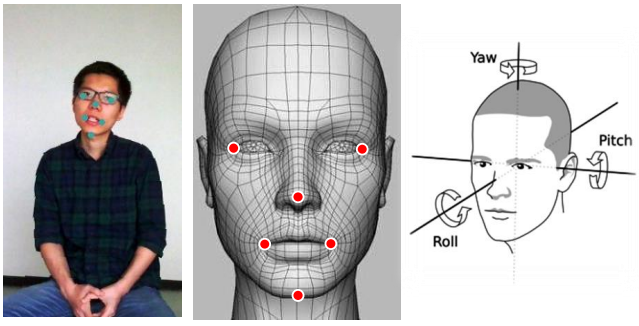

Figure 7: Facial key-points and head angles (the key points of the left image were detected from warped image using dlib; the middle image shows the default 3D key points; the right image illustrates the 3D head angles)

are denoted by r with two subscripts which represent the row and column index, respectively.

$$\begin{cases} \alpha = Atan(r_{32}/r_{33}) \\ \beta = Asin(-r_{31}) \\ \gamma = Atan(r_{21}/r_{11}) \end{cases}, \quad R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \quad (2)$$

where α, β, and γ denote the roll, pitch, and yaw angles, respectively. Then we calculated the Manhattan distance of two adjacent head angles to represent the head motion (HM) given by

$$HM_{i+1} = |\alpha_i - \alpha_{i+1}| + |\beta_i - \beta_{i+1}| + |\gamma_i - \gamma_{i+1}|, \quad (3)$$

where i and $i+1$ are two consecutive frames, and i is greater than or equal to zero. Note that the head angles with subscript i are calculated from the original image i, the image $i+1$ is the warped image with regard to the image i.

3.2.2. Gaze score

Social eye gaze played an important role in human-robot interaction [80]. Therefore, understanding the movements of the human gaze will contribute to enhancing human-robot engagement. The gaze score was calculated based on gaze direction. As the gaze direction and head pose are highly related to each other [81], we opted to calculate the gaze direction from the participant's head pose instead of analyzing movements of the eyes from the low-resolution images. Different from Fig. 6, the first image of each sentence was fixed as the reference image ($Image_i$), and the rest of the images of each sentence were warped to the reference image. All the head angles were calculated from the warped images. When the participant strictly faces the forehead camera of the robot, the roll, pitch, and yaw angles are 0°. The pitch and yaw angles fall within the closed interval of $[-\pi/4, +\pi/4]$. The gaze score describes the confidence in the fact that the participant is looking at the robot. As the gaze direction is highly related to the pitch and yaw angles, Eq. 4 shows how the gaze score of the frame i (i is greater than or equal to one.)

is calculated. As mentioned above, β and γ denote the pitch and yaw angle, respectively:

$$GS_i = 1 - \sqrt{\frac{\beta_i^2 + \gamma_i^2}{\beta_{max}^2 + \gamma_{max}^2}}, \tag{4}$$

where β_{max} and γ_{max} represent the maximum degree of the head pitch and yaw angle, respectively.

3.2.3. Body motion

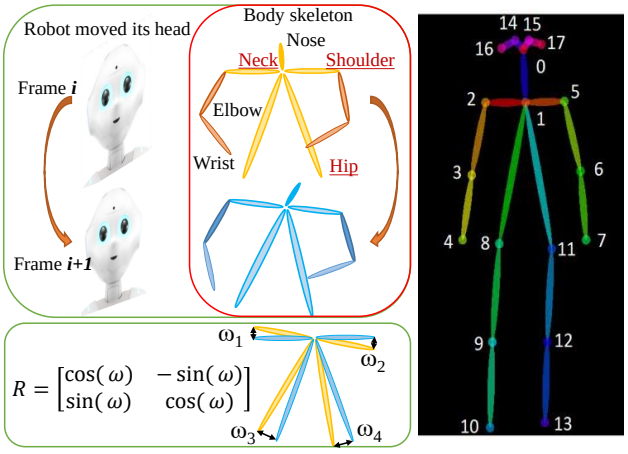

Figure 8: Adjusting the body pose of two successive images

The motion energy is acquired from a long period of time over the whole interaction. Computing motion energy with different pixels [38] of between images is not feasible when the images are blurry due to camera shake. The method of [82] was used to extract the skeleton of human body. We calculated body motion from two successive images, the original images ($Image_i$) and the warped image ($Image_{i+1}$) as shown in Fig. 6. With the neck as the center of rotation and the joints two shoulders, and two hips as the reference point, we calculated the angles $\omega_{1,2,3,4}$ to approximately compensate for the camera motion as shown in Fig. 8. The rotation angle ω is the mean of all the angles calculated from the angles mentioned above. Then, the second skeleton was rotated based on the rotation matrix in Fig. 8. Sometimes, the robot looked up and only the upper body could be captured by the camera. Therefore, the rotation angle was only calculated when it was possible to see the whole body in images. Finally, we took the neck as the center to overlap the skeletons of two frames to calculate the change of each joint.

Fig. 9 shows how to calculate the body motion from the overlapped skeleton. If the shoulders of two frames are overlapped, the triangle area constituted by two upper arms (from shoulder to elbow) in two consecutive frames i and $i+1$ from image sequence, can be calculated using the cross product (Eq. 5) of two vectors.

$$BM_{i+1}^{SE} = \frac{1}{2}\|\overrightarrow{SE}_i \times \overrightarrow{SE}_{i+1}\|, \tag{5}$$

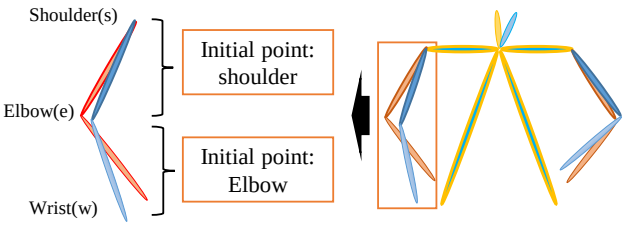

Figure 9: Example of calculating the upper arm motion

where BM_{i+1}^{SE} is body motion of the upper arm. SE is a vector which represents the upper arm from shoulder to elbow. And i is greater than or equal to zero. As the size of the human face will occupy different number of pixels according to the distance to the camera, the sum of all the triangle areas was standardized by dividing the size of the human face.

3.2.4. Pitch and energy

Pitch and energy are two well-known vocal features that are widely used in emotion recognition. Pitch, which is generated by the vibration of vocal cords, is perceived as $F0$ the fundamental voice frequency. Many different methods such as Simple Inverse Filter Tracking [83] (SIFT), Average Magnitude Difference Function (AMDF), and Auto-correlation Function [84] (ACF) were proposed to track the pitch. ACF, denoted by $acf_i(\tau)$, finds the second highest similarity between the signal and a series of shifted versions of itself, given by

$$acf_i(\tau) = \sum_{n=1}^{N-1-\tau} s_i(n)s_i(n+\tau), (0 \le \tau < N), \tag{6}$$

where $s(n)$ is the audio signal of the i-th frame, τ is the time delay, and N is the frame size.

Generally, the audio signal of each frame used to calculate voice pitch should contain more than two periods, and the pitch range of a human's voice is higher than $50Hz$. Given that an audio file with a sampling frequency is $16,000Hz$, we can calculate the range of the frame size N using Eq. 7:

$$\frac{16000}{50} \le \frac{N}{2}, N = 16000 \times T, \tag{7}$$

On the other hand, in Eq. 7, T is the time duration of the audio signal for one frame. Since the frame size N used in this study is 800, the time duration T is 50 millisecond. Finally, based on $acf_i(\tau)$, the pitch of the i-th frame can be calculated by Eq.8.

$$\begin{aligned} pp_i &= \arg\max_{\tau} \ (acf_i(\tau)), (20 \le \tau), \\ Pt_i &= 16000/\tau_i \end{aligned} \tag{8}$$

where pp_i is the second peak point of auto-correlation function $acf_i(\tau)$ in the i-th frame. We supposed that the highest pitch should be lower than $800Hz$. Therefore, τ is greater than or equal to 20. The sampling frequency $16,000Hz$ was divided by pp_i to calculate the voice pitch Pt_i of the i-th frame.

