CHAPTER 10

Computational Analysis of Affect, Personality, and Engagement in Human–Robot Interactions*

Oya Celiktutan*, Evangelos Sariyanidi†, Hatice Gunes‡
*Imperial College London, Electrical Engineering, Personal Robotics Lab, United Kingdom
†Centre for Autism Research, Philadelphia, PA, United States
‡University of Cambridge, Computer Laboratory, United Kingdom

Contents

Abstract

This chapter focuses on recent advances in social robots that are capable of sensing their users, and support their users through social interactions, with the ultimate goal of fostering their cognitive and socio-emotional wellbeing. Designing social robots with socio-emotional skills is a challenging research topic still in its infancy. These skills are important for robots to be able to provide physical and social support to human users, and to engage in and sustain long-term interactions with them in a variety of application domains that require human–robot interaction, including healthcare, education, entertainment, manufacturing, and many others. The availability of commercial

* The research reported in this chapter was completed while O. Celiktutan and E. Sariyanidi were with the Computer Laboratory, University of Cambridge, United Kingdom.

Computer Vision for Assistive Healthcare.
DOI: https://doi.org/10.1016/B978-0-12-813445-0.00010-1

robotic platforms and developments in collaborative academic research provide us with a positive outlook; however, the capabilities of current social robots are quite limited. The main challenge is understanding the underlying mechanisms of humans in responding to and interacting with real life situations, and how to model these mechanisms for the embodiment of naturalistic, human-inspired behavior in robots. Addressing this challenge successfully requires an understanding of the essential components of social interaction, including nonverbal behavioral cues such as interpersonal distance, body position, body posture, arm and hand gestures, head and facial gestures, gaze, silences, vocal outbursts, and their dynamics. To create truly intelligent social robots, these nonverbal cues need to be interpreted to form an understanding of the higher level phenomena including first-impression formation, social roles, interpersonal relationships, focus of attention, synchrony, affective states, emotions, personality, and engagement, and in turn defining optimal protocols and behaviors to express these phenomena through robotic platforms in an appropriate and timely manner. This chapter sets out to explore the automatic analysis of social phenomena that are commonly studied in the fields of affective computing and social signal processing, together with an overview of recent vision-based approaches used by social robots. The chapter then describes two case studies to demonstrate how emotions and personality, two key phenomena for enabling effective and engaging interactions with robots, can be automatically predicted from visual cues during human–robot interactions. The chapter concludes by summarizing the open problems in the field and discussing potential future directions.

Keywords

Social robotics, Human–robot interaction, Affective computing, Social signal processing, Personality computing, Computer vision, Machine learning

10.1 INTRODUCTION

Humanoid robots are being deployed in public spaces including hospitals [1], banks [2], and airports [3]. An increasing number of individuals needing companionship and psychological support push the need for socially assistive robotics. Socially assistive robotics focuses on building robots that can facilitate an effective interaction with their human users for the purpose of assisting them at the social and cognitive level, namely, aiding them to achieve their goals, managing their medical needs, or enhancing their overall well-being. In the context of heath care and therapy, there is a significant body of work on how Paro, a robotic seal, improves well-being and reduces depression and anxiety in elderly people [4]. KASPAR, Kinesics And Synchronization in Personal Assistant Robotics, is a child-sized humanoid robot designed to develop basic social interaction skills in children with autism through turn taking and imitation games [5]. SPRITE, Stewart Platform Robot for Interactive Tabletop Engagement, helps a group of people to complete a task by manipulating turn-taking patterns and the participants' attention, with the goal of increasing group cohesion [6].

In education, several studies have already shown the benefits of using robots in one-to-one tutoring sessions and classroom settings. Students performed better in mathematics when a robot tutored them [7], and were more emotionally expressive when engaged in an interactive educational task with a social robot than when performing the same task with a tablet [8]. Personalizing a robot's actions to individual differences has been shown to be compulsory for achieving good learning outcomes in several studies. Keepon, a tabletop robot, was made to provide personalized feedback using a skill assessment algorithm in [9]. To accommodate children's short attention spans, Nao, a child-sized humanoid robot, was programmed to offer breaks based on personalized timing strategies [10]. Similarly, in [11], Nao tutored language learning by adapting its feedback to the children's skills and observed behaviors.

User modeling, adaptation, and personalization are key to the effective deployment of social robots in real-world settings. The generic system of such a robot consists of three modules [12]: (1) the perception module; (2) the reasoning (intermediate) module; and (3) the action module. The perception module acquires information regarding the human user by capturing (multimodal) data through both the robot's sensors and the environmental sensors, and analyzes the human user's behaviors based on the information collected during interactions. The action module deals with the design and generation of behaviors for the robot. The reasoning (intermediate) module connects the perception and action modules to deliver robot behaviors that are shaped by the output of the perception module. In this chapter, we exclusively focus on the perception module, in particular from the perspective of affect and social signal analysis from visual cues.

Affective and social signals are integral parts of communication. 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 behaviors [13]. These nonverbal behaviors carry significant information regarding higher level social phenomena such as emotions, personality, and engagement. Recognizing and interpreting these signals is a natural routine for humans, and automatizing these mechanisms is necessary for robots to be successful in their interactions with humans.

The objective of this chapter is to present a survey of computational approaches to the analysis of affective and social signals, together with recent techniques used by social robots, to categorize the available algorithms and to highlight the latest trends. The chapter starts with representative techniques for the analysis of an individual's emotions, personality, and engagement state, three social phenomena that have been commonly studied in the area of affective and social signal processing (see Section 10.2). The chapter then focuses on summarizing the state of the art of robotic platforms endowed with the capability of analyzing these social phenomena. To provide concrete examples, the chapter presents two case studies to describe how a computational method can be built for predicting emotions and personality from visual cues during human–robot interactions (see Section 10.3). The chapter concludes by summarizing the open problems in the field and discusses potential solutions to these problems (see Section 10.4).

