DOI:
10.1109/TAFFC.2017.2737019

Document Version
Peer reviewed version

Link to publication record in King's Research Portal

Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement

Oya Celiktutan, Efstratios Skordos and Hatice Gunes

Abstract—In this paper we introduce a novel dataset, the Multimodal Human-Human-Robot-Interactions (MHHRI) dataset, with the aim of studying personality simultaneously in human-human interactions (HHI) and human-robot interactions (HRI) and its relationship with engagement. Multimodal data was collected during a controlled interaction study where dyadic interactions between two human participants and triadic interactions between two human participants and a robot took place with interactants asking a set of personal questions to each other. Interactions were recorded using two static and two dynamic cameras as well as two biosensors, and meta-data was collected by having participants fill in two types of questionnaires, for assessing their own personality traits and their perceived engagement with their partners (self labels) and for assessing personality traits of the other participants partaking in the study (acquaintance labels). As a proof of concept, we present baseline results for personality and engagement classification. Our results show that (i) trends in personality classification performance remain the same with respect to the self and the acquaintance labels across the HHI and HRI settings; (ii) for extroversion, the acquaintance labels yield better results as compared to the self labels; (iii) in general, multi-modality yields better performance for the classification of personality traits.

Index Terms—Multimodal interaction dataset, human-human interaction, human-robot interaction, personality analysis, engagement classification, benchmarking

1 Introduction

Humans exchange information and convey their thoughts and feelings through gaze, facial expressions, body language and tone of voice along with spoken words, and infer 60-65% of the meaning of the communicated messages from these nonverbal behaviours during their daily interactions with others [1]. These nonverbal behaviours carry a significant information regarding social phenomena such as personality, emotions and engagement. Social Signal Processing (SSP) is the emerging research domain that aims at developing machines with the ability to sense and interpret human nonverbal behaviours, and bridging the gap between human nonverbal behaviours and social phenomena [2]. The first step of building an automatic system to model nonverbal behaviours is the acquisition of diverse, richly labelled data. Although a significant effort has been put into this in recent years, availability of annotated mutimodal databases is still limited [3], especially for studying personality and engagement.

Personality computing has become increasingly prominent over the last decade due to its strong relationship with important life aspects, including not only happiness, physical and psychological health, occupational choice, but also quality of relationships with others [4]. Individuals' interactions with others are shaped by their personalities and their impressions regarding others' behaviours and personalities [5]. This has been shown to be the case also for interactions with social robots [6]. Automatic personality analysis from nonverbal behaviours has therefore various practical applications ranging from matching recruiters with candidates [7] to improving humans' interaction experience with socially intelligent robots [8], [9].

Nonverbal behaviours are significant predictors of personality. Voice contains cues to the extroversion trait [10]. Extroverted people talk more, in both number of words and total speaking time. They also talk louder and faster with fewer hesitations and more variable pitch. Gaze and head movements are also strongly correlated with personality. For example, *dominance* and *extroversion* are found to be related to holding a direct facial posture and long durations of eye contact during interaction, whereas *shyness* and *social anxiety* are highly correlated with gaze aversion [11]. Extroverted people are found to be more energetic, leading to higher head movement frequency, more hand gestures and more postural shifts than introverted people [12], [13].

Many methods have demonstrated that people use facial and head cues including head movements [8], [14], head pose [15], facial expressions [16] as a basis for judging personality. Bodily cues such as frequency of posture shifts, amount of body movement, and hand gestures have been shown to be useful in predicting personality [14], [17]. While most of the works have focused on combining audio and visual cues only, Abadi *et al.* [18] showed that electroencephalogram (EEG) and physiological signals can be as effective as facial cues. However, the integration of multimodal sensors (more than two modalities) for predicting personality has remained a less explored problem due to the fact that multimodal datasets for the analysis of personality are scarce.

- *O. Celiktutan and H. Gunes are with the Computer Laboratory of the University of Cambridge, Cambridge, United Kingdom. E-mail: {Oya.Celiktutan.Dikici, Hatice.Gunes}@cl.cam.ac.uk*
- *E. Skordos is with the Department of Computer Science, University College London, London, United Kingdom. E-mail: efstratios.skordos.15@ucl.ac.uk*

In this paper, we present the Multimodal Human-Human-Robot Interactions (MHHRI) dataset that features the following:

- The MHHRI dataset comprises natural interactions between two human participants and between the same participants with a small humanoid robot, resulting in data acquired from a total of 18 participants recorded over the course of 12 independent interaction sessions.
- Each interaction was recorded using a set of sensors including two first-vision cameras (also known as egocentric cameras), two Kinect depth sensors and two physiological sensors, which resulted in approximately 4 hours 15 minutes of fully synchronised multimodal recordings. Although personality analysis from static, third vision perspective has been studied extensively, social signal processing from wearable cameras is still an understudied research problem.
- The multimodal recordings are annotated with respect to the self-assessed Big Five personality traits and engagement. Differently from the existing approaches in personality computing [4], we also provide acquaintance-assessments with respect to the Big Five personality traits.

We used a portion of the MHHRI dataset for automatic analysis in [8], [9], both focusing on the human-robot interaction setting. In this paper, we present a complete description of the MHHRI dataset including the recording protocol, the data collected from multiple modalities and their annotations together with baseline classification results for both interaction settings, i.e., human-human interaction and human-robot interaction.

2 Related Work

2.1 Multimodal Interaction Databases

Previous multimodal interaction databases that are available for research purposes can be divided into two groups based on their interaction settings, namely, human-human interaction and human-robot interaction. These databases are summarised in Table 2.

Human-human interaction databases were extensively reviewed in [3]. Here we only present the multimodal databases that are most relevant to our work in terms of interaction setting. Rozgic [19] investigated approach-avoidance behaviours in dyadic interactions. Interaction partners were asked to have a discussion on one of the nine suggested topics ranging from cheating in a relationship to a drinking problem. The data capturing system was composed of (i) a motion capture system that tracked and recorded 23 positions on each participants' upper bodies, (ii) a camera array that recorded close-up and far-field views of the interaction, and (iii) a microphone array that recorded audio of each participant individually and the whole interaction. The recorded data was annotated in two ways. Audio was annotated by English speakers based on turn taking (successful and unsuccessful interruptions), speaker segmentation, dialogue acts (question, statement, back-channel) and transcription. Expert coders also provided annotations at the subject-interaction level using the Couple Interaction Rating System, regarding participants' attitude (positive vs. negative), presence of blame, acceptance and approach-avoidance. One limitation of this database is that it was based on role-play. Participants were asked to play a pre-defined role and discuss about the flow of the conversation prior to the recording. Aubrey *et al.* [21], [22] focused on naturalistic and spontaneous conversations between two participants, without assigning any roles to them. Different topics including liking or disliking different genres of music, literature, movies, were suggested but the participants were not restricted to these topics and were observed to deviate from these topics frequently. Cardiff Conversation Database (CCDb) database [21] was first introduced with the emphasis of head/facial gestures in dyadic interactions, where each participant was recorded using a 3D dynamic scanner, an RGB camera and a microphone. Approximately one third of the conversations were annotated for various social signals including (dis)agreement, backchanneling, head nodding, head shaking and various affective states including happiness, surprise, thinking, and confusion. In [22], this database was further extended using a more sophisticated data capturing system to incorporate depth, and was called *4D CCdb*.

Remote Collaborative and Affective Interactions (RECOLA) database by Ringeval *et al.* [23] aimed at measuring emotions (arousal and valence) and social signals in a dyadic interaction setting. Interaction partners were asked to perform a collaborative task, i.e. the widely used Mission Survival Task, at remote locations, and were recorded using a set of sensors. In particular, audio-visual data was captured by web cameras and microphones; and Electrodemal Activity (EDA) and Electrocardiogram (ECG) were recorded by electrodes placed on the participants' fingers, palms, and inner side of their ankles, where all these recordings were synchronised continuously in time. They also collected two sets of annotations: self-assessment and other-assessment. In addition to self-assessment, they recruited six external observers to have the clips assessed along arousal and valence continuously in time, using their in-house tool. They also collected annotations with respect to five social signals, namely, agreement, engagement, dominance, performance and rapport based on a Likert scale rating. There were 46 participants in total, however, only recordings from 23 participants are available for external use. One disadvantage of this database is the fact that recordings from both participants are not available, which hinders the full analysis of a dyadic interaction.

In addition to the abovementioned multimodal HHI databases, there are a number of relevant audio-visual databases [20], [24]. MAHNOB Mimicry database was created by Bilakhia *et al.* for investigating conversational dynamics, but exclusively focusing on mimicry occurrences [24]. Interaction partners were composed of a naive participant and a confederate, and either discussed a contemporary socio-political topic or negotiated a tenancy agreement. The conversations were recorded using a set of audio-visual sensors including far-field and head mounted microphones, three close-up cameras capturing head, upper body and full body as well as a profile oriented camera capturing the interaction. One third of the interaction recordings were annotated in a semi-automatic manner, for motor mimicry behaviours of head gestures, hand gestures, body movement and facial expressions. However, this database does not have annotations of personality and engagement.

TABLE 1
Summary of the publicly available multimodal interaction databases described in Section 2.