Now the average of the short-term energy can be calculated by the following equation:

$$En_i = \frac{1}{N}\sum_{n=1}^{N} s_i(n)^2, \tag{9}$$

where $s(n)$ is the audio signal of the i-th frame, and N is the frame size.

3.2.5. *Mel-frequency cepstral coefficient*

The frequency of the incoming sound can vibrate different spots of the human cochlea. Depending on the locations in the cochlea, different nerves were stimulated to inform the brain that some frequencies are present. Mel-Frequency Cepstral Coefficient (MFCC) was proposed based on this concept, since it is close to what humans hear actually. Then we investigate how these essential features affect people's perception of other people's personality traits. We used the method in [85] to calculate MFCC.

4. Feature fusion and classification models

In this section, we elaborate on Step 3 and Step 4 in Fig. 5 and the machine learning methods. We also compare our method to the baseline.

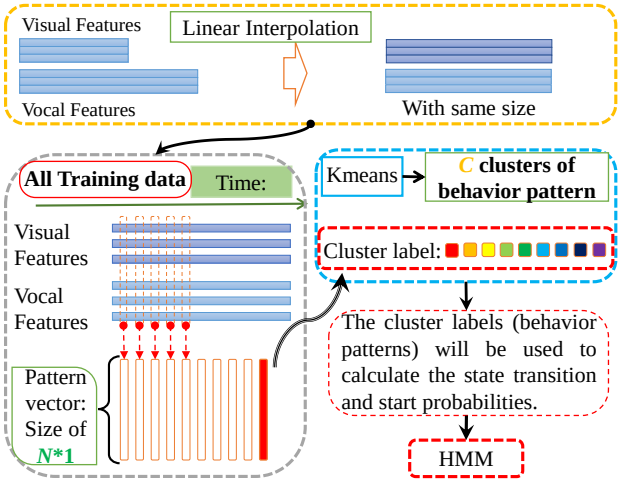

Figure 10: Linear interpolation and clustering behavior pattern

The visual and vocal features were extracted from each sentence as shown in Fig. 10. Due to the difference in sampling rate of the camera and microphone, we applied the linear interpolation to make visual and vocal features have the same length. We then have six nonverbal features composed of eighteen feature vectors defined in Table 4 (HM, GS, ME, Pt, En, and thirteen $MFCC$ feature vectors).

Testing all the combinations of eighteen feature vectors (the number of all the combinations is more than twenty thousand) is completely overwhelming. Therefore, all combinations were restricted to contain at most one MFCC feature vector. We used all eighteen features vectors as one combination for a simple comparison. In this study, 448 feature combinations (including the combination of all eighteen feature vectors) were tested. The parameter N in Fig. 10 indicates what features were used in a combination.

Once the combination of the features was decided, a feature matrix, each row of which represents a nonverbal feature, was generated. Each column of the feature matrix delineates patterns of behavior that were clustered by k-means [86]. In Fig. 10, the parameter C indicates the number of clusters or behavior patterns.

In order to determine the parameters of k-means, the relation of the total distances to the number of times that k-means was run with different centroid seeds (the abbreviation n_seeds was used to represent this parameter) was presented in Fig. 11. The results were acquired for eight clusters and all six nonverbal feature vectors, in which the first MFCC vector was used. Thirty thousand iterations for a single run was enough to make the clustering results converge to our dataset. In Fig. 11, the values shown in the vertical axis are in the hundred thousandths decimal place of the sum of distances. It can be seen that the sum of distances was minimized when n_seeds is larger than 360. Therefore, n_seeds was set to 400 in our study.

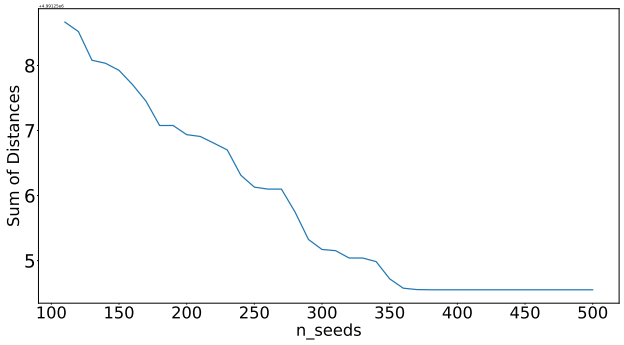

Figure 11: Relation of the total distances to the number of times that k-means was run with different centroid seeds

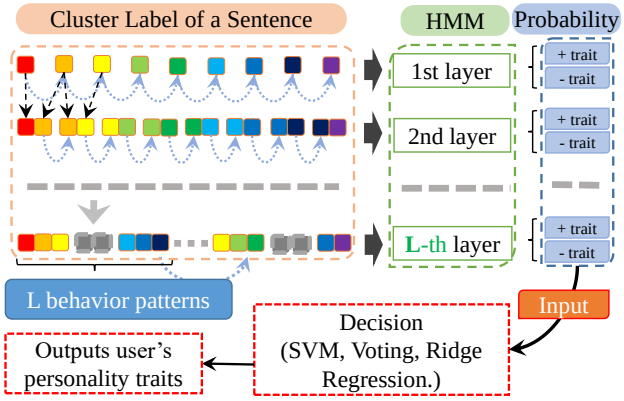

Figure 12: Approach to generating multiple layers of HMM and making decision

Based on the cluster labels, each sentence can be considered as the transition of a sequence of observable behavior patterns. Considering human behaviors in reality, the duration of each behavior varies. Therefore, we tried to combine

two or more successive behavior patterns to generate new transition sequences as shown in Fig. 12. Each sentence can generate several new transition sequences in which a state is a combination of up to L behavior patterns. In order to avoid the appearance of the isolated behavior pattern at the end of the transition sequence, the combined behavior patterns were slid with a step length of one behavior pattern. All the training data were divided into two parts, sentences of which the personality trait is positive or negative. In Fig. 12, $+trait$ is the prediction score or probability that the personality is high on this trait. $-trait$ is the prediction score or probability that the personality is low on this trait. We generated two dictionaries that contain all the state transition probabilities, and two dictionaries that contain the start probabilities of the sentences, both for binary personality traits (high versus low).

With the increase in the number of clusters, C and the number of combined behavior patterns L, the categories of transition states would increase dramatically. Consequently, some states would only exist in a positive or negative personality trait. The transition and start probabilities of these states were appended to the opposite dictionaries with a minimum probability. In testing, probabilities of the states that only existed in the testing data were assigned 1.

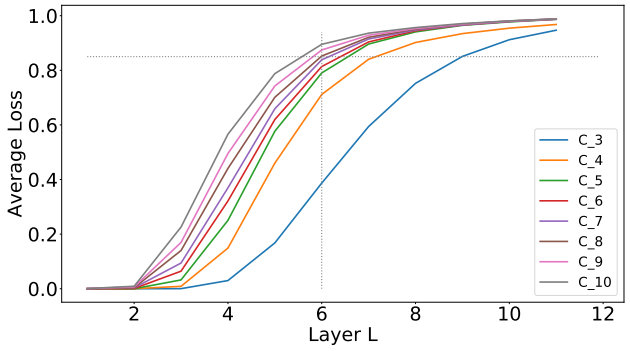

Figure 13: Relation of average loss to layers and clusters

We defined the average loss, which was calculated by averaging the ratio of the number of the appended states to the number of the states in stock, to show the relationship between the appended states and parameter C and L. In Fig. 13, the horizontal axis is the number of layers L ranging from 1 to 11. The vertical axis is the average loss. The number of clusters C was tested from 3 to 10. The results in Fig. 13 were obtained by a combination of all six feature vectors, in which the first MFCC vector was used, in terms of extroversion trait. There are no rigid requirements for the parameter C and L. In our study, the range of the parameter C is from three to eight, and the maximum L is six. Therefore, 63 combinations of the outputs of different layers were tested.

In the testing phase, each layer can provide two probabilities of the personality trait. In light of the previous work that predicted the leadership style [69], we adopted to use the same methods in our study. Therefore, voting, SVM, and Ridge Regression were used to classify the participants' personality traits. Voting method is a relatively easy for making a decision. The personality trait was considered as positive when the majority of the higher probabilities is $+trait$.