10.2 AFFECTIVE AND SOCIAL SIGNAL PROCESSING

In this section, we first introduce the state-of-the-art computer vision-based approaches to affective and social signal processing, and then review the prominent techniques used by the currently available social robots. We scope out and explore three social phenomena that are widely studied in this context: (i) emotion; (ii) personality; and (iii) engagement.

10.2.1 Emotion

Emotion (or affect) recognition has been one of the most active research areas across multiple disciplines ranging from psychology to computer science and social robotics. There have already been several extensive surveys on automatic emotion recognition from facial cues [14,15] and bodily cues [16].

Emotion recognition methods from facial cues aim at recognizing the appearance of facial actions or the expression of emotions conveyed by these actions, and usually rely on the Facial Action Coding System (FACS) [17]. FACS consists of facial Action Units (AUs), which are codes that describe certain facial muscle movements (e.g. AU 12 is lip corner puller). The temporal evolution of an expression is typically modeled with four temporal phases [17]: neutral, onset, apex, and offset. Neutral is the expressionless phase with no signs of muscular activity. Onset denotes the period during which muscular contraction begins and increases in intensity. Apex is a plateau where the intensity usually reaches a stable level. Offset is the phase of muscular action relaxation.

There have been two lines of approaches proposed in the literature that are associated with two models of emotions, namely, the categorical model and the dimensional model. The categorical model refers to the affect model developed by Ekman and his colleagues, who argued that the production and interpretation of certain expressions are hard-wired in our brains and are recognized universally (e.g. [18]). The emotions conveyed by these expressions are grouped into six classes, known as the *six basic emotions*: happiness, sadness, surprise, fear, anger, and disgust. AUs can be mapped to the six basic emotions. For example, using a simple rule-based method, happiness can be represented as a combination of AU6 (cheek raiser) and AU12 (lip corner puller) [14]. However, the categorical model is believed to be limited in its ability to represent the broad range of everyday emotions [19]. To represent a wider range of emotions, the dimensional approach is used to continuously model emotions in terms of affect dimensions [19]. The most established affect dimensions are arousal, valence, power, and expectation [19].

The categorical and dimensional models were evaluated in two prominent affect recognition challenges: Facial Expression Recognition and Analysis (FERA) [20,21] and Audio/Visual Emotion Challenges (AVEC) [22]. The FERA challenge evaluates AU detection/AU intensity estimation and discrete emotion classification for four basic emotions (anger, fear, happiness, sadness) and one nonbasic emotion (relief). The AVEC challenge evaluates dimensional emotion models along arousal and valence dimensions.

De la Torre et al. [23] addressed the AU detection problem using a personalized learning approach based on a Selective Transfer Machine (STM) that learns a classifier while simultaneously reweighting the training samples that are most relevant to the test subject. They extracted appearance features based on Scale-Invariant Feature Transform (SIFT) descriptors, from patches centered on the automatically detected facial landmarks. The proposed method achieved superior performance compared to the conventional classification methods such as Support Vector Machines (SVMs) for classifying five emotions on the FERA 2011 benchmark [20]. The recent trend for AU detection has been deep learning methods. Jaiswal and Valstar [24] simultaneously learned dynamic appearance and shape features within a time window using Convolutional Neural Networks (CNNs), and applied Bidirectional Long Short-Term Memory (BLSTM) networks on top of the time-windowed CNN features to model temporal relationships. The proposed method outperformed the previous approaches in the FERA 2015 challenge datasets [21].

Recent works adopting the dimensional model were characterized by combining visual data with different modalities, usually audio and physiological data, and employing BLSTM for predicting arousal and valence in a time-continuous manner [25,26]. For example, the winner of the AVEC 2015 challenge [27] combined two appearance features, namely, Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP), which were baseline features provided by the challenge organizers, and Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) together with geometric features computed from facial landmarks. Different feature types were fused using a model-level fusion strategy, where the outputs of single modality models were smoothed and combined using a second layer of BLSTM. Chen and Jin [26] proposed a multimodal attention fusion method that automatically assigns weights to different modalities according the current modality features and history information, which outperformed the traditional fusion strategies (e.g. early-fusion, model-level fusion, late-fusion) in the detection of valence in the same database.

Emotion Recognition in HRI. Emotion recognition methods used by social robots were extensively surveyed by Yan et al. in [12] and McColl et al. in [28]. Here, we only considered the prominent works that performed the recognition task by automatically extracting features from visual cues, and integrated the developed method on a robotic platform.

The categorical model of emotion has been the most widely adopted approach in the literature. Cid et al. [29] developed an emotion recognition system by extracting features based on the Facial Action Coding System (FACS) [17], and implemented it on a robotic head, Muecas [30], for an imitation task. For emotion recognition, they first applied a preprocessing step to the face image taken by Muecas to normalize the illumination and remove the noise, and a Gabor filter to highlight the facial features. From the processed face, a set of edge-based features were extracted and modeled using Dynamic Bayesian Networks to detect a total of 11 AUs. The detected AUs were used to represent four basic emotions including happiness, sadness, fear, and anger according to a rule-based approach, and were mapped on the Muecas robot to display the inferred emotion in real-time. In [31], the authors used similar visual features (i.e. Gabor filter responses) to enable the robot to learn facial expressions of emotion from interactions with humans through an online learning algorithm based on neural networks. The Muecas robot was able to learn all the emotions successfully, except for sadness. This was due to the large intra-class variability for sadness, namely, each person expressed sadness in a different manner.

In [32], Leo et al. developed an automatic emotion recognition system to measure the facial emotion imitation capability of children with Autism Spectrum Disorders (ASD). The R25 robot from Robokind [33], a small cartoon character-like robot, was first made to display a facial expression, and then the child was instructed to imitate the displayed facial expression while being analyzed through the camera located in R25's right eye. The emotion recognition method was based on a generic pipeline that consisted of four components: Viola–Jones face detection, face registration, Histogram of Gradient (HoG) face representation, and classification with SVMs. The method was tested via a study involving three children with ASD, and it achieved good emotion recognition performance, especially for happiness and sadness.