Database / Year	Setting / Num. of Partners	Num. of Subjects	Num. of Interaction Sessions	Duration of Recordings	Multimodality	Annotations
MMDI [19] 2010	HHI: 2	NA	30 (of 8 are annotated)	3 h (45 m annotated)	audio, video, motion capture data	speaker segmentation, transcription, dialogue acts, turn taking, subject-interaction level labels (attitude, presence of acceptance, approach-avoidance)
ELEA [20] 2011	HHI: 3-4	148	40 (of 3 are audio-only)	10 h	audio, video	self-assessed personality, power, dominance, leadership, perceived leadership, dominance, competence, likeness
CCDb [21] 2013	HHI: 2	16	30 (of 8 are annotated)	5 h (80 m annotated)	audio, video	back/front channelling, (dis)agreement, happiness, surprise, thinking, confusion, head nodding/shaking/tilting, utterance
4D CCDb [22] 2015	HHI: 2	4	17	34 m	audio, video, depth	same as above
RECOLA [23] 2013	HHI: 2	46 (of 23 are available)	23	3.8 h (2.9 h available)	audio, video, ECG, EDA	valence, arousal, agreement, dominance, performance, rapport, engagement, utterance
MAHNOB [24] 2015	HHI: 2	60 (of 12 are confederates)	54 (of 15 are annotated)	11 h 40 m	audio, video	motor mimicry behaviours of head gestures, hand gestures, body movement and facial expressions
Vernissage [25] 2013	HRI: 3	26	13	2 h 23 m	audio, video, motion capture	speech utterances, head-location, nodding, focus of attention, speech transcription
JOKER [26] 2015	HRI: 2	37	111	8 h	audio, video, depth	self-assessed personality, sense of humour
MHHRI 2016	HHI:2 HRI: 3	18	48 (746 short clips)	6 h (4 h 15 m are synchronised)	audio, video, depth, EDA, temperature, 3-axis wrist acceleration	self-/acquaintance-assessed personality, self-reported engagement

Emergent Leadership (ELEA) corpus [20] has been one of the most popular databases for the analysis of the Big Five personality traits and other social traits including leadership, dominance, competence and likeness. It was initially built to study leadership in small groups of 3-4 participants, and its relationship with participants' personality traits. Each group was asked to perform winter survival task, while being recorded using a set of cameras (four close-up views, two side views, one centre view) and microphones. Before performing the task, each participant completed the Big Five personality questionnaire and another questionnaire that measured self-perceived power, dominance and leadership. In addition, after the task, the participants were asked to score others in the group with respect to their perceived leadership, dominance, competence and likeness.

Multimodal HRI databases are still scarce. Vernissage corpus was collected by Jayagopi *et al.* [25], comprising interactions of two participants with a humanoid robot. The robot served as an art guide, introducing paintings to the participants and then quizzing them in art and culture. A Wizard-of-Oz setup was used to manage the dialogue and control the robot's gaze and nodding. The interactions were recorded using three external cameras, two close-up microphones and a VICON motion capture system in addition to the robot's built-in camera and microphone, and were annotated with a set of nonverbal cues including speech utterances, 2D head location and visual focus of attention.

JOKER system, a robotic platform, was designed by Devillers *et al.* [26] for eliciting laughter from a single participant interacting with a humorous robot. The participants tested three different platforms: autonomous, semi-autonomous and non-autonomous (Wizard of Oz setup - the robot was remotely controlled by a human), and reported their satisfaction about each of the platforms. During the interaction, participants were recorded using microphones, web cameras and Kinect depth sensors, which resulted in a total of 8 hours of multimodal recordings. Participants were also asked to fill in a personality questionnaire and the sense of humour scale questionnaire. The goal of the autonomous platform was to recognise the emotions of the participants based on audio cues (paralinguistic features) in order to endow the robot with a comprehension of the user's receptiveness to humour before producing an action.

2.2 Automatic Analysis of Personality and Engagement

Three of the most related works in automatic personality analysis utilised the ELEA corpus [20]. Aran and Gatica-Perez [14] combined audio and motion features with a set of high level features based on head, body, speaking activity and focus of attention for predicting personality impressions (i.e., personality assessed by others). The best classification results were achieved with feature-level fusion for extroversion (74.5%). In their follow-up study [27], they took into account similar features, but, in addition to individual-level features, they proposed a method to detect temporal co-occurrence patterns in the target's features and the group's features (e.g., the others' change of posture as the target

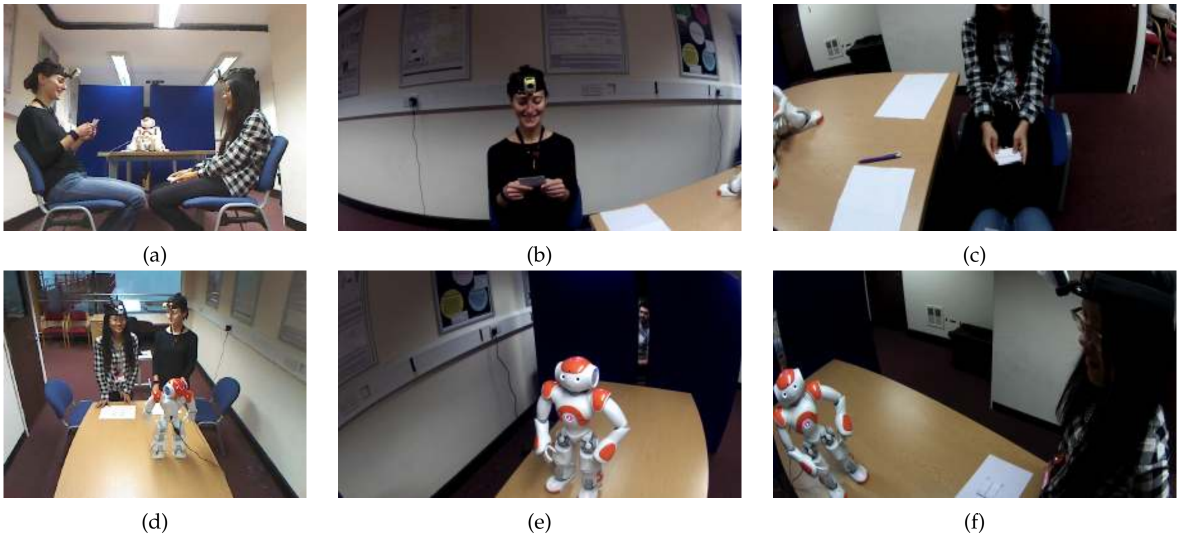
(a) (b) (c)
(d) (e) (f)

Fig. 1. The human-human interaction setup is shown in the first row: Simultaneously captured snapshots from (a) Kinect sensor; (b-c) ego-centric cameras that are placed on the forehead of the participants. The human-robot interaction is shown in the second row: Simultaneously captured snapshots from (d) Kinect depth sensor; (e-f) ego-centric cameras that are placed on the forehead of the participants.

speaks) and used these co-occurrence features to predict the personality of the target. While agreeableness was the trait most benefiting from co-occurrence features, the best classification accuracy was achieved with individual-level features for openness. Fang *et al.* [28] further explored similar features under three categories, namely, individual (related to only one participant), dyadic (related to a pair of participants) and group features (related to one participant with respect to the rest of the participants) for predicting self-assessed personality; unlike [27] dyadic and group features were extracted from audio clips only. Combining three feature categories yielded the best results as compared to using individual features alone, e.g., classification performance increased from 64.71% to 77.45% for extroversion. Individual and interpersonal features together were also found to be useful for predicting extroversion within the scope of HRI in [29] where Rahbar et al. combined similar individual features with interpersonal features including synchrony, dominance and proxemics.

There is another line of work examining the effect of personality matching (similar or complementary personality types) on the engagement state of the participants within the scope of HRI. Celiktutan and Gunes [8] investigated correlations between the self-assessed personality and interaction experience from first-person perspective, with respect to two robot conditions (extroverted vs. introverted). They found that perceived engagement with the extroverted robot is found to be positively correlated with participants' extroversion trait, indicating the *similarity rule*. They also extracted a set of low-level features from the first-person recordings in order to describe camera wearer's head movements and gaze patterns, and applied support vector regression in order to predict Big Five personality traits. This approach yielded state-of-the-art results for agreeableness, conscientiousness and extroversion. Salam *et al.* [9] expanded on this idea and further investigated the impact of the participants' personalities on their engagement states from the Kinect depth sensor recordings. Unlike [8], these recordings contained interactions between two participants and the robot from a static, third-person perspective. To do so, they collected personality labels and engagement labels from external observers using an online crowd-sourcing service, and extracted two sets of features, namely, individual and interpersonal features. They first applied Gaussian process regression for personality prediction. They then combined the predicted personality labels with the individual and interpersonal features to classify whether the participants were engaged or not, where the best results were achieved using individual features together with personality labels.

2.3 Our Work

The MHHRI dataset was built with the aim of studying personality simultaneously in dyadic human-human interactions and triadic human-human-robot interactions, and its relationship with perceived engagement. We conducted a controlled interaction study where, in the first stage, dyadic interactions between two human participants took place with the interactants asking a set of personal questions to each other. The second stage of the study was recorded during a triadic interaction between two participants and a humanoid robot, where the participants were asked to answer questions similar to the first stage, but posed by the robot. Sample snapshots are given in Figure 1.

The main contributions of the MHHRI dataset are highlighted in Table 2. The MHHRI dataset complements the existing databases along three main avenues:

- The MHHRI dataset incorporates two different interaction settings, namely, human-human interaction and human-robot interaction, whereas the previous databases exclusively focus on either human-human interaction or human-robot interaction.
- In addition to the static, third-vision cameras, the conversations were recorded using dynamic, first-person vision cameras. First-person vision provides the most relevant information for recognising social interactions [30] - people that the camera wearer interacts with tend to be cen-

TABLE 2
Comparison of the multimodal interaction databases with respect to the key features incorporated in the MHHRI dataset (NM: Num. of Modalities, IS: Interaction Setting, SC: Static Camera, DC: Dynamic Camera, PA: Personality Annot., EA: Engagement Annot.).