The formula of SVM [87] is given in Eq. 10.

$$y(x) = \sum_{m=1}^{M} a_m y_m \mathcal{K}(x, x_m) + b, \tag{10}$$

where $y(x)$ is the predicted label of the sample x. The data x_m and the corresponding label y_m were used to train a set of optimal Lagrange multipliers a_m. $\mathcal{K}(x, x_m)$ is the kernel function. We tested three different kernel functions: linear, RBF (radial basis function), and sigmoid given by

$$\mathcal{K}(x_i, x_j) = \begin{cases} x_i^T x_j\,, & Linear \\ e^{-\lambda \|x_i - x_j\|^2}\,, & RBF \\ tanh(\lambda x_i^T x_j)\,, & Sigmoid \end{cases} \tag{11}$$

where x_i and x_j are two data samples, and λ was chosen from [0.01, 0.05, 0.1, 0.5, 1, 5]. For training each SVM, the penalty parameter of the error term was chosen from [0.4, 1, 1.6, 2.2, 2.8, 3.4, 4].

While training the ridge regression, the inputs are the probability, the predicted value is the averaged personality trait score ranging from 1 to 5. The regression parameters were optimized by cross-validation methods. The regression parameters can be calculated by the following equation:

$$\omega = (X^T X + \gamma I)^{-1} X^T Y, \tag{12}$$

where X is the probability, I is an identity matrix, Y is the personality traits score, and γ is the ridge parameter defined by

$$\gamma = e^{0.5i-10} (i \in [0, 32], i \in \mathbb{N}). \tag{13}$$

We recruited 21 participants and each participant asked the robot about 10 to 20 questions. In total, 329 sentences of participants were collected. These sentences were used as training samples. The performance of voting, SVM, and ridge regression was evaluated by using the leave-one-out method, which means that every time one sentence was used for testing, and the rest were used for training.

In view of previous studies [35, 43], the statistical information such as mean, maximum, minimum, standard deviation, and variance of each nonverbal features can be easily used to classify the personality traits. On the other hand, zero-padding is also very popular in the field of signal processing [88]. Therefore, we padded zero to the end of each raw form nonverbal feature to separately generate the visual and vocal features with equal length. Moreover, different combinations of statistical features and zero-padded features were concatenated and tested. The same classification methods described above were applied to evaluate these two features. The feature combinations that yielded the best result of each trait were used as the baseline.

5. Experimental results and analysis

In this section, we presented the classification results and the comparison to the baseline. The results of the controlled experiment were also presented, where the visual features were extracted without compensating for the robot's camera motion.

5.1. Classification results

The mean score of personality traits of all participants was used as a cutoff point when we analyze the performance of the ridge regression classifier. The results of single features and combined features were presented separately.

5.1.1. Classification results of single features

The accuracy of every single feature for inferring five personality traits was analyzed. As the figures were too large to fit on one page, only the accuracy of every single feature for inferring extroversion was presented in Fig. 14. Each row represents different layers defined in Fig. 12, where each column shows a different classifier. In each sub-figure, the vertical axis indicates the classification accuracy and the horizontal axis shows the number of clusters to determine different behavior patterns defined in Fig. 10. The result of every single feature is distinguished by different colored solid or dashed lines. This part reports on the following findings obtained:

1) Increasing the number of hidden layers is helpful for achieving a higher accuracy. However, if the number of layers increased to a substantially large value, the accuracy will decrease;
2) With the increase of the number of layers, the number of clusters should be decreased, and vice versa. Increasing the number of clusters is helpful when the number of layers is small;
3) The less influential features can be filtered out with the increase of layers.

As shown in Fig. 14, the accuracy of SVM with RBF kernel is the lowest compared to the other four methods. And En apparently is the best feature for inferring Extroversion. It is also obvious according to the results of Openness and Emotional Stability. Both GS and En are good at inferring Openness. HM, GS, ME, and En are good at inferring Emotional Stability when using SVM with three different kernels. In the ridge regression and voting, ME outperforms the other features in inferring Emotional Stability.

However, according to the accuracy of the single feature for inferring Conscientiousness and Agreeableness, the aforementioned findings were not as clear as in the other three traits. It was found that the other four classification methods did not provide high accuracy, except for the sigmoid kernel SVM. On the other hand, it is also difficult to draw any conclusions about which single features we have filtered out by increasing the number of layers. Although ME provided an extremely high accuracy by the sigmoid kernel SVM in inferring conscientiousness, we hardly observe any patterns. A similar situation appears to En by the sigmoid kernel SVM and $MFCC_2$ by voting that provide higher accuracy for inferring agreeableness, without showing any notable patterns. When we review the three findings mentioned above, we realized that increasing the number of layers or clusters also increases the number of behavior patterns (Figs. 10 and 12) and the average loss (Fig. 13). In other words, less information of each sentence remained useful with the increase in the number of layers or clusters. Therefore, it causes a decrease in accuracy when the number of layers or clusters increases. On the other hand, increasing the diversity of behavior patterns properly improves the classification accuracy, which is in accordance with the findings 1 and 2. In layer 1, the results of the influential features are bad. As the behavior patterns in the first layer are independent of each other, some of which could be deceptive. However, while the number of layers was increased, some deceptive patterns could be removed by incorporating successive patterns. We believe that the features that provided high accuracy match the personality trait well. Therefore, increasing the number of layers or clusters has less effect on the classification performance of these features. Thus, the finding 3 can be explained.

5.1.2. Classification results of combined features and baseline comparison

The highest accuracies of each method were presented in Table 5, where C denotes the number of clusters, F the index of feature combinations, and L the index of layer combinations, respectively. The results of all_18 were acquired by combining all eighteen nonverbal features. The best results were presented in the second row. The third row is the results of the controlled experiments, where the same feature and layer combinations without camera motion compensation. The best results of each personality trait were shown in bold. The italic figure indicates the cases that the accuracy of the controlled experiment is higher than that of our proposed method. The details of the feature and layer combination of the best results were presented in Table 6. The last column of Table 5 shows the baseline results, where 0_{pad} denotes the results of the zero-padding features, and Sta the results of statistical features. The training methods were omitted here due to the space limitations.

We found that the results of extroversion, openness, and emotional stability in Table 5 are highly correlated with the results of single features. As we mentioned above, the sigmoid kernel SVM did not provide accurate results as to extroversion and openness, which is in accordance with the results in Table 5. Likewise, the results of emotional stability provided by five different methods are pretty similar in Table 5. The results of single features in inferring emotional stability are also pretty similar. In [89], the authors statistically analyzed the correlations between nonverbal patterns and personality traits self-report questionnaire, where *Eye contact* and *Raise voice* are considered as basically the same as our proposed features GS and En, and personality traits. In [90], the author not only summarized research on relationships between nonverbal cue and personality traits from the