Among works adopting the dimensional model of emotion, Castellano et al. [34] focused on valence of an affect, representing it with three discrete states: positive, neutral, and negative. They designed an affect-sensitive robotic game companion, with the goal of detecting these three states and selecting an empathic strategy for the robot to display. For this purpose, they combined visual features including smiling gestures and gaze patterns with contextual information such as game state and game evolution. For detecting smiles, first an off-the-shelf application was used to estimate head pose and track facial landmark points, and then a geometry-based descriptor was defined based on the spatial locations of the facial landmarks with respect to the head pose. The developed method was integrated onto the iCat platform, a desktop user-interface robot with animated facial expressions [35] to test with children during the course of a chess game. Schacter et al. [36] focused on the prediction of both arousal and valence dimensions. They extracted geometry-based features from facial landmarks that were detected using Constrained Local Models [37], and applied Support Vector Regression (SVR) for prediction. The proposed method was tested using the onboard camera of their in-house robot called Social Robot Brian.

In this chapter, we exclusively focus on facial cues. However, body postures and hand gestures are important sources of information, especially in the context of HRI, when facial cues cannot be observed reliably. Most of the emotion recognition methods from bodily cues has relied on real-time skeleton tracking algorithm of Kinect depth sensor [38]. Wang et al. [39]

aimed at modeling arousal and valence dimensions in a time-continuous manner. They captured visual recordings using a Kinect depth sensor during the course of a game of Snakes and Ladders played by a child against the Nao robot [40]. Nao's behaviors were manipulated to display either competitive or supportive behaviors in order to elicit different emotional responses from the participated children. From these recordings, they modeled bodily expressions using the 3D skeleton tracking algorithm, and skeletal representations were used to extract two types of features: (i) a set of low-level features comprising spatial distances between hands, elbows, and shoulders, and the angles between the spine and the upper arms, and the orientation of the shoulders; (ii) a set of high-level features describing body movement activity and power, body spatial extension, and head bending. These features were then used to train Online Recursive Gaussian Processes for real-time emotion recognition from bodily expressions, where they found that the valence dimension was more difficult to model than the arousal dimension.

10.2.2 Personality

Individuals' interactions with others are shaped by their personalities and their impressions regarding others' behaviors and personalities [41]. This has also been shown to be the case for interactions with social robots [42]. The traditional approach to describing personality is the trait theory that focuses on the measurement of general patterns of behaviors, thoughts, and emotions, which are relatively stable over time and across situational contexts [43]. The Big Five Model is currently the dominant paradigm in personality research which defines traits along five broad dimensions: *extroversion* (assertive, outgoing, energetic, friendly, socially active), *neuroticism* (a tendency to negative emotions such as anxiety, depression, or anger), *openness* (a tendency to changing experience, adventure, new ideas), *agreeableness* (cooperative, compliant, trustworthy), and *conscientiousness* (self-disciplined, organized, reliable, consistent).

There are two strategies coupled with two main problems in automatic personality analysis [44], which are personality recognition (prediction of actual personality) and personality perception (prediction of personality impressions). In both problems, the commonly used method to measure Big Five personality traits is the Big Five Inventory (BFI) [45]. In personality recognition, an individual is asked to fill in the BFI which aims to assess personal behavioral tendencies, i.e. how an individual sees herself in the way she approaches problems, likes to work, deals with feelings, and manages relationships with others. In personality prediction, external observers are asked to view a video of the individual and rate the individual along the Big Five personality dimensions based on thin slices of behavior ranging from 10 seconds to several minutes. However, employing observers to carry out this tedious task is in itself a problem. A number of researchers [46,47] obtained manual annotations through online crowd-sourcing services such as the Amazon Mechanical Turk (MTurk) service. Typically, several folds of independent ratings are run since there is rarely full agreement between the raters.

Nonverbal behaviors are significant predictors of personality. Gaze and head movement are 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 [48]. Extroverted people are found to be more energetic, leading to higher head movement frequency, more hand gestures, and more posture shifts than introverted people [49,50]. Research has demonstrated that these nonverbal behaviors can be reliably modeled from visual cues for predicting personality.

Among the works focusing on facial and head cues, Joshi et al. [51] investigated varied situational contexts using audio-visual recordings of conversations between a human and four different virtual characters using the SEMAINE corpus [52]. The SEMAINE corpus comprises audio-visual recordings of interactions between human participants and four different virtual characters. Facial cues were extracted using the pyramid of HoG, which counts the gradient orientations in the whole face and in the localized portions. The mean and the standard deviation of the histograms accumulated from all the frames were fed into SVMs for regression. The visual features used yielded the best prediction accuracy for *conscientiousness* among the Big Five personality traits.

High-level features were taken into account by Teijeiro-Mosquera et al. [46] using videos from Youtube, so-called “video blogs”, with annotations generated through the MTurk service. They detected facial expression of emotions (e.g. anger, happiness, fear, sadness) on a frame-by-frame basis and extracted emotion activity cues from sequences either by thresholding or by using an HMM-based method. These features were then fed into SVMs for predicting the five traits. Their results showed that facial expressions were a strong predictor of *extroversion*.