Database	NM	IS	SC	DC	PA	EA
MMDI [19]	3	HHI	✓	✗	✗	✗
ELEA [20]	2	HHI	✓	✗	✓	✓
CCDb [21], [22]	3	HHI	✓	✗	✗	✗
RECOLA [23]	4	HHI	✓	✗	✗	✓
MAHNOB Mimicry [24]	2	HHI	✓	✗	✗	✗
Vernissage [25]	3	HRI	✓	✓	✗	✗
JOKER [26]	3	HRI	✓	✗	✓	✗
MHHRI	6	HHI & HRI	✓	✓	✓	✓

tred in the scene, and are less likely to be occluded when captured from a co-located, first person perspective rather than from a static, third-person perspective. Social signal processing with respect to first-person vision cameras is still an understudied research problem.

- Similar to the ELEA corpus [20], the MHHRI dataset offers both personality and engagement labels. However, while the ELEA corpus [20] comprises audio-visual recordings only, the MHHRI dataset provides fully synchronised recordings of six different data modalities ranging from visual to physiological.

In order to demonstrate the usability of the MHHRI dataset, we present baseline results for personality and engagement classification by extracting a set of generic features from each modality and using Support Vector Machines. In particular, we aim at addressing the following research questions: (i) Does interaction setting (HHI or HRI) have an impact on the prediction of participants' personalities?; (ii) Do acquaintance assessments provide better performance than self-assessments for predicting participants' personalities?; and (iii) Does fusion of multiple modalities yield better performance than using a single modality for predicting participants' personalities and engagement states?

3 Multimodal Human-Human-Robot Interaction (MHHRI) Database

In this section, we describe the interaction scenarios and the data collection procedure, and summarise the collected data.

3.1 Tasks

3.1.1 Human-Human Interaction

The participant were guided into the experimental room with controlled lighting where they were asked to sit on a chair located near the edge of a table as shown in Figure 1-a, and were provided with an informed consent form that explained the overall goal of the study and how the interactions and the recordings would proceed. The participants were then asked to fill in a pre-study questionnaire, which asked them about demographic information and further information about their general behavioural preferences (see Section 3.2.2).

Each participant was provided with a set of cards with five questions in total (see Table 4). The session started with each of the participants describing themselves briefly, and continued with them asking one of the provided questions alternately. The participant on the left hand side always asked questions 1, 4 and 7 in Table 4, while the participant on the right hand side asked questions 2, 5 and 8. Some questions were common, i.e., questions 3 and 6. For question 6, the participants showed a cartoon about robots to their partner, and asked them to comment on the cartoon based on their subjective interpretation. While most of the questions aimed at measuring participants' previous experience with robots and their positive/negative attitudes towards robots, question 1 and 2 were more personal and asked about their good and bad memories to induce different emotional states. Once they went through all the questions, they were told to remain seated for the second stage.

3.1.2 Human-Robot Interaction

The second stage took place between the robot and the two participants as shown in Figure 1-d. For the robot test-bed, we used the humanoid robot Nao developed by Aldebaran Robotics[1] with the technical details of NaoQi version 2.1, head version 4.0 and body version 25. The robot was controlled remotely as a Wizard-of-Oz setup during the interaction. To manage the turn taking, an experimenter (i.e., operator) that sat behind a poster board, operated the robot using a computer, robot's camera as well as the other cameras placed in the experimental room.

The robot was initially seated and situated on the table. The interaction session was initiated by the robot standing up on the table and greeting the participants as illustrated in Figure 1-d. The participants were exposed to two different types of robot personalities. As previous literature suggests [12], [13], extroverts tend to be more energetic such that they have a louder and faster speaking style coupled with a higher hand or facial gesture frequency and more posture changes, whereas introverts tend to have more hesitations accompanied by gaze aversion. We manipulated the robot's behaviours and generated two types of personality, i.e., extroverted robot and introverted robot. While the extroverted robot displayed hand gestures and talked faster and louder, the introverted robot was hesitant, less energetic and exhibited no hand gestures in the course of the interaction. Table 3 summarises the difference in behaviour between the two personality types and provides example statements associated with each personality type from the robot's repertoire.

The conversation was performed in a structured routine and was led by the robot. The robot asked personal questions to each of the participants one-by-one as given in Table 4. The robot initiated the conversation by greeting the participants and by asking neutrally "You, on my right, could you please stand up? Thank you! What is your name?". Then the robot continued by asking about their occupations, feelings and so on, by specifying their names at each turn. After the interaction took place, each participant filled in a post-study questionnaire to evaluate their engagement (see Section 3.2.2). The study was approved by the Queen Mary University of London Ethics Committee,

[1] https://www.aldebaran.com/en

TABLE 3
Manifestation of two personality types through robot's verbal and nonverbal behaviours.

Type	Verbal	Nonverbal
EXT.	"Would you like me to dance for you?"; "It is amazingly exciting!"	Displays hand gestures, posture shifts; talks faster, louder
INT.	"Hmm . . . well, ok . . . would you like me to play music for you?"; "Not good ..."	Displays an almost static posture; talks slower, lower

TABLE 4
Questions used in the course of human-human interaction (HHI) and human-robot interaction (HRI).

ID	*HHI Questions*
1	Can you tell me about the best memory you have or the best event your have experienced in your life? Why?
2	Can you tell me about an unpleasant or sad memory you have had in your life? Why?
3	Have you ever watched a movie with robots? What is your favourite one? Why?
4	What has your experience with robots been?
5	Have you heard about humanoid robots? If yes, can you tell me about them?
6	(Show the cartoon and then ask:) In your opinion what does this cartoon try to communicate? What does it tell you about robots?
7	Do you think humanoid robots are unethical? Do you see them as dangerous?
8	What is your wildest expectation from a robot?
ID	***HRI Questions***
1	How has your day been <*participant's name*>?
2	How do you feel right now/about being here <*participant's name*>?
3	What do you do for a living? Do you like your job?
4	<*participant's name*>, I have a personal question for you. Is there something you would like to change in your life?
5	Can you tell me about the best memory you have or the best event you have experienced in your life?
6	Can you tell me about an unpleasant or sad memory you have had in your life?
7	What are your feelings toward robots? Do you like them?
8	Have you watched Wall-e? Do you like it?

and the participants were reimbursed with £10 for their time.

3.2 Data Collection

3.2.1 Sensor Data

A total of 18 participants (9 female), mostly graduate students and researchers, partook in our experiment. Each interaction session lasted from 10 to 15 minutes, and was recorded using multiple sensors. First-person videos were recorded using two Liquid Image ego-centric cameras[2] placed on the forehead of both participants and the robot's camera. The whole scene was also captured using two static Microsoft Kinect depth sensors (version 1)[3] placed opposite to each other (see Figure 1-a and d), resulting in RGB-D recordings. Sound was recorded via the microphones built in the ego-centric cameras. In addition to the audio-visual recordings, the participants were asked to wear a Q Sensor by Affectiva[4], which records the physiological signals, namely, electrodermal activity (EDA), skin temperature and 3-axis acceleration of the wrist.

We collected approximately 6 hours of multimodal recordings over the course of 12 interaction session, 7 of which were with the extroverted robot and 5 were with the introverted robot. Each session involved two participants, therefore some participants took part more than once provided that they had a different partner, and were exposed to different robot personalities. To ensure this, in the first 7 interaction sessions, the participants were exposed to the extroverted robot, and in the remaining 5 interaction sessions they were exposed to the introverted robot. Prior to the data collection, the physiological sensors and Kinect sensors were synchronised with the computer clock used in the experiments. However, the ego-centric recordings and the robot's camera were unsynchronised with the other cameras. For this reason, the experimenter switched on and off the light before each session started. This co-occurred appearance change in the cameras was used to synchronise the multiple videos (i.e., videos taken from 2 ego-centric cameras, 2 Kinect depth sensors and robot's camera) in time. Basically we calculated the amount of appearance change between two successive frames based on gray-level histograms.

For further analysis, we segmented each recording into short clips using one question and answer window. In the HHI task, each clip contains participants asking one of the questions listed in Table 4 to their interaction partners. Similarly, in the HRI task, each clip comprises the robot asking a question to one of the participants and the target participant responding accordingly. This yielded 290 clips of HHI, 456 clips of HRI, and 746 clips in total for each data modality. However, Q sensor did not work during one of the sessions, resulting in 276 physiological clips of HHI. Each clip has a duration ranging from 20 to 120 seconds, resulting in a total of 4 hours 15 minutes of fully synchronised multimodal recordings. Figure 2 illustrates simultaneously captured snapshots from the ego-centric clips and EDA clips, from the HHI setting.

3.2.2 Annotation Data

The participants were asked to complete two questionnaires separately, one was prior to the interaction session (pre-study questionnaire) and the other one was after the interaction session (post-study questionnaire). All measures are on a 10-point Likert scale (from very low to very high).