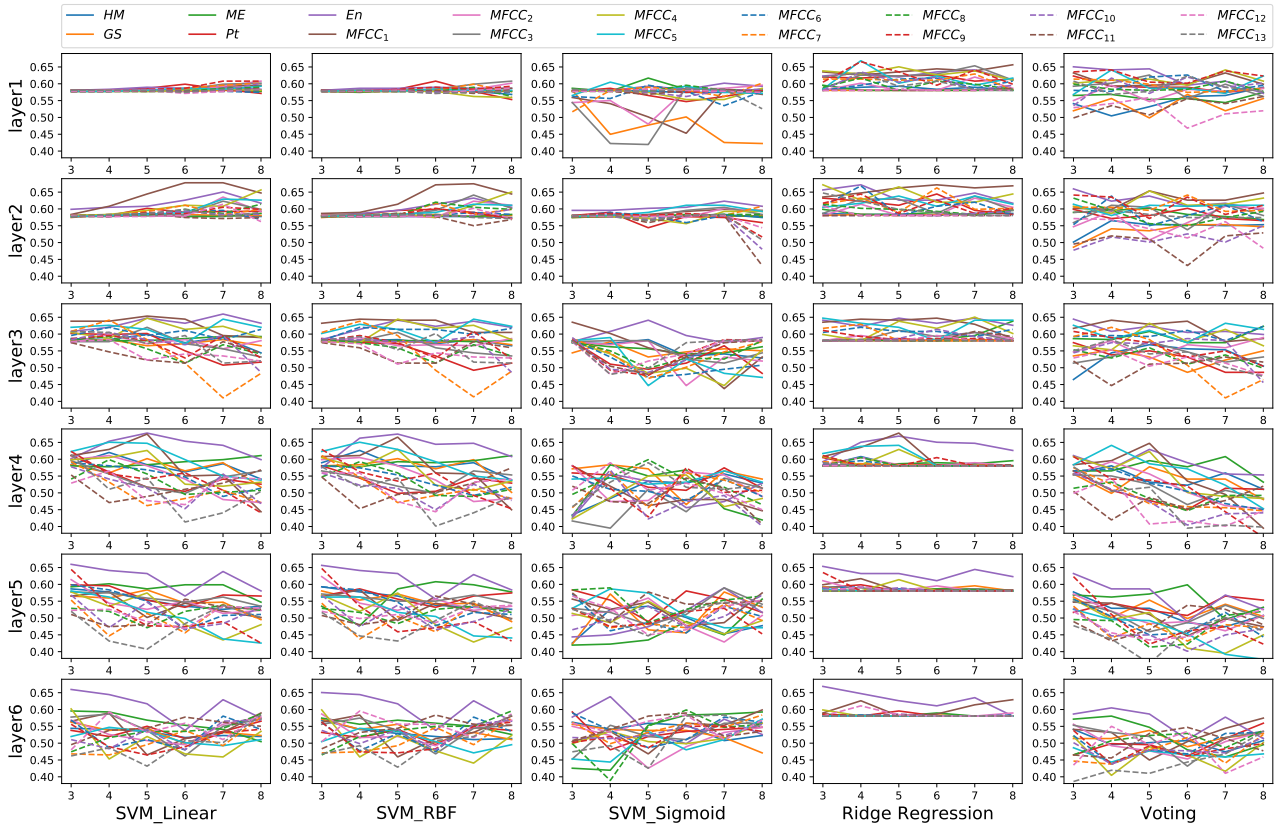

Figure 14: Accuracy of each single feature for inferring Extroversion

Table 5
Highest accuracies for Big Five Personality Traits with different feature combinations and parameters VS best of baseline

Personality Trait	SVM			Ridge Regression	Voting	all_18	Baseline	
	Linear	RBF	Sigmoid				0_{pad}	Sta
Extroversion	C7F441L2	C7F431L2	C3F193L9	C4F133L2	C4F142L7	C7L2_Ridge	F133	
	0.7508	0.7477	0.7173	**0.7629**	0.7538	0.6869	0.7325	0.6748
	0.7568	0.7447	0.6049	0.7629	0.7508	0.6717	0.7325	0.6748
Openness	C7F364L7	C7F375L2	C3F202L26	C7F411L1	C5F370L7	C8L23_Voting	F364	
	0.8237	0.8146	0.7690	0.8146	0.8207	0.6687	0.7112	0.7781
	0.8024	0.8055	0.5502	0.7994	0.8024	*0.6717*	0.7264	0.7781
Emotional Stability	C7F379L2	C3F311L3	C4F172L12	C4F142L11	C7F26L1	C8L1_RBF	F142	
	0.7872	0.7842	0.7751	**0.7994**	0.7660	0.7568	0.7599	0.7477
	0.7325	0.7447	0.7325	0.7964	0.7204	0.7325	0.7812	0.7416
Conscientiousness	C7F354L7	C5F238L12	C3F3L8	C5F244L11	C7F362L2	C4L41_RBF	F3	
	0.7173	0.6930	**0.9149**	0.7052	0.7021	0.6353	0.6383	0.6109
	0.6109	0.6748	0.7690	0.6474	0.6383	*0.6444*	0.5805	0.5562
Agreeableness	C8F425L7	C8F425L7	C3F302L7	C8F425L7	C7F82L23	C3L1_Sigmoid	F302	
	0.6960	0.6778	**0.9210**	0.6900	0.7325	0.7964	0.6444	0.5532
	0.6960	0.6687	0.5532	*0.6960*	0.6049	0.7447	0.6231	0.5623

self-report, which was named cue validity, but also the evaluation of external observers, which was named cue utilization. These works support our results, which will be detailed below.

En provided the best results when increasing the number of layers for inferring extroversion with all four methods, except for the sigmoid kernel SVM. The results provided by the linear and RBF kernel SVM, ridge regression, and voting method in Table 5 achieved high accuracies by the feature combinations that include *En*. As mentioned in [89], *"High levels on the extroversion scale will correlate with a high tendency to raise the voice to emphasize something"*. Similarly, [90] showed that some studies supported that *loudness of voice* affects both cue validity and utilization. In

brief, the observers used *loudness of voice* to infer the co-communicator's extroversion, and extroversion also affects *loudness of voice*.

The same situation emerged in openness. The results with the RBF kernel SVM are relatively poor compared to the other four methods. Moreover, GS and En are the best features in inferring openness. These two points were supported by the results in Table 5. Similarly, the correlation analysis in [89] suggested that *individuals scoring high on the openness scale also might look back at the co-communicator while being in a conversation*. On the other hand, there is a somewhat weak correlation between *Raise voice* and openness. However, it was found that *individuals that score high on the openness scale feel comfortable when others raise their voices*. It could be conjectured that people who scored high on openness would tend to raise their voices to inspire the co-communicator to raise their voices. In [90], there was only one study showing that *loudness of voice* has effects for both utilization and validity. *eye contact* showed less obvious effects on openness. However, it is opposite of the observer viewpoint.

Table 5 shows that the feature combination with the highest accuracy of the emotional stability by the linear kernel SVM consists of HM, GS, ME, En, and $MFCC_{11}$, therein, single feature HM, GS, ME, and En also yielded good results. Similarly, based on the results of single features on inferring emotional stability, GS, ME, and En by the RBF kernel SVM, GS and En by the sigmoid kernel SVM, ME by ridge regression and voting are in accordance with the results in Table 5. In [89], they revealed that neuroticism, which is contrary to emotional stability, is highly associated with *Eye contact* and *Raise voice*. Their investigation result is in line with our results obtained by SVM with three kernels. Our results also revealed that the body motion ME is somehow highly related to emotion stability. [90] described negative aspects between *head movements* and neuroticism with regard to cue validity, and positive aspects with regard to cue utilization. *Loudness of voice* showed effects on both validity and utilization. *eye contact* showed less obvious effects on cue validity. The effects on cue utilization of *eye contact* are clear. The effects of *body movement* on emotional stability in terms of both validity and utilization are not obvious.

The highest classification accuracy for conscientiousness was obtained by ME, on the condition that the number of clusters is 3 and the number of layers is 1 or 3, which is in line with the highest accuracy obtained by the sigmoid kernel SVM. In agreeableness, except for the feature combination that yielded the highest accuracy by the sigmoid kernel SVM, it is the same feature combination used in the linear and RBF kernel SVM, and ridge regression. All the combinations of $F425$, $F301$, and $F82$ contain En. This is partially supported by the investigation of [89] [90], where individuals that score high on agreeableness do not raise their voices to emphasize something and also showed the effects of *head movements*, *eye contact*, and *body movement* in terms of cue utilization on both conscientiousness and agreeableness.