Another line of work has focused on the fusion of facial/head cues and bodily cues at the feature level. Aran and Gatica-Perez [53] used recordings from the ELEA corpus [54] involving three or four participants performing a Mission Survival task [55]. They represented the visual cues by extracting two types of features, namely, activity features and attention features. The participants' heads and bodies were tracked in videos, and optical flow was computed from tracked head and body parts, yielding the binary occurrence of head/body activity at a specific time instant and the amount of activity. Activity features were then computed by aggregating the occurrences and amount of activity over the whole sequence, which included head/body activity length, head/body activity turns, standard deviations of head/body activity in x and y directions, etc. In addition to head/body activity features, simple statistics were calculated from weighted Motion Energy Images (MEI) in order to encapsulate the whole body activity over time. Attention features were extracted based on the visual focus of attention analysis during interactions, which included attention given while speaking/listening, attention received while speaking/listening, and visual dominance ratio. Ridge regression was used both for the prediction of *extroversion* level and for the binary classification of *extroversion*, *agreeableness*, and *openness*. For both regression and classification, the best results were achieved by combining all the features. However, the prominent visual features were attention features and MEI statistics in the classification of *extroversion*.

From human–virtual character interactions [52], Celiktutan and Gunes [56] modeled the face/head and body movements by extracting three sets of features: (i) spatial and spatio-temporal appearance features (e.g. Zernike moments, gradient and optical flow); (ii) geometric features (e.g. spatio-temporal configuration of facial landmark points, horizontal and vertical trajectories over time); and (iii) hybrid features (e.g. the fusion of local appearance and motion information around facial landmark points). These features were then used in conjunction with Long Short-Term Memory Networks for predicting personality traits continuously in space and time, which yielded the highest coefficient of determination (R^2) for *conscientiousness* using the face appearance features and for *neuroticism* and *openness* using the body appearance features.

Personality Prediction in HRI. Incorporating human personality analysis to adapt a robot's behavior for engaging a person in an activity is a fundamental component of social robots [57,47]. One prominent work by Rahbar et al. [58] focused on the prediction of the *extroversion* trait only, when a participant was interacting with the humanoid iCub [59], a robot shaped like a four-year-old child. They combined individual features and interpersonal features that were extracted from Kinect recordings. More explicitly, the individual features included the participant's quantity of motion computed from the depth images. The interpersonal features modeled synchrony and dominance between the movements of iCub and the participant, and also proxemics (i.e. the distance between iCub and the participant). They achieved the best F-measure when they fused individual and interpersonal features at the feature level using Logistic Regression.

Some works focused on the robot's personality to improve the quality of the human experience with the robot: humans tend to be attracted by characters that have either matching personality traits (similarity rule) or non-matching personality traits (complementarity rule) [60]. Salam et al. [47] investigated the impact of the participants' personalities on their engagement states from the Kinect depth sensor recordings. These recordings contained interactions between two participants and Nao [40], a small humanoid robot. They extracted two sets of features, namely, individual and interpersonal features, similarly to [58]. Individual features described the individual behaviors of each participant, e.g. body activity computed from articulated pose and motion energy images, body appearance, etc. Interpersonal features characterized the interpersonal behaviors of the participants with respect to each other and the robot. These include the visual focus of attention (VFOA), the global quantity of movement, the relative orientation of the participants, the relative distance between the participants, and the relative orientation of the participants with respect to the robot. 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. The best results were achieved using individual features together with personality labels.

Despite its importance, automatic personality analysis as a part of a social robot has been scarce; indeed, to the best of our knowledge, there has been no system that is integrated onto a robot, and performs real-time analysis of personality in the course of interaction. In [61], Celiktutan et al. used a real-time implementation of their method of personality prediction from nonverbal cues [56], and demonstrated this system, called MAPTRAITS, together with the Nao robot. Using a Wizard of Oz setup, Nao asked the participants a predefined set of questions about their jobs, hobbies, and memories while the MAPTRAITS system (running on a PC) analyzed the participants' personalities in real-time using a camera placed on a tripod.

The predicted personality scores were displayed to each participant instantaneously on a screen; however, no quantitative analysis was conducted.

10.2.3 Engagement

Engagement is the process by which interactors start, maintain, and end their perceived connection to each other during an interaction [62]. When individuals interact with each other, they display affective and social signals that give away information regarding their engagement states (i.e. intention to engage, engagement, and disengagement).

Most of the methods for predicting engagement have focused on observable visual cues including social gaze patterns, facial gestures, and body posture. Although these cues were manually annotated, Kapoor et al. [63] exploited features based on facial gestures and body posture in order to predict the level of interest of a child who was solving a puzzle. In particular, facial gestures were coded in terms of manually annotated facial action units associated with upper face muscle movements around the eyes, eyebrows, and upper cheeks, and body posture was determined using a sensor chair. Their results showed that body posture alone was more informative than facial gestures, yielding a better classification performance with Hidden Markov Models (HMMs). Oertel and Salvi [64] only relied on features extracted from manually annotated social gaze patterns to model group involvement and individual engagement in game-based group interactions. They divided the social gaze patterns into four groups, namely, looking at another participant, looking away, looking down, and eyes closed, that were converted into a matrix for each participant. They then extracted group-level features and individual-level features from these matrices. While group-level features modeled interpersonal dynamics such as mutual gaze, individual features intended to capture individual differences in gaze behaviors. Good classification results were obtained with Gaussian Mixture Models (GMMs) for detecting the high level of group involvement and group forming/getting familiar with each other, whereas the low-level group involvement was classified poorly.

Peters et al. [65] focused on automatic gaze estimation and shared attention detection from a web camera during interactions with a virtual agent. They first estimated head pose and gaze by automatically detecting and tracking facial landmark points. The user's head and gaze directions were then mapped on the computer screen in order to model the level of attention and the level of engagement. While the level of attention was measured in terms of gaze fixations onto the virtual objects on the screen (including the virtual agent itself), the scene background, or outside of the screen, the level of engagement was defined as how much the user looks at the relevant objects in the scene at the appropriate times.