The pre-study questionnaire aims to assess personal behavioural tendencies, i.e., how an individual sees herself in the way she approaches problems, likes to work, deals with feelings and emotions, and manages relationships with others, along the widely known Big Five personality traits [4]. These five personality traits are extroversion (assertive, outgoing, energetic, friendly, and socially active), neuroticism (having tendency to negative emotions such as anxiety, depression or anger), openness (having tendency to changing experience, adventure, and new ideas), agreeableness (cooperative, compliant, and trustworthy) and con-

[2] www.liquidimageco.com/products/model-727-the-ego-1
[3] en.wikipedia.org/wiki/Kinect
[4] http://qsensor-support.affectiva.com/

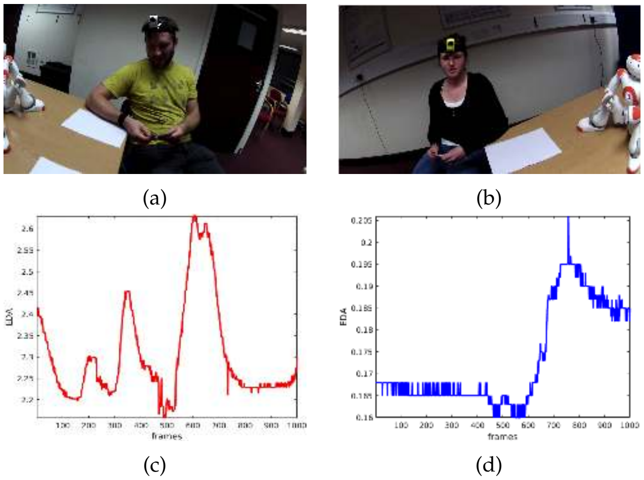

Fig. 2. Examples from the human-human interaction (HHI) recordings: (a-b) first-person views (left: Participant A; right: Participant B); (c-d) EDA signals from the biosensors for one segmented clips (left: Participant A; right: Participant B).

scientiousness (self-disciplined, organized, reliable, consistent). The commonly used method to measure these traits is the Big Five Inventory [31]. In our study, we used the BFI-10 [32], the short version of the Big-Five Inventory, which has been used in other studies, e.g., [4], [9]. Each item contributes to the score of a particular trait.

The post-study questionnaire consists of six items listed in Table 5, which evaluate the participants' engagement with the human participant and the robot, and measures their impressions about the robot's behaviours and abilities.

Although the general trend in the personality computing has been collecting personality impressions at zero-acquaintance, where targets are recorded and then recorded clips are viewed and rated with respect to the personality traits by strangers, Kenny *et al.* [33] reported the highest level of agreement among judges when the judges and the target were highly acquainted, i.e., the participants had known each other for an extended time, except for extroversion. In our case, the participants were also acquainted, being part of the same research group and working together in an open space office environment for over one year. For this reason, in addition to the pre-study and post-study questionnaires, we asked participants to assess others' personalities using the BFI-10 [32] after the study took place. This resulted in 9 to 12 acquaintance ratings per participant.

3.3 Analysis of Annotations

Prior to generating ground truth personality labels, we examined the validity of acquaintance assessments through measuring intra-agreement and inter-agreement, and their correlation with self-assessments (hereafter referred to as self-acquaintance agreement). Intra-agreement (also known as internal consistency) evaluates the quality of personality judgements based on correlations between different questionnaire items that contribute to measuring the same personality trait by each acquaintance. We measured intra-agreement in terms of standardised Cronbach's α. The resulting α coefficient ranges from 0 to 1; higher values are associated with higher internal consistency and values less than 0.5 are usually unacceptable [34].

TABLE 5
Post-study questionnaire.

ID	Question	Measured Aspect
1	I enjoyed the interaction with the human.	Engagement
2	I found the robot behaviour realistic.	Believability
3	I thought the robot was being positive.	Valence
4	I thought the robot was assertive and social.	Extroversion
5	I thought the robot was being supportive.	Empathy
6	The robot could successfully recognize me.	Capability
7	I enjoyed the interaction with the robot.	Engagement

TABLE 6
Analysis of acquaintance assessments: Intra-agreement in terms of Cronbach's α (good reliability > 0.70 is highlighted in bold); (b) Inter-agreement in terms of ICC(1,k) (at a significance level of $^{**}p < 0.01$ and $^{***}p < 0.001$); (c) Self-acquaintance agreement in terms of Pearson Correlation (at a significance level of $^{**}p < 0.01$ and $^{***}p < 0.001$).

	Intra-agr.	Inter-agr.	Self-acquaintance agr.
EXT	**0.76**	0.79***	0.17**
AGR	0.69	0.73***	0.09
CON	**0.71**	0.62**	0.28***
NEU	**0.76**	0.58**	0.02
OPE	0.15	0.25	0.10

Inter-agreement refers to the level of consensus among acquaintances. We computed the inter-agreement in terms of Intra-Class Correlation (ICC) [35]. ICC assesses the reliability of the acquaintances by comparing the variability of different ratings of the same target to the total variation across all ratings and all targets. We used ICC(1,k) as in our experiments each target subject was rated by a different set of k acquaintances. ICC(1,k) measures the degree of agreement for ratings that are averages of k independent ratings on the target subjects.

Self-acquaintance agreement measures the similarity between the personality judgements made by self and acquaintances. We computed self-acquaintance agreement in terms of Pearson correlation and tested the significance of correlations using Student's t distribution. Correlation was computed between the target's self-reported responses and the mean of the acquaintances' scores per trait.

In Table 6, we presented intra-agreement, inter-agreement and self-acquaintance agreement for each trait. We obtained a high level of intra-/inter-agreement for all personality traits except for openness, which demonstrates the validity of acquaintance assessments for training automatic predictors. However, we obtained significant correlations between self- and acquaintance-ratings for extroversion and conscientiousness only. Finally, we computed ground-truth personality labels from acquaintances' assessments by simply averaging over all raters.

As explained in Section 3.1.2, we manipulated the robot's behaviours to display a certain personality type. We manually inspected the participants' perception of the robot's personality with respect to these conditions, namely, extroverted robot versus introverted robot, where high scores were associated with perceived extroversion and low scores with perceived introversion. For the extroverted robot, the participants were unanimous in that the robot was extro-

TABLE 7
Pearson correlations between the participants' Big Five personality traits and their engagement at a significance level of $^{**}p < 0.01$ and $^{***}p < 0.001$: (a) human-human interaction (HHI) setting; (b) human-robot interaction (HRI) setting. (EXT: extroversion, AGR: agreeableness, CON: conscientiousness, NEU: neuroticism, OPE: openness, SELF: self labels, ACQ: acquaintance labels.)

(a) HHI	**Extroverted Partner**		**Introverted Partner**	
	SELF	ACQ	SELF	ACQ
EXT	-0.39	-0.36	0.62	0.56
AGR	0.17	0.41	0.42	$\mathbf{0.67}^{**}$
CON	-0.26	-0.32	0.74	0.09
NEU	0.38	0.45	0.62	-0.02
OPE	0.33	0.07	-0.15	-0.07

(b) HRI	**Extroverted Robot**		**Introverted Robot**	
	SELF	ACQ	SELF	ACQ
EXT	$\mathbf{0.76}^{**}$	$\mathbf{0.68}^{**}$	-0.25	-0.31
AGR	0.43	$\mathbf{0.82}^{***}$	-0.08	0.52
CON	0.40	0.35	0.49	0.00
NEU	-0.01	0.10	0.13	0.05
OPE	0.10	0.28	-0.05	-0.34

verted as all scores were greater than 5 (with a mean centred at 7), however, for the introverted robot, their responses varied with a mean centred at 5 (please refer to the Supplemental Material for details).

An in-depth analysis of personality and its relationship with engagement was done for the HRI setting in [9], by extracting a rich set of visual features and applying classification/regression methods. Here we investigated the possible links between the Big Five personality traits of the participants and their engagement with their partners with respect to the extroversion / introversion trait of the human partner in the HHI setting (Table 7-a) and with respect to the extroversion / introversion trait of the robot in the HRI setting (Table 7-b). In Table 7, we presented Pearson correlation results together with significance levels for two types of personality labels, namely, self labels and acquaintance labels. More explicitly, we classified the interaction partners into two classes (e.g., extroverted vs. introverted) based on both self labels and acquaintance labels.

Looking at the HHI setting (see Table 7-a), we did not observe any significant correlation in the extroverted human-partner condition. Cuperman and Ickes [5] indicated that more agreeable people reported having more enjoyable interactions. Conforming to this finding, we found large correlations between agreeableness and engagement for both the self labels and the acquaintance labels in the introverted human partner condition. Especially, this yielded a significant correlation of 0.67 with $p < 0.01$ for the acquaintance labels.

Looking at the HRI setting (see Table 7-b), for the extroverted robot condition, engagement with the robot was found to be significantly correlated with participants' extroversion trait, which validates the similarity rule [6], [36]. A study of agreeableness reported that more agreeable people showed strong self-reported *rapport* when they interacted with a virtual agent [37]. We observe that perceived engagement with the robot was highly correlated with the agreeableness trait of the participants, similarly to the HHI setting. No significant correlation was obtained for the introverted robot condition.

4 Baseline Feature Extraction

This section describes a set of generic features that we use for personality and engagement classification, which can be categorised based on the modality types as follows.

Audio Features (AF). We extracted two sets of vocal features using the speech feature extraction tool described in [38] to model audio cues. This tool uses a two-level HMM to segment each speech signal into *voiced* and *unvoiced* segments, and group the voiced segments into speaking and non-speaking segments. Prosodic style was then represented using nine types of features, including, voicing rate, formant frequency, confidence in formant frequency, spectral entropy, value of largest autocorrelation peak, location of largest autocorrelation peak, number of autocorrelation peaks, energy and time derivative of energy, mel-frequency cepstral coefficients (MFCC). For example while voicing rate is associated with speaking rate and fluency, high energy indicates loudness. Except for voicing rate, these features were computed at a frame rate of 62.5 Hz and were summarised over the whole clip by calculating the mean and the standard deviation. Voicing rate was defined as the ratio of the overall duration of voiced segments to the total duration of speech. This resulted in a total of 24 features per clip.