Table 6
A part of feature combinations and layer combinations

	Feature Combination	
F	3	$[ME]$
	26	$[HM, ME, Pt]$
	82	$[HM, GS, En, MFCC_2]$
	133	$[En, MFCC_4]$
	142	$[ME, En, MFCC_4]$
	172	$[GS, En, MFCC_5]$
	193	$[HM, MFCC_6]$
	202	$[GS, ME, MFCC_6]$
	238	$[ME, En, MFCC_7]$
	244	$[HM, ME, En, MFCC_7]$
	302	$[ME, En, MFCC_9]$
	311	$[GS, ME, En, MFCC_9]$
	354	$[GS, MFCC_{11}]$
	362	$[GS, ME, MFCC_{11}]$
	364	$[GS, En, MFCC_{11}]$
	370	$[HM, GS, En, MFCC_{11}]$
	375	$[GS, ME, En, MFCC_{11}]$
	379	$[HM, GS, ME, En, MFCC_{11}]$
	411	$[HM, GS, ME, En, MFCC_{12}]$
	425	$[HM, En, MFCC_{13}]$
	431	$[Pt, En, MFCC_{13}]$
	441	$[ME, Pt, En, MFCC_{13}]$
	Layer Combination	
L	1	$[1st]$
	2	$[2nd]$
	3	$[3rd]$
	7	$[1st, 2nd]$
	8	$[1st, 3rd]$
	9	$[1st, 4th]$
	11	$[1st, 6th]$
	12	$[2nd, 3rd]$
	14	$[2nd, 5th]$
	23	$[1st, 2nd, 4th]$
	26	$[1st, 3rd, 4th]$

Referring to [89] and [90], the results of our experiments are supported by social science research. We also noticed that $MFCC$ contributed significantly to improving the classification accuracy. However, the relationship between $MFCC$ and personality trait estimation needs to be further investigated with a specific experimental design and setup. Table 5 also showed that most results of the feature combinations that contain visual features with camera motion compensation are better than without camera motion compensation. Excepts $F441$ for inferring extroversion and $F425$ for inferring agreeableness, the results of visual features without motion compensation are slightly higher than or equal to the results of visual features with motion compensation. $F133$ and $F431$ are all vocal features, therefore, their results are the same.

Table 7 showed the best results for each personality traits that acquired by single visual features. The results of the visual feature with camera motion compensation were presented in the first row, those without camera motion compensation were given in the second row. It can be noted that

Table 7
Average accuracy for Big Five Personality Traits with visual nonverbal features

Personality Trait	Extroversion	Openness	Emotional Stability	Conscientiousness	Agreeableness
HM	**0.6565**	0.6991	**0.7356**	0.6444	0.6353
	0.6261	**0.7082**	0.7264	**0.6474**	**0.6565**
GS	0.6201	0.7508	**0.7508**	**0.6869**	0.6322
	0.6444	**0.7538**	0.7416	0.6778	**0.7143**
ME	0.6778	0.6505	0.7173	**0.9149**	**0.6748**
	0.7143	**0.7173**	**0.7234**	0.769	0.6109

Table 8
MSE and R^2 scores of Extroversion and Emotional-Stability

Personality Trait	Extroversion	Emotional Stability
MSE	0.248	0.389
R^2	0.024	0.196

the visual feature with camera motion compensation did not always provide better results. However, the results of combining visual features with motion compensation with vocal features were better as shown in Table 5. It was understood that individuals' voices did not match their visual nonverbal behaviors, if the visual features were extracted without compensating for camera motion. On the other hand, combined features can provide better results than single features and all features all_18, comparing Table 5 and 7.

Moreover, compared to the baseline, our proposed feature fusion method outperformed the baseline method.

5.2. Regression evaluation

It was conjectured that using the probabilities to calculate the regression of personality traits does not have any explicit physical meanings. However, based on Table 5, the classification results of ridge regression of Extroversion and Emotional-Stability were surprisingly good. We calculated the Mean Squared Error (MSE) values and coefficient of determination (R^2) to evaluate the ridge regression of the Extroversion and Emotional Stability. MSE was calculated using the regression results of the group of parameters that provided the highest classification accuracy. Referring to Table 5, the regression results of our proposed methods for inferring Extroversion ($C4F133L2$) and Emotional Stability ($C4F142L11$) were used.

R^2 was calculated based on the following equation:

$$R^2 = 1 - \frac{\sum_{i=1}^{S} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{S} (Y_i - \overline{Y}_i)^2}, \tag{14}$$

where S is the total number of the samples, Y_i is the mean score of the personality trait of the sample i, $\hat{Y}_i$ is the regression score of the personality trait of the sample i, and $\overline{Y}_i$ is the average score of the trait. Note that since the R^2 score is relatively small, the results of regression model did not fit the data perfectly.

However, the classification accuracies of the ridge regression on extroversion and emotional stability were the highest compared to other classifiers. Therefore, we took the regression result of extroversion as an example of why the R^2 score is small. The scatter plot of extroversion is shown in Fig. 15, where the orange dots are the ground-truth label and the blue dots are the prediction scores, respectively. The orange solid line is the mean score of all participants, and the blue solid line is the mean score of all the prediction scores. It can be seen that the prediction scores on extroversion are distributed around the mean score.

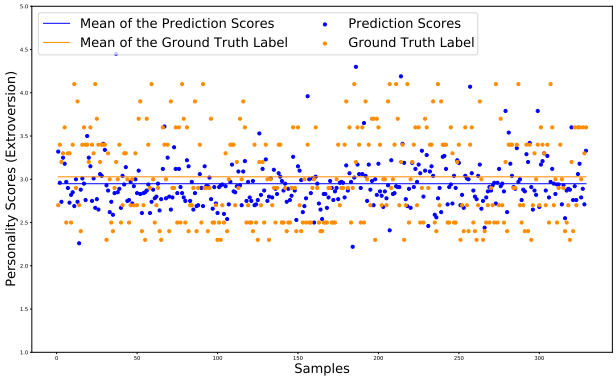

Figure 15: Relation of average loss to layers and clusters

5.3. Classification results by optimizing hyper-parameters using training data

In the previous subsection, the classification results were acquired based on the testing data. The parameters (combination of features, number of clusters, and combination of layers) were fixed in the beginning. In the following, the parameters were considered as the hyperparameters in the learning phase. The procedure for training the model was explained in the pseudo-code in Algorithm 1. In brief, all the samples were divided into three parts: one test sample, 20% validation data, and 80% training data. The classifier will be trained with the training data according to different parameter combinations (different combinations of features, different number of clusters, and different combinations of layers). The parameters providing the highest classification accuracy on the validation data would be recorded to test the testing data. Finally, the final accuracy on the testing data was presented, as well as the parameters that provided the highest classification accuracy on validation data.

Algorithm 1: Training with Hyper-parameters

Input: Nonverbal features: X;
The corresponding personality trait labels: Y;
Number of samples: N
Output: Accuracy of test data: Acc;
Number of time that the parameter was used: Par_usage

for $i = 1$ *to* N **do**
  # Leave-one-out;
  $Test\_x, Test\_y = X_i, Y_i$;
  $Train\_data, Train\_label = X_{(not\ i)}, Y_{(not\ i)}$;
  for $j = 1$ *to* 5 **do**
    # 5-folder cross validation;
    $Vali_x = Train\_data_{(1/5)}$;
    $Vali_y = Train\_label_{(1/5)}$;
    $Train_x = Train\_data_{(4/5)}$;
    $Train_y = Train\_label_{(4/5)}$;
    initialize validation accuracy: $Vali_{acc}$;
    for F *in* $Feature\ combinations$ **do**
      for C *in* $Number\ of\ clusters$ **do**
        for L *in* $Layer\ combinations$ **do**
          Training classifier with $Train_x, Train_y$;
          Classifier: $Clf_{(F,C,L)}$;
          Testing by using the validation data $Vali_x, Vali_y$;
          Update $Vali_{acc}$;
  $F, C, L = argmax(Vali_{acc})$;
  Update $Par\_usage \leftarrow F, C, L$;
  Predicted label: $Pred_y = Clf_{(F,C,L)}(Test\_x)$;
Compute Acc by $Pred_y$ and $Test_y$;
return Acc, Par_usage;

The classifiers of SVM with RBF kernel and the sigmoid kernel were not on a par with the linear SVM classifier. Therefore, only the results of linear SVM, ridge regression, and voting classifiers were presented in Table 9. The highest classification accuracy on each trait also was highlighted in bold. The classification accuracies on extroversion, openness, and emotional stability in Table 9 were not as high as the classification accuracies in Table 5. However, the differences of the classification accuracies on conscientiousness and agreeableness are notable between Table 9 and Table 5.

During the training and testing, we analyzed the number of times the above-mentioned parameters were used. Instead of counting the combination of the features or the combination of the layer, we counted the number of times that each single feature was used. For instance, if the feature combinations of $[HM, GS, En]$ and $[ME, En]$ were used, En would be counted twice. Fig. 16 showed the number of times that the nonverbal feature was used by each classifier on extroversion.