There is another line of research investigating the impact of personality on engagement in human–virtual character interactions. Cerekovic et al. [66] considered two virtual agents from the SEMAINE System [52], namely, Obadiah and Poppy. While Obadiah was gloomy and neurotic with low variation in speech and a flat tone, Poppy was cheerful and extroverted with frequent gestures and head nods. They measured the engagement level of each participant along three dimensions: quality, rapport, and likeness. In order to predict the levels of these three dimensions, they took into account both audio-visual features and manually annotated personality trait labels collected from external observers. As visual features, they computed the distribution of body leans and frequency of shifts from one body posture to another using the 3D skeleton tracking information from the Kinect depth sensor. Similar features were computed for manually annotated hand gestures, and facial expressions were modeled using an off-the-shelf facial expression recognition toolbox. They achieved the best results when they combined nonverbal features with personality scores. They found that extroverted people tended to like the neurotic agent, whereas people that score high on *neuroticism* liked the cheerful agent, supporting the interpersonal complementarity rule [60].

Engagement Prediction in HRI. Understanding the user's engagement is important to ensure that the user maximally benefits from an activity conducted with the assistance of the robot, particularly in health-related applications and education settings. In [67], Sanghvi detected engagement states during a chess game played by a child and iCat [34]. In order to detect whether the child was engaged or not, the child's body silhouette was first extracted, and then a set of features was extracted based on the posture and body movements. These features included (i) body lean angle, a measure of the orientation of the child's upper body when playing the game with the robot; (ii) slouch factor, a measure of the curvature of the child's back; (iii) quantity of motion, a measure of the amount of detected motion from the extracted silhouette; and (iv) contradiction index, a measure of the degree of contraction and expansion of the upper body. Using the extracted features in conjunction with ADTree and OneR classifiers yielded a high accuracy for engagement classification.

In [68], Benkaouar and Vaufreydaz proposed a multimodal approach for recognizing nonverbal cues and inferring engagement in a home environment where they used a Kinect depth sensor mounted onto a mobile robot called Kompai from Robosoft [69]. They extracted three sets of visual features: (i) proxemics features such as distance to the robot, speed from the recorded depth data; (ii) face location and face size from the recorded RGB data; and (iii) positions of stance, hips, torso, and shoulders, and their relative rotations from the tracked skeletons. The most relevant features yielding the best engagement detection accuracy were selected using the Minimum Redundancy Maximum Relevance method. Their results showed that shoulder rotation, face position and size, relative distance, and speed played an important role in engagement detection.

Salam and Chetouani [70] conducted a study in a triadic HRI scenario to investigate to what extent it is possible to infer an interactor's engagement state starting from the cues of the others in the interaction. They considered two set of features from two human participants and a robot. Each participant's features were composed of manually annotated social cues including head nods, visual focus of attention (VFOA), head pose, face location, and utterances. In addition to these cues, they extracted simple features over time, e.g. VFOA shifts, sliding windows statistics of head pose and face location, etc. The robot's features comprised utterances, addressee (addressing the speech to an interactor), and the topic of the speech. These features were used, both singly and in pairwise combinations (i.e. combining features of both participants, or combining a participant's features with the robot's features), in conjunction with SVMs for engagement classification. Their results showed that in a multiparty interaction, the cues of the other interactors can be used to infer the engagement state of the individual in question, which suggests that inter-personal context plays an important role in engagement classification.

10.3 TWO CASE STUDIES

In this section, we describe two automatic methods for modeling emotion and personality in interactions with a robot. First, we present a novel AU detection method. AU detection has been a popular research problem in computer science; however, there are fewer works performing AU detection in the context of HRI. Differently from [30], for more robust AU detection, our method combines shape and appearance information, and exploits differential features with respect to an individual's neutral face. Then, we introduce how this method can be implemented on the humanoid robot Nao in real-time and can be used in live public demonstrations.

Second, we describe a pipeline for automatic prediction of an individual's personality in the course of their interactions with Nao, from experimental study design to data collection and feature extraction. Despite its importance, there are only a few works performing automatic personality prediction in the context of HRI. Additionally, most of these works investigate the relationship between the personality traits and engagement state based on self reports, which might not be available in real-life applications. Here, we show that personality can be predicted from a set of low-level features extracted from videos captured from a first-person perspective.

10.3.1 Automatic Emotion Recognition

In this chapter, we introduce a novel method for detecting Action Units (AUs) in video sequences, and present comparative figures on a state-of-the-art database. We also demonstrate how this approach can be used for public engagement at various events (e.g. science festivals).

10.3.1.1 Action Unit Detection Methodology

There has been a significant body of work in the area of automatic AU detection. Recently, Sariyanidi et al. [15] highlighted the importance of two practices: (i) combining shape and appearance features, which yields better performance because they carry complementary information, and (ii) using differential features that describe information with respect to a reference image (i.e. a neutral face in the case of emotion recognition). The main advantage of the differential features is to place greater emphasis on the facial action by reducing person-specific appearance cues.

Feature Extraction. In the light of these insights, we extracted four types of features, namely, shape, appearance, differential-appearance (hereafter δ-appearance) and differential-shape (hereafter δ-shape) as follows. Shape features were obtained by concatenating the vertical and horizontal coordinates of the facial landmarks that were estimated using the Supervised Descent Method (SDM) in [71]; δ-shape features were computed by subtracting the shape representation of a given facial image from the shape representation that was computed from a facial image, of the same subject, with a neutral expression.

Appearance features were extracted using the Quantized Local Zernike Moments (QLZM) method [72]. The use of this method was previously demonstrated for affect recognition based on both the categorical and dimensional models of emotion [72]. The QLZM method consists of two steps: (i) computing local Zernike moments to describe image discontinuities at various scales and orientations, and (ii) performing non-linear encoding and pooling to improve the robustness against image noise and translation. Here we computed the appearance features in a part-based manner. Using the estimated facial landmarks, we first cropped three square patches that contained the left eye, right eye, and mouth, and then computed the QLZM histograms from each part.