Physiological Features (PF). Each physiological clip comprises 5 measurements including Electrodermal Activity, skin temperature and 3-axis acceleration of the right wrist over time. As a preprocessing step, we applied a person-dependent normalisation method. We scaled each measurement into the range of $[0, 1]$ independently, where maximum values were computed by taking into account all the measurements pertaining to a specific participant. We downsampled each clip such that each clip had a sampling rate of 8kHz, and applied a 3rd order Butterworth filter with a cutoff frequency of 1kHz in order to remedy the noisy measurements. We then extracted a set of simple features including maximum, minimum, mean, standard deviation, mean absolute difference of first order and second order derivatives over time, which resulted in $6 \times 5\ measurements = 30$ features per physiological clip.

First Person Vision (FPV) Features. First-person vision directly reflects where target participants are looking at, and how they are interacting with their interactants, as shown in Figures 1-(b-c, e-f) and 2-(a-b). For modelling head movements and visual focus of attention, we extracted two types of features from the target participants' first-person vision recordings, namely, head motion signatures (FPV-HMS) and low-level descriptors (FPV-LLD).

In order to compute FPV-HMS features, we first estimated the camera motion using the method in [39]. Prior to this, we applied a robust person detector in order to eliminate local motion vectors as each FPV clip contained a participant interacting with another participant. For the person detector, we used the Fast Region-based Convolutional Neural Networks (Fast R-CNN) [40]. While training details can be found in [40], during detection, Fast R-CNN took as input an entire image and a list of object proposals (only person and background in our case) to score. For each test Region of Interest (ROI), the forward pass delivered a class posterior probability distribution and a set of predicted bounding-box offsets relative to its associated RoI. Then a detection confidence for each RoI was assigned using the estimated class posterior probabilities, and non-maximum suppression was performed to eliminate redundant, overlapping bounding boxes. In our experiments, we only took into account the bounding boxes that had a detection confidence larger than 0.9. Since Fast R-CNN performed person detection frame by frame, we further applied a filtering operation in order to smooth the estimations and deal with the missing values.

We used the estimated bounding-boxes together with the method in [39] for generating global motion vectors. Briefly, we found the correspondences between two frames and then estimated homography by combining two complementary approaches. First, speeded-up robust features (SURF) were extracted, and were matched based on the nearest neighbourhood rule. Secondly, motion vectors were computed based on the optical flow algorithm with polynomial expansion, and were refined using the good-features to track criterion. Once the candidate matches were obtained, the homography between two frames was estimated using the random consensus method (RANSAC), and was used to compute motion vectors at each pixel for every frame. Following [41], we averaged these vectors over all pixels in a frame, which resulted in two dimensional (i.e., horizontal and vertical) global motion features for each frame.

In addition to the global motion features, we computed blur and illumination features similarly to [8]. Rapid head movements might lead to significant motion in the first-person videos, which can be characterised by motion blur. Large scene changes, indicating large head rotations, may also cause drastic illumination changes. Blur features were computed based on the no-reference blur estimation algorithm of [42]. Given a frame, this algorithm yielded two values, vertical (BLUR-Ver) and horizontal blur (BLUR-Hor), ranging from 0 to 1 (the best and the worst quality, respectively). We also calculated the maximum blur (BLUR-Max) over the vertical and the horizontal values. For illumination, we simply calculated the mean (ILLU-Mean) and the median (ILLU-Med) of the pixel intensity values per frame. In total, we obtained 7 features (2 global motion, 3 blur and 2 illumination) per frame, and summarised these features over time by computing simple statistics as explained before, namely, mean, standard deviation etc., resulting in $6 \times 7\ features = 42$ FPV-HMS features per clip.

We already used the low-level descriptors (FPV-LLD) for predicting personality in the HRI setting in [8], and found it to be useful specifically for predicting agreeableness. Here, we extracted the same set of features from the HHI first-person videos as well, and compared FPV-LLD features with FPV-HMS features. Briefly, in addition to the blur and illumination features described above, we extracted optical flow features using the SIFT flow algorithm proposed in [43]. First, we computed a dense optical flow estimate for each frame, and converted the vertical and horizontal flow estimates of a pixel into magnitude and angle. Then, we calculated a set of statistical and frequency features from the magnitude values and angle values (for details, please refer to [8]). Together with blur and illumination features, this resulted in 40 FPV-LLD features per clip.

Second-Person Vision (SPV) Features. In Figure 2, considering Participant B as a target participant, FPV refers to what Participant B sees from her perspective (see Figure 2-(a)), while SPV refers to how Participant B is seen from Participant A's perspective (see Figure 2-(b)). In addition to FPV features, we therefore extracted SPV features to capture the bodily cues of the target participants from their interactants' perspective. Bodily cues such as frequency of posture shifts, amount of body movement, hand gestures have been proven to be useful in predicting personality [14], [17]. We extracted the improved dense trajectory features [39] as they were shown to be well suited to describe the participants from their partners' perspective without tracking them explicitly by Yonetani *et al.* [41]. The improved dense trajectory method is a standard method to extract features for action recognition from third-person perspective, e.g., as illustrated in Figure 1-(a) and (d). As partially explained above, features were densely sampled and tracked based on the optical flow fields with the good-features to track criterion over a short period of time (e.g., 15 frames). Then traditional descriptors including the Histograms of Oriented Gradients (HOG), Histograms of Optical Flows (HOF) and Motion Boundary Histograms (MBH) were extracted along trajectories and encoded using improved Fisher Vectors [44]. In our experiments, we used the default parameters as provided in [39] (e.g., the length of the trajectory, sampling stride neighbourhood size, etc.), and obtained 96-dimensional HOG per feature point (we did not take into account HOF and MBH features in this paper). We constructed a visual dictionary using the implementation in [45], where we set the number of clusters to 32. Since the durations of the clips varied, we applied encoding to the groups of 50 feature points neighbouring in time. After FV encoding, this resulted in $96 \times 32 \times 2 = 6144$-dimensional SPV-HOG features. During classification, we aggregated classifier outputs over all groups based on the majority voting approach to obtain a final decision per clip.

5 Baseline Classification Experiments

We performed baseline classification experiments to investigate the following hypotheses:

- *H1*. Certain personality traits are easier to classify in the HHI setting as compared to the HRI setting. Our results in [17] showed that situational context (e.g., interacting with different partners) has an impact on the prediction of personality.
- *H2*. Classifiers trained with labels obtained from self-assessments (hereafter, self labels) work better than classifiers trained with labels obtained from acquaintances' assessments (hereafter, acquaintance labels) for predicting certain personality traits. Self-acquaintance agreement results (see Section 3.3) show that correlations between self- and acquaintance-labels are too low except for conscientiousness, and hence indicate that self- and acquaintance-labels denote different sources of information.
- *H3*. Combining features that are extracted from different data modalities yields better performance than using a single feature type/data modality for predicting personality and engagement.

5.1 Experimental Results

We approached the personality/engagement prediction problem as a binary classification problem. We used conventional Support Vector Machines (SVMs) with a Radial Basis Function kernel in conjunction with AF, PF and FPV features and with a linear kernel in conjunction with SPV features, where we optimised the parameters in a subject-independent fashion. More explicitly, we evaluated the classification performance using a double leave-one-subject-out cross validation approach: each time we used all the data from one subject for testing, and all the data from the remaining 17 subjects for training and validation, and selected the best parameters on the training and validation sets using leave-one-subject-out cross validation. This is a common practice to ensure better generalizability of the trained models to the unseen subjects.

We trained a separate classifier with each feature type described in Section 4 for each personality trait and engagement, using two sets of labels, namely, self labels and acquaintance labels. Given the labels ranging from 1 to 10, we binarised them with respect to the mean computed over training samples for each (outer) cross validation fold, and grouped participants into two classes (i.e., low and high) per personality trait/engagement. We further fused classifier decision values (referred to as decision-level fusion) in order to examine the impact of combining different data modalities on the classification performance.

In Tables 8 and 9, we presented the classification results with self labels and acquaintance labels in terms of F-Score for HHI and HRI settings, respectively. We tabulated the engagement classification results with self labels only in Table 10.

5.2 Discussion

Our classification results partially support H2 and H3. In contrast to what we hypothesised in H1, classification results show a similar trend in the classification of personality traits across the HHI setting and the HRI setting. Below, we discuss our findings in comparison to the previous works.

(H1) HHI versus HRI. In [17], we showed that situational context had an impact on the prediction of personality as the participants showed a different facet of their personalities when they were interacting with different virtual agents. However, comparing Tables 8 and 9, we could not observe any difference between the two interaction contexts (i.e., HHI versus HRI). In other words, in both HHI and HRI settings, extroversion, openness and neuroticism were classified better with the acquaintance labels, and agreeableness and conscientiousness with the self labels. However, the overall performance was found to be slightly lower in the HRI setting as compared to the HHI setting. The same trend can also be observed for engagement classification in Table 10.

Although performance trends with respect to the personality traits were similar, different feature types / combinations contributed differently in the HRI setting. For example, while physiological features (PF) yielded the best performance for agreeableness in the HHI setting (mean F-Score = 0.62, see Table 8-(a)), combining audio features with first person vision low level descriptors (AF + FPV-LLD) resulted in the best performance in the HRI setting (mean F-Score = 0.57, see Table 9-(a)). For extroversion, the physiological features together with the second person vision HOG (PF + SPV-HOG) features performed better in the HHI setting (mean F-Score = 0.70, see Table 8-(a)), while audio features (AF) alone yielded better results than all the feature combinations in the HRI setting (mean F-Score = 0.60, see Table 9-(b)). This was due to the fact that interaction style and recording setup were different in the HHI and HRI settings. For example, in the HHI setting, the participants were sitting and looking face-to-face, whereas they were standing in the HRI setting, and their attention was shared between the robot and the other participant. Due to the same reason, second-person vision HOG (SPV-HOG) features did not work in the HRI setting, and therefore were not included in this paper. The best results were obtained with acquaintance labels where mean F-Score was 0.36 for extroversion in the HRI setting.