Table 9
Maximum accuracies for Big Five Personality Traits

Personality Trait	SVM Linear	Ridge Regression	Voting
Extroversion	**0.7356**	0.6930	0.6930
Openness	**0.7872**	**0.7872**	0.7568
Emotional Stability	0.7568	**0.7629**	0.7203
Conscientiousness	**0.6383**	0.6018	0.6292
Agreeableness	0.5957	0.6292	**0.6444**

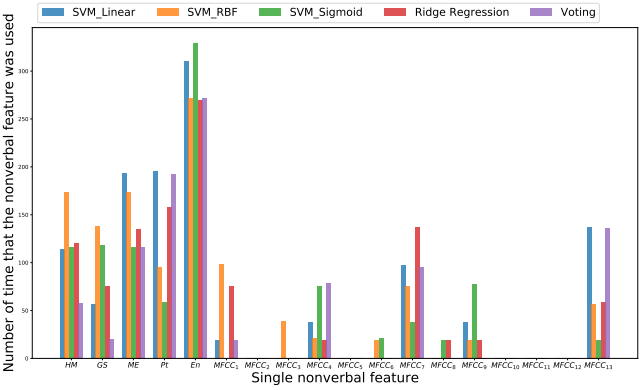

Figure 16: Number of time that the nonverbal feature was used by each classifier on extroversion

Fig. 16 showed that the En was used most frequently, which is in line with the previous finding. The same method was applied to analyze the other four personality traits. Specifically, it was observed that GS and En were frequently used for classifying openness. GS, En, and $MFCC_4$ were frequently used for classifying emotional stability. The patterns on conscientiousness and agreeableness were not as clear as the other three traits. Following the analysis, we found taht the En was used most frequently on all five traits.

6. Conclusions and future work

Several important issues of understanding human personality traits in social human-robot interaction have been addressed based on our experiments involving human participants. A new algorithmic framework was proposed to deal with the robot (or camera) posture change and multi-modal feature fusion problem toward improving the human personality traits classification accuracy. It was demonstrated that selecting the right set of multi-modal features can improve the performance of inferring human personality traits. Our model is able to deal with the data with variable lengths. Notably, visual features that were extracted with camera motion compensation could not always provide good results. Once these visual features were combined with vocal features, their results outperformed the same combinations in which the visual features were extracted without camera motion compensation.

The multi-layer HMM model in our framework showed some interesting phenomena. It can be used to filter out less

influential features, by which we can fuse some features with purpose. The relationships between nonverbal cues and extroversion, openness, and emotional stability were clearer and more straightforward than the relationships between nonverbal cues and conscientiousness and agreeableness. Recent social science studies showed many evidences that supported our findings.

As future work, we will focus on two main aspects: feature extraction and model improvement. Currently, our visual nonverbal features mainly describe the magnitude of the movements. Inspired from the methods applied to social science, we designed some methods for extracting describable nonverbal features or cues. A human can interact with a robot while standing and approaching it, or sitting. The nonverbal cues such as *closed arms*, *self-touch*, and *facial expression* will be extracted and used to analyze human personality traits. On the other hand, the number of combined successive behavior patterns was fixed. The system will be extended to include a varying number of combined successive behavior patterns. It is also well known that the personality traits will likely become apparent over time. Therefore, robots need to update their impression of personality traits whenever they are interacting with humans in an incremental fashion.

Acknowledgments

The authors are grateful for financial support from the Air Force Office of Scientific Research under AFOSR-AOARD/ FA2386-19-1-4015 and the Shibuya Science, Culture, and Sports Foundation 2019 Grant Program.

References

[1] G. W. Allport, Personality: A Psychological Interpretation. Holt (1937).

[2] A. Vinciarelli, G. Mohammadi, A survey of personality computing, 2014. doi: 10.1109/TAFFC.2014.2330816.

[3] G. M. Hurtz, J. J. Donovan, Personality and job performance: The big five revisited, 2000. doi: 10.1037/0021-9010.85.6.869.

[4] R. Mõttus, G. McNeill, X. Jia, L. C. Craig, J. M. Starr, I. J. Deary, The associations between personality, diet and body mass index in older people, Health Psychology 32 (2013) 353–360. URL: http://www.ncbi.nlm.nih.gov/pubmed/21928903. doi: 10.1037/a0025537.

[5] D. J. Ozer, V. Benet-Martínez, Personality and the Prediction of Consequential Outcomes, Annual Review of Psychology 57 (2006) 401–421. doi: 10.1146/annurev.psych.57.102904.190127.

[6] D. C. Funder, The Personality Puzzle (6th ed.). New York, NY: Norton., 2013.

[7] J. S. Uleman, S. Adil Saribay, C. M. Gonzalez, Spontaneous Inferences, Implicit Impressions, and Implicit Theories, Annual Review of Psychology 59 (2008) 329–360. doi: 10.1146/annurev.psych.59.103006.093707.

[8] J. Willis, A. Todorov, First impressions: Making up your mind after a 100-ms exposure to a face, Psychological Science 17 (2006) 592–598. doi: 10.1111/j.1467-9280.2006.01750.x.

[9] C. Y. Olivola, A. Todorov, Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences, Journal of Experimental Social Psychology 46 (2010) 315–324. doi: 10.1016/j.jesp.2009.12.002.

[10] L. P. Satchell, From photograph to face-to-face: Brief interactions change person and personality judgments, Journal of Experimental Social Psychology 82 (2019) 266–276. doi: 10.1016/j.jesp.2019.02.010.

[11] D. Levy, Love and Sex with Robots., HarperCollins Publishers. New York, 2009.

[12] R. Richards, C. Coss, J. Quinn, Exploration of relational factors and the likelihood of a sexual robotic experience, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 10237 LNAI, Springer Verlag, 2017, pp. 97–103. doi: 10.1007/978-3-319-57738-8_9.

[13] A. Rossi, K. Dautenhahn, K. L. Koay, M. L. Walters, The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario, Paladyn 9 (2018) 137–154. doi: 10.1515/pjbr-2018-0010.

[14] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, M. Schröder, Bridging the gap between social animal and unsocial machine: A survey of social signal processing, 2012. URL: http://ieeexplore.ieee.org/document/5989788/. doi: 10.1109/T-AFFC.2011.27.

[15] T. Minato, M. Shimada, H. Ishiguro, S. Itakura, Development of an android robot for studying human-robot interaction, in: Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science), volume 3029, 2004, pp. 424–434. doi: 10.1007/978-3-540-24677-0_44.

[16] V. Ng-Thow-Hing, P. Luo, S. Okita, Synchronized gesture and speech production for humanoid robots, 2010, pp. 4617–4624. doi: 10.1109/IROS.2010.5654322.

[17] B. Bruno, N. Y. Chong, H. Kamide, S. Kanoria, J. Lee, Y. Lim, A. K. Pandey, C. Papadopoulos, I. Papadopoulos, F. Pecora, A. Saffiotti, A. Sgorbissa, Paving the way for culturally competent robots: A position paper, in: RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication, volume 2017-Janua, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 553–560. doi: 10.1109/ROMAN.2017.8172357. arXiv:1803.08812.

[18] N. T. Viet Tuyen, S. Jeong, N. Y. Chong, Emotional Bodily Expressions for Culturally Competent Robots through Long Term Human-Robot Interaction, in: IEEE International Conference on Intelligent Robots and Systems, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 2008–2013. doi: 10.1109/IROS.2018.8593974.

[19] A. Aly, A. Tapus, A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction, ACM/IEEE International Conference on Human-Robot Interaction (2013) 325–332. . doi: 10.1109/HRI.2013.6483606.

[20] C. Nass, K. M. Lee, Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction, Journal of Experimental Psychology: Applied 7 (2001) 171–181. doi: 10.1037/1076-898X.7.3.171.

[21] S. Woods, K. Dautenhahn, C. Kaouri, R. te Boekhorst, K. L. Koay, M. L. Walters, Are robots like people?: Relationships between participant and robot personality traits in human–robot interaction studies, Interaction StudiesInteraction Studies Social Behaviour and Communication in Biological and Artificial Systems 8 (2007) 281–305. doi: 10.1075/is.8.2.06woo.