We computed δ-appearance features using the Gabor motion energy filters [73], where we adopted a part-based representation similarly to the appearance features. We used Gabor motion energy filters to describe the motion between a given face image of a subject and the subject's neutral face image. One advantage of using the Gabor representation over using simpler representations (e.g. difference between neutral and apex phases) is its robustness to illumination variations. During the on-the-fly tests, we ensured that we had the neutral face of human subjects by asking them to stand still and make a neutral face in front of the camera prior to beginning a test session.

Note that each AU can occur either in the upper part or in the lower part of the face. For example, AU1 (inner brow raiser) occurs in the upper part and AU25 (lips part) occurs in the lower part. Therefore, when detecting an AU, we took into account the above-mentioned four features extracted either from the upper part or from the lower part of face. For shape and δ-shape features, this resulted in a 60-length feature vector, corresponding to the landmarks associated with eyes and eyebrows, and a 38-length feature vector, corresponding to the landmarks associated with mouth. For the appearance and δ-appearance features, this resulted in a 800-length and a 512-length feature vector, respectively, computed from the left and right eye patches, and a 400-length and a 256-length feature vector, respectively, computed from the mouth patch.

Decision Fusion. We trained four binary SVM classifiers, each in conjunction with one of the above-mentioned feature types, for each AU. The final AU detection decision was given by fusing the outputs of the four individual classifiers. Specifically, we adopted the *consensus fusion* approach, where an AU was detected based on the condition that all four classifiers were in full agreement. The advantage of the consensus fusion approach is that it yields a low False Alarm Rate (FAR). The downside is that it can also lead to a low True Positive Rate (TPR) because the consensus cannot be reached even when one of the classifiers misses an AU. To address this issue, we decreased the AU detection threshold for each classifier, where we empirically set the threshold to 0.95 TPR on the training dataset. This also increased the False Positive Rate (FPR) of the individual classifiers, but the overall FPR of the consensus fusion approach was low, as shown in the next section.

10.3.1.2 Experimental Results

In this work, we focused on a total of seven AUs, namely, inner brow raiser (AU1), outer brow raiser (AU2), brow lowerer (AU4), cheek raiser (AU6), lip corner puller (AU12), lips parted (AU25), and jaw drop (AU26). For these AUs, we evaluated the performance of the proposed AU detection pipeline using the MMI Facial Expression dataset [74], one of the most widely used benchmark datasets in the field.

Experimental Setup. For each AU, we trained an SVM classifier using the one-vs-all approach, namely, positive samples were the images where the AU was displayed, and the negative samples were all the other images where the AU was not displayed, including neutral samples. We used a linear c-SVM [75] and fixed the c parameter to $c = 10^{-3}$.

We used the MMI Facial Expression [74] database, which contains a total of 329 video sequences with annotations provided for the temporal segments of onset, apex, and offset. In order to increase the number of training samples, we selected multiple frames from the apex segment. Subjects often displayed eye movements or small head movements; therefore, the frames extracted from the apex segment were not identical. Similarly, in order to create negative samples, for δ-appearance and δ-shape representations, we randomly picked pairs of frames with neutral expressions. This resulted in a total of 6349 training samples; however, some AUs (e.g. AU1, AU12) have a relatively small number of samples. We handled the data imbalance issue by limiting the number of negative samples. More explicitly, for each AU, we formed 20% of the training samples from the positive samples, 40% from the negative samples with neutral faces, and 40% from the negative samples with nonneutral faces.

Results. We evaluated AU detection performance using five-fold subject-independent cross validation. Table 10.1 presents AU detection results with respect to the four individual features, and their combination via the consensus fusion approach in terms of (a) the alternative forced choice (2AFC) metric [76], (b) the TPR, and (c) the FPR. The 2AFC metric can be defined as the area A underneath the receiver-operator characteristic (ROC) curve, and an upper bound for the uncertainty of the A statistic for n_p positive and n_n negative samples, $s = \sqrt{A(1-A)/\min\{n_p, n_n\}}$.

Table 10.1 AU detection performance in terms of (a) the alternative forced choice (2AFC) score, (b) the false positive rate (FPR), and (c) the true positive rate (TPR). Bold text indicates the best (i.e. highest) score

	AU1	**AU2**	**AU4**	**AU6**	**AU12**	**AU25**	**AU26**
(a) 2AFC							
Shape	0.74	0.53	0.67	0.61	0.79	0.73	0.53
Appearance	0.74	0.73	0.65	0.78	0.82	0.78	0.67
δ-shape	0.78	0.67	0.71	0.74	0.78	0.82	0.64
δ-appearance	0.90	**0.92**	**0.87**	0.82	0.92	**0.89**	0.78
Fusion	**0.91**	0.89	0.78	**0.87**	**0.93**	0.86	**0.79**
(b) FPR							
Shape	0.41	0.87	0.49	0.77	0.40	0.44	0.77
Appearance	0.45	0.46	0.50	0.35	0.31	0.32	0.58
δ-shape	0.41	0.62	0.46	0.42	0.45	0.30	0.51
δ-appearance	0.15	0.12	0.21	0.28	0.12	0.17	0.35
Fusion	**0.02**	**0.03**	**0.04**	**0.12**	**0.06**	**0.02**	**0.11**
(c) TPR							
Shape	0.89	0.93	0.82	1.00	0.98	0.90	0.83
Appearance	0.92	0.92	0.80	0.90	0.94	0.88	0.93
δ-shape	0.98	0.96	0.87	0.90	1.00	0.93	0.79
δ-appearance	0.96	0.96	0.95	0.91	0.96	0.95	0.91
Fusion	0.84	0.81	0.61	0.86	0.92	0.73	0.68

Looking at the AFC scores (Table 10.1(a)), the best performing individual feature is the δ-appearance feature, and the consensus fusion achieves a higher AFC score than the δ-appearance feature for four AUs (AU1, AU6, AU12, AU26) out of seven AUs. The main advantage of the consensus fusion is the low FPR, as given in Table 10.1(b) (the corresponding TPRs are provided in Table 10.1(c)). We also used the best performing trained models in the real-time demonstration.