(H2) Self versus acquaintance labels. Supervising the classifiers with the self labels (see Table 8-(a)) yielded the best results for extroversion (mean F-Score = 0.70 with PF + SPV-HOG), agreeableness (mean F-Score = 0.62 with PF), conscientiousness (mean F-Score = 0.59 with PF + FPV-HMS) and neuroticism (mean F-Score = 0.61 with FPV-HMS + SPV-HOG). Fang *et al.* [28] used ELEA corpus [20] that comprises human-human interactions, similar to our HHI setting. Although the classification results were given in terms of accuracy, Fang *et al.* [28] obtained the best results with self labels for conscientiousness and neuroticism, followed by agreeableness and extroversion. On the other hand, Abadi *et al.* [18] focused on personality classification in a completely different scenario through the analysis of participants' implicit responses to emotional videos, and obtained the lowest performance for conscientiousness (mean F-Score = 0.53). Extroversion and openness were classified with a higher accuracy, yielding respective mean F-Score = 0.70 and mean F-Score = 0.69. As expected, our results were more in line with the results in [28].

Supervising the classifiers with the acquaintance labels (see Table 8-(b)) significantly improved the classification of openness (mean F-Score = 0.61 with FPV-LLD + SPV-HOG), which has not been reported in the literature before. Using the acquaintance labels was on a par with using the self labels for the classification of extroversion (mean F-Score = 0.68 with PF + SPV-HOG) and neuroticism (mean F-Score = 0.60 with AF + SPV-HOG). Similarly, Aran and Gatica-Perez obtained the highest classification accuracy for extroversion in [14], where external observers provided zero-acquaintance personality assessments for the clips in the ELEA corpus [20]. Indeed, Kenny *et al.* [33] suggested that even complete strangers can share a common sense of judging extroversion, however, for the other traits, the highest level of agreement among judges is reached when the judges and the target were highly acquainted.

Looking at the HRI setting (see Table 9), for extroversion, openness and neuroticism, the acquaintance labels yielded better results, while the self labels would be a better choice for predicting agreeableness and conscientiousness. Overall, our classification results partially supported H2.

(H3) Single-modality versus multi-modality. In the HHI setting, combining features extracted from different modalities boosted the classification performance for conscien-

TABLE 8
Personality classification of HHI clips: (a) self labels; and (b) acquaintance labels. The results are presented in terms of F Score; the best results with respect to the average performances (last column) are highlighted in bold. (AF: Audio Features, PF: Physiological features, FPV-HMS: FPV head motion signatures, FPV-LLD: FPV low level descriptors, SPV-HOG: SPV histogram of gradients).

Features	EXT		AGR		CON		NEU		OPE		Ave.
(a) Self	low	high	low	high	low	high	low	high	low	high	
AF	0.38	0.58	0.59	0.48	0.70	0.39	0.79	0.21	0.36	0.52	0.50
PF	0.25	0.68	**0.64**	**0.59**	0.57	0.37	0.43	0.27	0.42	0.46	0.47
FPV-HMS	0.31	0.64	0.74	0.36	0.69	0.31	0.75	0.28	0.24	0.62	0.49
FPV-LLD	0.34	0.42	0.77	0.41	0.61	0.36	0.50	0.37	0.24	0.69	0.47
SPV-HOG	0.33	0.79	0.68	0.66	0.59	0.09	0.66	0.09	0.35	0.60	0.48
AF+PF	0.32	0.64	0.43	0.47	**0.58**	**0.54**	0.49	0.36	0.39	0.49	0.47
AF+FPV-HMS	0.35	0.62	0.42	0.34	**0.56**	**0.54**	0.24	0.34	0.31	0.58	0.43
AF+FPV-LLD	0.36	0.51	0.39	0.34	**0.52**	**0.51**	0.45	0.34	0.32	0.62	0.44
AF+SPV-HOG	**0.65**	**0.58**	0.58	0.56	0.68	0.34	**0.58**	**0.59**	0.57	0.58	0.57
PF+FPV-HMS	0.28	0.67	0.43	0.40	**0.60**	**0.57**	0.54	0.39	0.35	0.55	0.48
PF+FPV-LLD	0.30	0.58	0.41	0.40	**0.57**	**0.53**	0.67	0.40	0.36	0.59	0.48
PF+SPV-HOG	**0.74**	**0.65**	0.36	0.60	0.63	0.45	0.64	0.40	0.51	0.73	0.57
FPV-HMS+SPV-HOG	**0.68**	**0.64**	0.50	0.58	0.70	0.37	**0.58**	**0.64**	0.49	0.65	0.58
FPV-LLD+SPV-HOG	0.45	0.66	**0.56**	**0.59**	0.52	0.47	0.53	0.52	0.43	0.71	0.54
(b) Acquaintance											
AF	0.48	0.28	0.31	0.61	0.55	0.35	**0.50**	**0.55**	0.55	0.41	0.46
PF	0.60	0.04	0.70	0.00	0.78	0.34	0.28	0.51	0.28	0.41	0.40
FPV-HMS	0.59	0.42	0.43	0.54	0.54	0.31	0.63	0.35	0.29	0.65	0.47
FPV-LLD	0.42	0.34	0.56	0.43	0.68	0.33	0.43	0.30	0.40	0.51	0.44
SPV-HOG	0.60	0.73	0.60	0.60	0.77	0.44	0.71	0.63	0.58	0.72	0.64
AF+PF	**0.57**	**0.64**	0.58	0.43	0.40	0.45	0.39	0.53	**0.52**	**0.64**	0.52
AF+FPV-HMS	**0.55**	**0.55**	0.37	0.58	0.45	0.42	0.57	0.47	0.52	0.49	0.50
AF+FPV-LLD	**0.58**	**0.56**	0.45	0.54	0.46	0.43	0.53	0.41	**0.53**	**0.51**	0.50
AF+SPV-HOG	0.53	0.64	0.47	0.61	0.61	0.41	**0.61**	**0.59**	0.53	0.68	0.57
PF+FPV-HMS	**0.63**	**0.64**	0.61	0.37	0.46	0.48	0.48	0.45	0.55	0.51	0.52
PF+FPV-LLD	**0.64**	**0.63**	0.59	0.39	0.46	0.48	0.47	0.41	**0.55**	**0.52**	0.51
PF+SPV-HOG	**0.63**	**0.72**	0.66	0.40	0.61	0.49	0.52	0.56	0.56	0.68	0.58
FPV-HMS+SPV-HOG	**0.61**	**0.64**	0.52	0.57	0.65	0.45	0.66	0.50	0.56	0.55	0.57
FPV-LLD+SPV-HOG	**0.62**	**0.66**	0.58	0.52	0.64	0.45	0.57	0.46	**0.57**	**0.64**	0.57

TABLE 9
Personality classification of HRI clips: (a) self labels; and (b) acquaintance labels. The results are presented in terms of F Score; the best results with respect to the average performances (last column) are highlighted in bold. (AF: Audio Features, PF: Physiological features, FPV-HMS: FPV head motion signatures, FPV-LLD: FPV low level descriptors, SPV-HOG: SPV histogram of gradients).

Features	EXT		AGR		CON		NEU		OPE		Ave.
(a) Self	low	high	low	high	low	high	low	high	low	high	
AF	0.42	0.24	0.50	0.46	**0.69**	**0.47**	0.64	0.31	0.40	0.54	0.47
PF	0.26	0.53	0.77	0.38	0.41	0.37	0.32	0.41	**0.45**	**0.63**	0.45
FPV-HMS	0.25	0.59	0.57	0.39	0.56	0.37	0.80	0.30	0.23	0.64	0.47
FPV-LLD	0.34	0.50	0.52	0.36	0.52	0.38	0.64	0.30	0.29	0.46	0.43
AF+PF	0.35	0.42	0.56	0.34	**0.56**	**0.50**	0.67	0.45	0.42	0.59	0.49
AF+FPV-HMS	0.35	0.45	0.66	0.43	**0.60**	**0.49**	0.39	0.38	0.33	0.59	0.47
AF+FPV-LLD	0.35	0.46	**0.65**	**0.48**	**0.59**	**0.48**	0.44	0.38	0.31	0.55	0.47
PF+FPV-HMS	0.26	0.56	0.43	0.33	**0.60**	**0.52**	0.56	0.39	0.36	0.63	0.46
PF+FPV-LLD	0.29	0.54	0.49	0.41	**0.58**	**0.50**	0.55	0.39	0.33	0.58	0.47
(b) Acquaintance											
AF	**0.50**	**0.69**	0.29	0.50	0.69	0.33	0.28	0.35	0.45	0.56	0.46
PF	0.52	0.27	0.69	0.42	0.59	0.21	**0.67**	**0.62**	0.24	0.56	0.48
FPV-HMS	**0.52**	**0.53**	0.26	0.37	0.42	0.42	0.57	0.42	0.31	0.52	0.43
FPV-LLD	**0.52**	**0.59**	0.22	0.53	0.58	0.34	0.40	0.39	0.55	0.44	0.46
AF+PF	0.52	0.47	0.55	0.47	**0.55**	**0.52**	0.50	0.48	0.49	0.46	0.50
AF+FPV-HMS	0.63	0.36	0.27	0.44	**0.58**	**0.47**	0.44	0.38	0.56	0.43	0.46
AF+FPV-LLD	0.57	0.36	0.26	0.47	**0.59**	**0.47**	0.43	0.38	**0.50**	**0.46**	0.45
PF+FPF-HMS	0.43	0.54	0.53	0.39	**0.61**	**0.53**	**0.62**	**0.53**	0.50	0.46	0.51
PF+FPV-LLD	0.42	0.48	0.46	0.45	**0.62**	**0.51**	0.56	0.48	0.46	0.48	0.49

tiousness when self labels were used (see Table 8-(a)), and for extroversion and openness when the acquaintance labels were used (see Table 8-(b)). Another personality trait that benefited from multi-modal fusion was neuroticism: classification with acquaintance labels yielded better results in the HRI setting. However, in the HRI setting, better results were achieved with individual features for conscientiousness when self labels were used (see Table 9-(a)), and for extroversion and neuroticism when the acquaintance labels were used (see Table 9-(b)).