[22] A. Tapus, M. J. Mataric, Socially assistive robots: The link between personality, empathy, physiological signals, and task performance, in: AAAI Spring Symposium - Technical Report, volume SS-08-04, 2008, pp. 133–140. URL: https://aitopics.org/doc/conferences:58A799BA/.

[23] E. Park, D. Jin, A. P. Del Pobil, The law of attraction in human-robot interaction, International Journal of Advanced Robotic Systems 9 (2012). doi: 10.5772/50228.

[24] D. Byrne, W. Griffitt, Similarity and awareness of similarity of personality characteristics as determinants of attraction, Journal of Experimental Research in Personality 3 (1969) 179–186.

[25] K. M. Lee, W. Peng, S. A. Jin, C. Yan, Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human-robot interaction, Journal of Communication 56 (2006) 754–772. URL: https://academic.oup.com/joc/art

icle/56/4/754-772/4102572. doi: 10.1111/j.1460-2466.2006.00318.x.

[26] K. Isbister, C. Nass, Consistency of personality in interactive characters: Verbal cues, non-verbal cues, and user characteristics, International Journal of Human Computer Studies 53 (2000) 251–267. doi: 10.1006/ijhc.2000.0368.

[27] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, M. Chetouani, Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions, IEEE Access 5 (2017) 705–721. doi: 10.1109/ACCESS.2016.2614525.

[28] A. Aly, A. Tapus, Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human–robot interaction, Autonomous Robots 40 (2016) 193–209. doi: 10.1007/s10514-015-9444-1.

[29] T. Santamaria, D. Nathan-Roberts, Personality measurement and design in human-robot interaction: A systematic and critical review, in: Proceedings of the Human Factors and Ergonomics Society, volume 2017-Octob, Human Factors an Ergonomics Society Inc, 2017, pp. 853–857. doi: 10.1177/1541931213601686.

[30] S. L. Müller, A. Richert, The big-five personality dimensions and attitudes to-wards robots: A cross sectional study, in: ACM International Conference Proceeding Series, Association for Computing Machinery, New York, New York, USA, 2018, pp. 405–408. URL: http://dl.acm.org/citation.cfm?doid=3197768.3203178. doi: 10.1145/3197768.3203178.

[31] U. Morsunbul, Human-robot interaction: How do personality traits affect attitudes towards robot?, Journal of Human Sciences 16 (2019) 499–504. doi: 10.14687//jhs.v16i2.5636.

[32] M. R. Barrick, M. K. Mount, the Big Five Personality Dimensions and Job Performance: a Meta-Analysis, Personnel Psychology 44 (1991) 1–26. URL: http://doi.wiley.com/10.1111/j.1744-6570.1991.tb00688.x. doi: 10.1111/j.1744-6570.1991.tb00688.x.

[33] T. Yarkoni, Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers, Journal of Research in Personality 44 (2010) 363–373. doi: 10.1016/j.jrp.2010.04.001.

[34] J. B. Hirsh, J. B. Peterson, Personality and language use in self-narratives, Journal of Research in Personality 43 (2009) 524–527. doi: 10.1016/j.jrp.2009.01.006.

[35] O. Aran, D. Gatica-Perez, One of a kind: Inferring personality impressions in meetings, in: ICMI 2013 - Proceedings of the 2013 ACM International Conference on Multimodal Interaction, 2013, pp. 11–18. doi: 10.1145/2522848.2522859.

[36] S. Okada, O. Aran, D. 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, Association for Computing Machinery, Inc, New York, New York, USA, 2015, pp. 15–22. URL: http://dl.acm.org/citation.cfm?doid=2818346.2820757. doi: 10.1145/2818346.2820757.

[37] J. I. Biel, D. Gatica-Perez, The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs, IEEE Transactions on Multimedia 15 (2013) 41–55. doi: 10.1109/TMM.2012.2225032.

[38] Z. Shen, A. Elibol, N. Y. 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, Institute of Electrical and Electronics Engineers Inc., 2019, pp. 402–407. doi: 10.1109/ICARM.2019.8834279.

[39] Z. Shen, A. Elibol, N. Y. Chong, Understanding nonverbal communication cues of human personality traits in human-robot interaction, IEEE/CAA Journal of Automatica Sinica 7 (2020) 1465–1477. doi: 10.1109/JAS.2020.1003201.

[40] P. K. Atrey, M. A. Hossain, A. El Saddik, M. S. Kankanhalli, Multimodal fusion for multimedia analysis: A survey, Multimedia Systems 16 (2010) 345–379. URL: http://link.springer.com/10.1007/s00530-010-0182-0. doi: 10.1007/s00530-010-0182-0.

[41] A. K. Katsaggelos, S. Bahaadini, R. Molina, Audiovisual Fusion: Challenges and New Approaches, in: Proceedings of the IEEE, volume 103, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 1635–1653. doi: 10.1109/JPROC.2015.2459017.

[42] T. Baltrusaitis, C. Ahuja, L. P. Morency, Multimodal Machine Learning: A Survey and Taxonomy, 2019. URL: http://arxiv.org/abs/1705.09406. doi: 10.1109/TPAMI.2018.2798607. arXiv:1705.09406.

[43] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, M. Zancanaro, Multimodal Recognition of Personality Traits in Social Interactions, in: ICMI'08: Proceedings of the 10th International Conference on Multimodal Interfaces, ACM Press, New York, New York, USA, 2008, pp. 53–60. URL: http://portal.acm.org/citation.cfm?doid=1452392.1452404. doi: 10.1145/1452392.1452404.

[44] J. A. Mioranda-Correa, I. Patras, A multi-task cascaded network for prediction of affect, personality, mood and social context using EEG signals, in: Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Institute of Electrical and Electronics Engineers Inc., 2018, pp. 373–380. doi: 10.1109/FG.2018.00060.

[45] A. V. Nefian, L. Liang, X. Pi, X. Liu, K. Murphy, Dynamic Bayesian networks for audio-visual speech recognition, Eurasip Journal on Applied Signal Processing 2002 (2002) 1274–1288. doi: 10.1155/S1110865702206083.

[46] S. M. Anzalone, G. Varni, S. Ivaldi, M. Chetouani, Automated Prediction of Extraversion During Human–Humanoid Interaction, International Journal of Social Robotics 9 (2017) 385–399. doi: 10.1007/s12369-017-0399-6.

[47] L. W. Morris, Extraversion and Introversion: An Interactional Perspective, Hemisphere Pub. Corp., Washington; New York, 1979.

[48] L. R. Goldberg, An Alternative "Description of Personality": The Big-Five Factor Structure, Journal of Personality and Social Psychology 59 (1990) 1216–1229. doi: 10.1037/0022-3514.59.6.1216.

[49] L. R. Goldberg, A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models., In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), Personality Psychology in Europe, 7 (1999) 7–28.

[50] S. D. Gosling, P. J. Rentfrow, W. B. Swann, A very brief measure of the Big-Five personality domains, Journal of Research in Personality 37 (2003) 504–528. doi: 10.1016/S0092-6566(03)00046-1.

[51] P. T. Costa, R. R. McCrae, Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI), Psychological Assessment Resources, Lutz Fla., 1992.

[52] R. R. McCrae, P. T. Costa, A contemplated revision of the NEO Five-Factor Inventory, Personality and Individual Differences 36 (2004) 587–596. doi: 10.1016/S0191-8869(03)00118-1.

[53] L. R. Goldberg, The Development of Markers for the Big-Five Factor Structure, Psychological Assessment 4 (1992) 26–42. doi: 10.1037/1040-3590.4.1.26.

[54] S. M. Geramian, S. Mashayekhi, M. T. B. H. Ninggal, The Relationship Between Personality Traits of International Students and Academic Achievement, Procedia - Social and Behavioral Sciences 46 (2012) 4374–4379. doi: 10.1016/j.sbspro.2012.06.257.

[55] H. D. Bui, N. Y. Chong, An Integrated Approach to Human-Robot-Smart Environment Interaction Interface for Ambient Assisted Living, in: Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO, volume 2018-Septe, IEEE Computer Society, 2019, pp. 32–37. doi: 10.1109/ARSO.2018.8625821.