Real-Time Demonstration. We performed the real-time implementation using C++. For the initial face detection in each session, we used the Viola–Jones face detector [77] and then tracked the face using the SDM method [71]. We redetected the face when tracking failed. The real-time implementation was integrated onto the Nao robot as shown in Fig. 10.1. The computational power of the Nao robot did not allow us to run the AU detection algorithm in real-time. For this reason, we used a pair of external cameras plugged into a laptop (Intel Core i6, 16 GB RAM), and ran the AU detection algorithm on the laptop. As shown in Fig. 10.1, these cameras were attached to Nao's head using custom 3D printed glasses. AU

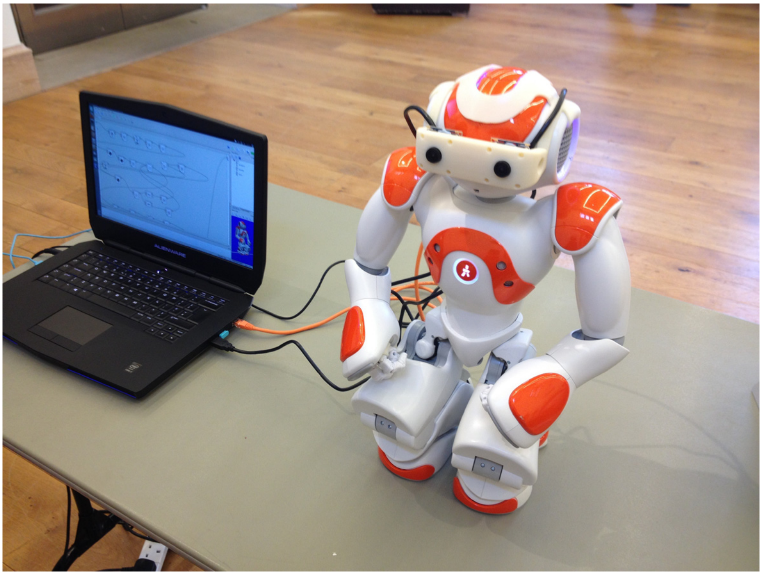

Figure 10.1 The robotic platform used during real-time public demonstrations.

detection from the robot's point of view is shown in Fig. 10.2. Vertical and horizontal bars indicate the head pose, and the color green is associated with frontal or nearly frontal head poses that yield more reliable AU detection. The detected AUs are highlighted in blue on the left-hand side of each frame; for example, AU1 and AU2 are detected in Fig. 10.2A.

We demonstrated the real-time AU detection method through face-to-face interactions with the Nao robot in a series of public engagement events. For this purpose, we designed an interactive game where Nao asked participants to help him improve his emotional intelligence by displaying facial expressions of emotion, such as happiness, sadness, etc. The participant could choose to display any AU such as pulling lip corners up (smile), pulling eyebrows up (surprise), dropping the mouth/chin (surprise), lowering the eyebrows (frown), etc. To collect the neutral face that was needed for the δ-appearance and δ-shape representations, we asked the participant at the beginning of the session to stand still and look at the camera. Since the neutral face was collected only for the frontal face, we did not take into account AUs detected in the non-frontal faces.

As illustrated in Fig. 10.2, Nao attempted to recognize each AU displayed by the participant, and inferred the expressed emotion based on the rule based approach, and then asked the participant for feedback in the

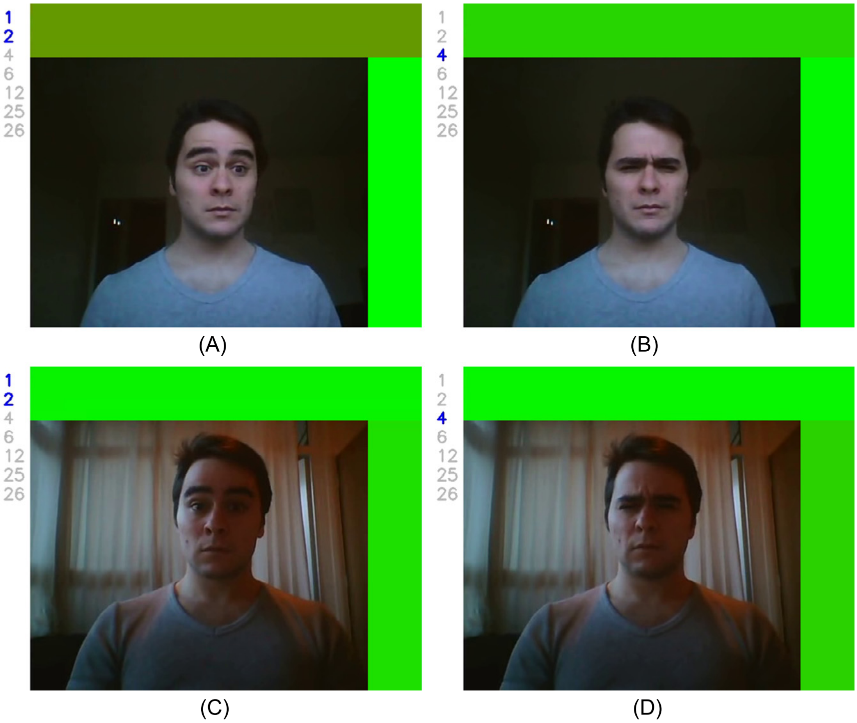

Figure 10.2 AU detection results under different illumination conditions, i.e. (A–B) vs (C–D). Vertical and horizontal bars indicate the head rotation; the color green is associated with frontal/nearly frontal head poses. The detected AUs in each face image are highlighted in blue: (A, C) AU1 and AU2; (B, D) AU4. (For interpretation of the colors in this figure, the reader is referred to the web version of this chapter.)

form of whether the recognized emotion was correct or not. However, an online learning algorithm was not considered, similarly to [31]. Sample images from the Cambridge Science Festival that took place in Cambridge, United Kingdom, on March 13, 2017,[1] are given in Fig. 10.3. The images illustrate the moment that one of the participants from the public displayed different facial expressions of emotions.