For personality classification, combining multiple features was shown to provide the highest performance in [14] and [28]. On the other hand, [18] obtained the best results with the individual features. Overall, our personality classification results partially supported H3. However, the advantage of multi-modality over single-modality highly depends on the interaction setting, the labels used and the personality trait predicted.

Our engagement classification results did not support H3. Individual features always yielded better performance (see Table 10) for engagement. Although they modelled the perceived engagement, and applied fusion at the feature level, we further compared our results with the engagement classification results in [9]. In [9], individual features alone performed on a par with combining them with other features in the HRI setting (mean F-Score = 0.60). Similarly, as shown in Table 10, we obtained the best results using first person vision head motion signatures (FPV-HMS) only (mean F-Score = 0.59). In the HHI setting, audio features (AF) yielded by far the best results (mean F-Score = 0.65). We conjecture that this might be due to the fusion approach employed. Decision-level fusion has many advantages over feature-level fusion, such as ensuring easier scalability among multiple modalities, and enabling the usage of the most suitable method for modelling each modality individually. However, decision-level fusion fails to utilise the feature-level correlations among the modalities [46].

TABLE 10
Engagement classification with self labels: HHI versus HRI. The results are presented in terms of F Score; the best results with respect to the average performances (last column) are highlighted in bold. (AF: Audio Features, PF: Physiological features, FPV-HMS: FPV head motion signatures, FPV-LLD: FPV low level descriptors).

Features	HHI			HRI		
	low	high	ave.	low	high	ave.
AF	**0.58**	**0.72**	**0.65**	0.19	0.72	0.45
PF	0.52	0.57	0.54	0.48	0.15	0.31
FPV-HMS	0.41	0.48	0.45	**0.42**	**0.77**	**0.59**
FPV-LLD	**0.58**	**0.58**	**0.58**	0.21	0.70	0.45
AF+PF	**0.55**	**0.64**	**0.59**	0.40	0.51	0.45
AF+FPV-HMS	0.49	0.60	0.55	**0.31**	**0.74**	**0.53**
AF+FPV-LLD	**0.52**	**0.60**	**0.56**	**0.27**	**0.73**	**0.50**
PF+FPF-HMS	0.47	0.53	0.50	**0.46**	**0.53**	**0.50**
PF+FPV-LLD	0.51	0.54	0.52	**0.41**	**0.60**	**0.50**

6 Conclusions

In this paper we introduced the MHHRI dataset that consists of natural interactions between two participants and interactions between the same participants with a small humanoid robot. The MHHRI dataset complements the previous multimodal databases by incorporating six different data modalities, two interaction settings, and two sets of labels (i.e., personality and engagement). The MHHRI dataset also features recordings from a first person perspective in addition to the conventional third person perspective, and personality labels provided by acquaintances in addition to the self-assessments. Once this paper is published, the dataset will be made available to the research community[5].

In order to provide insights into the research questions studied in the scope of the MHHRI dataset, we presented baseline results for personality and engagement classification. We extracted a set of generic features from three modalities (i.e., audio, physiological, and RGB video), and applied Support Vector Machines (SVMs) for classification. We also combined the SVM outputs via decision-level fusion. Our results showed that (i) for predicting the same personality trait, different feature types / combinations were found to be more useful in the HRI setting as compared to the HHI setting, however performance trends with respect to the self and acquaintance labels remained the same across the HHI and HRI settings; (ii) for extroversion, openness and neuroticism, the acquaintance labels yielded the best results, while agreeableness and conscientiousness were better modelled with the self labels; and (iii) multi-modality, namely, combining features from different modalities, yielded better performance for the classification of personality traits in general. However, single modality always worked better for the classification of engagement.

Below we elaborate the lessons learned during the data collection and analysis, and provide some potential future directions.

6.1 Limitations

Altough the MHHRI dataset has a multitude of benefits over the existing multimodal databases, it has some shortcomings that are described below.

- *Small sample size.* The MHHRI dataset is rich in terms of the number of clips, however, it comprises a relatively small number of subjects. In other words, there are 746 short clips from a total of 18 subjects, where each subject has approximately 40 clips on average.
- *Order effect.* In our experiments, all the participants went through the same conditions in the same order, namely, a dyadic interaction between two human participants was always followed by a triadic interaction between two human participants and a robot. This might have affected the flow of the human-robot interactions, namely, the human participants might have adapted to the questions during the human-human interactions. In our follow-up studies related to personality, we always randomised the order of interactions [47].
- *Subject-level engagement annotations.* In our experiments, post-study questionnaire was completed by the participants after both the human-human and human-robot interactions took place. The interaction with the robot might have affected the perceived engagement during the interaction with the human partner. Apart from this, evaluating the whole interaction sequence by assigning a single value for the engagement state might not be adequate due to the reason that an individual's engagement state changes over time depending on the contextual factors, e.g., the question being asked. Therefore, a better strategy would be to collect engagement annotations at the clip-level, rather than subject-level, continuously in time and space, namely, by asking external observers to use a slider with a continuous scale and annotate the clips in a time-continuous manner as in [48].
- *Lens distortion.* The first-person (ego-centric) cameras that we used during the data collection were equipped with wide-angle lenses, which resulted in lens (barrel) distortion in the recordings (e.g., Figure 1-f). Although this distortion might not have implications for social interaction analysis, as the person that the camera wearer interacts with tends to be in the centre [49], a recent work showed that barrel distortion might hinder global motion estimation [50]. However, lens distortion can be corrected using an off-the-shelf technique (please refer to the Supplemental Material for details).

6.2 Future Directions

In this paper, we provided baseline results of human-human and human-robot interactions for personality and

[5] http://www.cl.cam.ac.uk/research/rainbow/projects/mhhri/

engagement classification as a proof of concept. We believe that the MHHRI dataset will be useful to the social signal processing community along multiple directions that can be summarised as below.

- *Multimodal fusion.* In this paper, we presented the classification results for combining two different modalities / feature types as we could not observe any improvement in the classification performance when more than two feature types were taken into account. More sophisticated fusion strategies based on, for instance, canonical correlation analysis (CCA) (see [51] for fusing face and body cues at the feature level using CCA) should be investigated to fully exploit the information coming from multiple modalities.
- *Joint modelling.* Zhang *et al.* [52] recently showed that joint prediction of self-reported emotion and perceived emotion improved the performance over the prediction of self-reported emotion only. They used multi-task feature learning algorithm in order to perform joint prediction. Our classification results also suggest that self labels and acquaintance labels are complementary to one another, and joint modelling might be useful to leverage the complementary information from both meta-data.
- *Contextual information.* Contextual information such as social context (e.g, HHI vs. HRI), identity (e.g., female or male), interpersonal context (e.g, other people's relationships with the target participant and their personalities) offers insights for the analysis of an individual's personality traits and engagement states. In this respect, one promising direction would be to examine the identity effects, e.g., gender. Another promising direction would be to investigate the interpersonal context. For example, in [9], we found that combining individual features (e.g., pose, body movement) with interpersonal features (e.g., visual focus attention, interpersonal distance) yielded slightly better results in the prediction of personality and classification of engagement. In [28], Fang *et al.* also showed that the performance significantly improved when individual features were combined with dyadic features and one-versus-all features that model the general relationship between the target participant's nonverbal behaviours and group behaviours. Using the MHHRI dataset, this can be further investigated by extracting interpersonal features such as synchrony from physiological signals, in addition to the audio-visual cues. Another interpersonal feature to be explored could be micro-action and reaction patterns from first-person videos as proposed in [53].
- *Temporal modelling.* We used static representations that were aggregated over the whole clip together with a generic classifier (e.g., SVM). However, interactions change over time, and are composed of a series of concurrent and sequential dynamics. Temporal classification methods such as Long Short-Term Memory Networks (LSTM) are powerful tools for modelling such temporal relationships as in [17].
- *Future directions.* The MHHRI dataset comprises human-human interactions and human-robot interactions of the same participants, which can be used to empirically study the differences between human-human interactions and human-robot interactions. An interesting research problem would be to explore what social phenomenon to study and which technique to use with the sparse data.

ACKNOWLEDGMENTS

This work was funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (Grant Ref: EP/L00416X/1). The authors would like to thank Kubra Cengiz for her help during the process of data collection.

REFERENCES

[1] J. K. Burgoon, L. K. Guerrero, and K. Floyd, *Nonverbal Communication*, Boston, MA: Allyn and Bacon, 2009.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing," *Image Vision Comput.*, vol. 27, no. 12, pp. 1743–1759, Nov. 2009.

[3] A. Ěerekoviæ, "An insight into multimodal databases for social signal processing: acquisition, efforts, and directions," *Artificial Intelligence Review*, vol. 42, no. 4, pp. 663–692, 2014.