[56] N.-Y. Chong, F. Mastrogiovanni (Eds.), Handbook of Research on Ambient Intelligence and Smart Environments, Advances in Computational Intelligence and Robotics, IGI Global, 2011. URL: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-61692-857-5. doi: 10.4018/978-1-61692-857-5.

[57] N. Ambady, F. J. Bernieri, J. A. Richeson, Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream, Advances in Experimental Social Psychology 32 (2000) 201–271. doi: 10.1016/s0065-2601(00)80006-4.

[58] D. Sanchez-Cortes, O. Aran, M. S. Mast, D. Gatica-Perez, A nonverbal behavior approach to identify emergent leaders in small groups, IEEE Transactions on Multimedia 14 (2012) 816–832. doi: 10.1109/TMM.2011.2181941.

[59] M. G. Frank, A. Maroulis, D. J. Griffin, Nonverbal Communication: Science and Applications, 2013.

[60] P. Patompak, S. Jeong, I. Nilkhamhang, N. Y. Chong, Learning Prox-

emics for Personalized Human–Robot Social Interaction, International Journal of Social Robotics 12 (2020) 267–280. doi: 10.1007/s12369-019-00560-9.

[61] J. T. Webb, Interview synchrony: An investigation of two speech rate measures in an automated standardized interview, Studies in dyadic communication (1972) 115–133.

[62] Z. Zafar, S. Hussain Paplu, K. Berns, Automatic Assessment of Human Personality Traits: A Step Towards Intelligent Human-Robot Interaction, in: IEEE-RAS International Conference on Humanoid Robots, volume 2018-Novem, IEEE Computer Society, 2019, pp. 670–675. doi: 10.1109/HUMANOIDS.2018.8624975.

[63] G. Mohammadi, A. Vinciarelli, M. Mortillaro, The voice of personality: Mapping nonverbal vocal behavior into trait attributions, in: SSPW'10 - Proceedings of the 2010 ACM Social Signal Processing Workshop, Co-located with ACM Multimedia 2010, 2010, pp. 17–20. doi: 10.1145/1878116.1878123.

[64] A. Guidi, C. Gentili, E. P. Scilingo, N. Vanello, Analysis of speech features and personality traits, Biomedical Signal Processing and Control 51 (2019) 1–7. doi: 10.1016/j.bspc.2019.01.027.

[65] O. Kampman, E. J. Barezi, D. Bertero, P. Fung, Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction, in: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), volume 2, 2018, pp. 606–611. URL: http://arxiv.org/abs/1805.00705. arXiv:1805.00705.

[66] R. D. P. Principi, C. Palmero, J. C. Junior, S. Escalera, On the Effect of Observed Subject Biases in Apparent Personality Analysis from Audio-visual Signals, IEEE Transactions on Affective Computing (2019) 1–14. . doi: 10.1109/taffc.2019.2956030. arXiv:1909.05568.

[67] D. Sanchez-Cortes, O. Aran, D. B. Jayagopi, M. Schmid Mast, D. Gatica-Perez, Emergent leaders through looking and speaking: From audio-visual data to multimodal recognition, Journal on Multimodal User Interfaces 7 (2013) 39–53. doi: 10.1007/s12193-012-0101-0.

[68] J. C. S. J. Junior, Y. Güçlütürk, M. Perez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. V. Gerven, R. V. Lier, S. Escalera, First Impressions: A Survey on Vision-based Apparent Personality Trait Analysis, IEEE Transactions on Affective Computing (2019) 1–20. doi: 10.1109/taffc.2019.2930058. arXiv:1804.08046.

[69] C. Beyan, F. Capozzi, C. Becchio, V. Murino, Prediction of the leadership style of an emergent leader using audio and visual nonverbal features, IEEE Transactions on Multimedia 20 (2018) 441–456. doi: 10.1109/TMM.2017.2740062.

[70] E. Murphy-Chutorian, M. M. Trivedi, Head pose estimation in computer vision: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 607–626. doi: 10.1109/TPAMI.2008.106.

[71] K. Fornalczyk, A. Wojciechowski, Robust face model based approach to head pose estimation, in: Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017, Institute of Electrical and Electronics Engineers Inc., 2017, pp. 1291–1295. doi: 10.15439/2017F425.

[72] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications (2017). URL: http://arxiv.org/abs/1704.04861. arXiv:1704.04861.

[73] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, SSD: Single shot multibox detector, volume 9905 LNCS, Springer Verlag, 2016, pp. 21–37. URL: http://arxiv.org/abs/1512.02325. doi: 10.1007/978-3-319-46448-0_2. arXiv:1512.02325.

[74] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110. doi: 10.1023/B:VISI.0000029664.99615.94.

[75] M. A. Fischler, R. C. Bolles, Random sample consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Communications of the ACM 24 (1981) 381–395. URL: http://portal.acm.org/citation.cfm?doid=358669.358692. doi: 10.1145/358669.358692.

[76] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2004. doi: 10.1017/cbo9780511811685.

[77] D. E. King, Dlib-ml: A machine learning toolkit, Journal of Machine Learning Research 10 (2009) 1755–1758.

[78] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 Faces In-The-Wild Challenge: database and results, Image and Vision Computing 47 (2016) 3–18. URL: http://dx.doi.org/10.1016/j.imavis.2016.01.002. doi: 10.1016/j.imavis.2016.01.002.

[79] Z. Zhang, A flexible new technique for camera calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 1330–1334. doi: 10.1109/34.888718.

[80] H. Admoni, B. Scassellati, Social Eye Gaze in Human-Robot Interaction: A Review, Journal of Human-Robot Interaction 6 (2017) 25–63. doi: 10.5898/jhri.6.1.admoni.

[81] R. Stiefelhagen, J. Zhu, Head orientation and gaze direction in meetings, in: Conference on Human Factors in Computing Systems - Proceedings, ACM Press, New York, New York, USA, 2002, pp. 858–859. URL: http://portal.acm.org/citation.cfm?doid=506443.506634. doi: 10.1145/506621.506634.

[82] Z. Cao, T. Simon, S. E. Wei, Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, volume 2017-Janua, 2017, pp. 1302–1310. URL: http://arxiv.org/abs/1812.08008. doi: 10.1109/CVPR.2017.143. arXiv:1812.08008.

[83] J. D. Markel, The SIFT Algorithm for Fundamental Frequency Estimation, IEEE Transactions on Audio and Electroacoustics 20 (1972) 367–377. doi: 10.1109/TAU.1972.1162410.

[84] X. D. Mei, J. Pan, S. H. Sun, Efficient algorithms for speech pitch estimation, in: Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, ISIMP 2001, 2001, pp. 421–424. doi: 10.1109/isimp.2001.925423.

[85] M. Xu, L. Y. Duan, J. Cai, L. T. Chia, C. Xu, Q. Tian, HMM-based audio keyword generation, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3333 (2004) 566–574. URL: http://link.springer.com/10.1007/978-3-540-30543-9_71. doi: 10.1007/978-3-540-30543-9_71.

[86] J. Macqueen, Some methods for classification and analysis, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, volume 233, 1967, pp. 281–297. URL: http://projecteuclid.org/bsmsp.

[87] C. M. Bishop, Pattern recognition and machine learning, Information science and statistics, Springer, New York, NY, 2006. URL: https://cds.cern.ch/record/998831.

[88] E. R. Pacola, V. I. Quandt, P. B. N. Liberalesso, S. F. Pichorim, H. R. Gamba, M. A. Sovierzoski, Influences of the signal border extension in the discrete wavelet transform in EEG spike detection, Revista Brasileira de Engenharia Biomedica 32 (2016) 253–262. URL: http://dx.doi.org/10.1590/2446-4740.01815. doi: 10.1590/2446-4740.01815.

[89] M. Jensen, Personality traits and nonverbal communication patterns, International Journal of Social Science Studies 4 (2016). doi: 10.11114/ijsss.v4i5.1451.

[90] S. M. Breil, S. Hirschmüller, S. Nestler, M. Back, Contributions of Nonverbal Cues to the Accurate Judgment of Personality Traits (2019). URL: https://psyarxiv.com/mn2je/. doi: 10.31234/osf.io/mn2je.