[4] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[5] R. Cuperman and W. Ickes, "Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts disagreeables," *Journal of Personality and Social Psychology*, vol. 97(4), pp. 667–684, June 2009.

[6] A. Aly and A. Tapus, "A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human-robot interaction," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2013.

[7] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante, , and S. Escalera, *ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results*, pp. 400–418, 2016.

[8] O. Celiktutan and H. Gunes, "Computational analysis of human-robot interactions through first-person vision: Personality and interaction experience," in *IEEE Int. Sym. on Robot and Human Interactive Communication*, 2015, pp. 815–820.

[9] H. Salam, O. Celiktutan, I. Hupont, H. Gunes, and M. Chetouani, "Fully automatic analysis of engagement and its relationship to personality in human-robot interactions," *IEEE Access*, vol. PP, no. 99, pp. 1–1, 2016.

[10] M. L. Knapp and J. A. Hall, *Nonverbal Communication in Human Interaction*, Wadsworth: Cengage Learning, Bosten, USA, 2010.

[11] R. J. Larsen and T. K. Shackelford, "Gaze avoidance: Personality and social judgments of people who avoid direct face-to-face contact," *Personality and Individual Differences*, vol. 21, no. 6, pp. 907 – 917, 1996.

[12] R. E. Riggio and H.S. Friedman, "Impression formation: The role of expressive behavior," *J. of Personality and Social Psychology*, vol. 50, no. 2, pp. 421–427, 1986.

[13] R. Lippa, "The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: A lens model analysis," *J. of Res. in Personality*, vol. 32, no. 1, pp. 80–107, 1998.

[14] O. Aran and D. Gatica-Perez, "One of a kind: Inferring personality impressions in meetings," in *ACM Int. Conf. on Multimodal Interaction*, 2013.

[15] R. Subramanian, Y. Yan, J. Staiano, O. Lanz, and N. Sebe, "On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions," in *ACM Int. Conf. on Multimodal Interaction*, 2013.

[16] J. I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: predicting personality from facial expressions of emotion in online conversational video.," in *ACM Int. Conf. on Multimodal Interaction*, 2012.

[17] O. Celiktutan and H. Gunes, "Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability," *EEE Trans. on Affective Computing*, vol. 8, no. 1, pp. 29–42, Jan 2017.

[18] M. K. Abadi, J. A. M. Correa, J. Wache, H. Yang, I. Patras, and N. Sebe, "Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos," in *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2015.

[19] V. Rozgic, B. Xiao, A. Katsamanis, B. Baucom, P. Georgiou, and S. S. Narayanan, "A new multichannel multimodal dyadic interaction database," in *InterSpeech*, Sept. 2010.

[20] D. Sanchez-Cortes, O. Aran, and D. Gatica-Perez, "An audio-visual corpus for emergent leader analysis," in *ACM Int. Conf. on Multimodal Interaction Workshops*, 2011.

[21] A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven, "Cardiff conversation database (ccdb): A database of natural dyadic conversations," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 277–282.

[22] J. Vandeventer, A. J. Aubrey, P. L. Rosin, and A. D. Marshall, "4d cardiff conversation database (4d ccdb): a 4d database of natural, dyadic conversations," in *AVSP*. 2015, pp. 157–162, ISCA.

[23] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2013.

[24] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic, "The MAHNOB mimicry database: A database of naturalistic human interactions," *Pattern Recognition Letters*, vol. 66, pp. 52 – 61, 2015.

[25] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J. M. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, and D. Gatica-Perez, "The vernissage corpus: A conversational human-robot-interaction dataset," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*, March 2013, pp. 149–150.

[26] L. Devillers, S. Rosset, G. D. Duplessis, M/ A. Sehili, L. Bechade, A. Delaborde, C. Gossart, V. Letard, F. Yang, Y. Yemez, B. B. Turker, M. Sezgin, K. E. Haddad, S. Dupont, D. Luzzati, Y. Esteve, E. Gilmartin, and N. Campbell, "Multimodal data collection of human-robot humorous interactions in the joker project," in *Int. Conf. on Affective Computing and Intelligent Interaction*, Washington, DC, USA, 2015, pp. 348–354.

[27] S. Okada, O. Aran, and D. Gatica-Perez, "Personality trait classification via co-occurrent multiparty multimodal event discovery," in *ACM Int. Conf. on Multimodal Interaction*, 2015.

[28] S. Fang, C. Achard, and S. Dubuisson, "Personality classification and behaviour interpretation: An approach based on feature categories," in *ACM Int. Conf. on Multimodal Interaction*, 2016, ICMI 2016, pp. 225–232.

[29] F. Rahbar, S. M. Anzalone, G. Varni, E. Zibetti, S. Ivaldi, and M. Chetouani, "Predicting extraversion from non-verbal features during a face-to-face human-robot interaction," in *Social Robotics*, pp. 543–553. Springer, 2015.

[30] A. Fathi, J. K. Hodgins, and L. M. Rehg, "Social interactions: A first-person perspective," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2012.

[31] O. P. John and S. Srivastava, "Big five inventory (BFI)," *Handbook of personality: Theory and research*, vol. 2, pp. 102–138, 1999.

[32] B. Rammstedt and O. P. John, "Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German," *Journal of Research in Personality*, vol. 41, no. 1, pp. 203–212, 2007.

[33] A. D. Kenny, L. Albright, T. E. Malloy, and A. D. Kashy, "Consensus in interpersonal perception: Acquaintance and the big five," *Psychological Bulletin*, vol. 116, pp. 245 – 258, 1994.

[34] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affective Computing*, vol. 3, no. 1, pp. 5–17, Jan. 2012.

[35] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychology Bull.*, Jan. 1979.

[36] S. Buisine and J. C. Martin, "The influence of user's personality and gender on the processing of virtual agents' multimodal behavior," *Advances in Psychology Research*, vol. 65, pp. 1–14, 2009.

[37] S.-H. Kang, J. Gratch, N. Wang, and J. H. Watt, "Agreeable People Like Agreeable Virtual Humans," in *Lecture Notes in Computer Science*, 2008, pp. 253–261.

[38] A. Petland, "Social dynamics: Signals and behavior," Tech. Rep., MIT Media Laboratory, 2004.

[39] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *IEEE Int. Conf. on Computer Vision*, Dec. 2013, pp. 3551–3558.

[40] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., pp. 91–99. 2015.

[41] R. Yonetani, K. M. Kitani, and Y. Sato, "Ego-surfing first person videos," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2015, pp. 5445–5454.

[42] F. Crete, T. Dolmiere, P. Ladret, and M. Nicolas, "The blur effect: perception and estimation with a new no-reference perceptual blur metric," *Electronic Imaging*, vol. 6492, 2007.

[43] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 978–994, 2011.

[44] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *Europ. Conf. on Computer Vision*, 2010, ECCV'10, pp. 143–156.

[45] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[46] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 6, pp. 345–379, Nov. 2010.

[47] P. A. Bremner, O. Celiktutan, and H. Gunes, "Personality perception of robot avatar tele-operators in solo and dyadic tasks," *Frontiers in Robotics and AI*, vol. 4, pp. 16, 2017.

[48] O. Celiktutan and H. Gunes, "Continuous prediction of perceived traits and social dimensions in space and time," in *IEEE Int. Conf. on Image Processing*. IEEE, 2014, pp. 4196–4200.

[49] M.S. Ryoo and L. Matthies, "First-person activity recognition: What are they doing to me?," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 2013.

[50] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Biologically inspired motion encoding for robust global motion estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1521–1535, 2017.

[51] H. Gunes, C. Shan, S. Chen, and Y. Tian, "Bodily expression for automatic affect recognition," *Emotion Recognition: A Pattern Analysis Approach*, pp. 343–377, 2015.

[52] B. Zhang, G. Essl, and E. Mower Provost, "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?," in *ACM Int. Conf. on Multimodal Interaction*, 2016, ICMI 2016, pp. 217–224.

[53] R. Yonetani, K. M. Kitani, and Y. Sato, "Recognizing micro-actions and reactions from paired egocentric videos," in *IEEE Conf. on Computer Vision and Pattern Recognition*, June 2016.

Oya Celiktutan is a postdoctoral researcher in the Computer Laboratory, University of Cambridge, United Kingdom. She received her PhD degree in Electrical and Electronics Engineering from the Bogazici University, Istanbul, Turkey, in 2013. Her research interests centre around computer vision, machine learning and their applications to the areas of human-robot interaction, human-computer interaction, affective computing and personality computing.

Efstratios Skordos received his Diploma in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece in 2011. He worked as a research assistant at the Electrical Engineering Department of Imperial College London (2012-2014) and at the Multimedia and Vision Research Group of Queen Mary University London (2014). He is currently a PhD student in the Computer Science Department of University College London. His research interests are computer vision and machine learning.

Hatice Gunes (SM'16) is currently an Associate Professor (Senior Lecturer) in the Department of Computer Science and Technology, University of Cambridge, UK. Her research expertise is in the areas of affective computing, social signal processing, human behaviour analysis, and human robot interaction. She is the President-Elect of the Association for the Advancement of Affective Computing (AAAC) and the Chair of the Steering Board of IEEE Transactions on Affective Computing. Dr Gunes is an Associate Editor of the IEEE TRANSACTIONS on AFFECTIVE COMPUTING, the IEEE TRANSACTIONS on MULTIMEDIA, and the Image and Vision Computing Journal. She is the General Co-Chair of ACII19 and the Program Co-Chair of IEEE FG17.