Here, we presented a real-life demonstration of the proposed affect analysis approach in a science communication scenario. However, this approach can be utilized in a health scenario, where, similarly to [32], the robot would provide assistance to children with Autism Spectrum Disorders for improving their facial emotion expression/recognition capability.

[1] https://www.sciencefestival.cam.ac.uk/events/teach-me-emotional-intelligence.

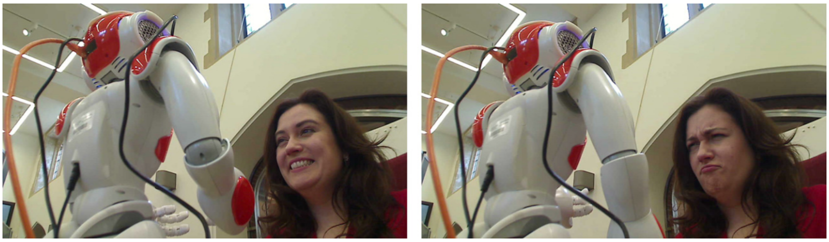

Figure 10.3 Photos from the public demonstration at the Cambridge Science Festival (Image copyright: University of Cambridge).

10.3.2 Automatic Personality Prediction

Several studies have shown that the success of social robots highly depends on assessing and responding to the user's personality (see Section 10.2.2). In this section, we describe how to build an automatic predictor of user personality during human–robot interactions as originally presented in [57]. We also investigate the impact of the participant's personality and the robot's personality on the human–robot interaction.

10.3.2.1 Personality Analysis Methodology

Data Collection and Annotation. To model the user's personality, we designed an experimental study involving interactions between two human participants and a robot, and collected audio-visual data using a set of first-person vision cameras (also called egocentric cameras) and annotation data by asking participants to complete BFI personality questionnaires [45].

We recruited participants from graduate students and researchers to take part in our experiment. The flow of interaction between the two participants and the robot was structured as follows. 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. The robot initiated the conversation by asking neutrally, "You, on my right, could you please stand up? Thank you! What is your name?" Then the robot continued by asking each of the participants about their occupations, feelings, and so on, by specifying their names at each turn.

We used the Nao robotic platform with the technical details of NaoQi version 2.1, head version 4.0, and body version 25. The robot was controlled remotely in a Wizard of Oz setup during the interaction. To manage the turn taking, an experimenter (i.e. operator), who was seated out of sight behind a sheet of poster board, operated the robot using a computer, the robot's camera, and the other cameras placed in the experimental room. To

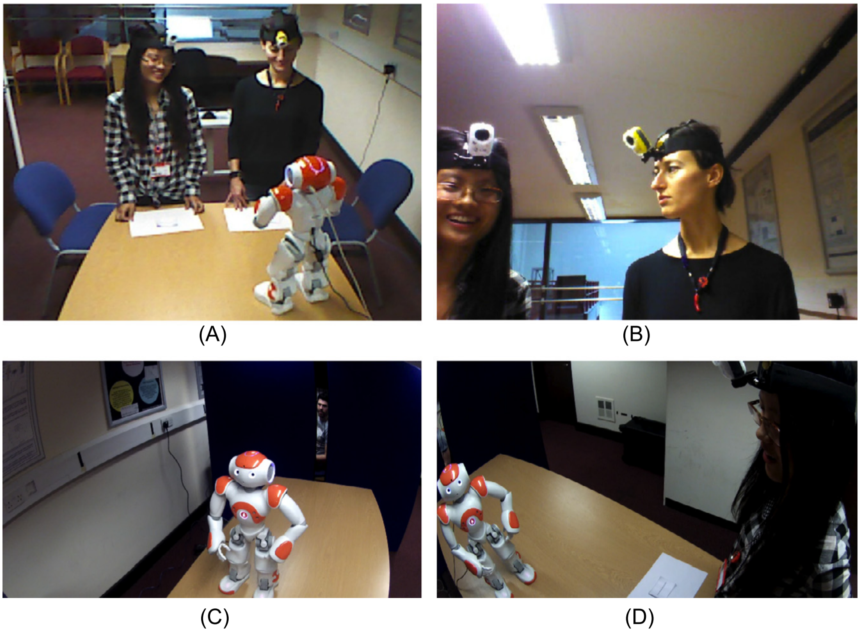

Figure 10.4 (A) The human–robot interaction setup. (B–D) Simultaneously captured snapshots from the first-person videos: the robot's camera (B) and the ego-centric cameras placed on the foreheads of the participants (C–D).

examine the importance of the robot's personality in the HRI, the robot was made to exhibit either extroverted or introverted personality. Following the previous literature [49,50], we manipulated the robot's behaviors to generate the two types of personality. 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.

A total of 18 participants (9 female and 9 male) took part in our experiment. Each interaction session lasted from 10 to 15 minutes and was recorded from different camera views. First-person videos were recorded using two Liquid Image ego-centric cameras[2] placed on the forehead of each participant and the robot's camera. The whole scene was also captured using two static Microsoft Kinect depth sensors (version 1)[3] as shown in Fig. 10.4A, resulting in RGB-D recordings. Sound was recorded via the microphones built into the ego-centric cameras.

[2] www.liquidimageco.com/products/model-727-the-ego-1.

[3] en.wikipedia.org/wiki/Kinect.

We recorded 12 interaction sessions and collected approximately 6 hours of multimodal recordings. Each session involved two participants, resulting in 24 individual recordings (some participants took part more than once provided that they had a different partner and were exposed to different robot personalities). The ego-centric recordings and the robot's camera were unsynchronized with the Kinect cameras. For this reason, the experimenter switched the light off and on before each session started. This co-occurred appearance change in the cameras was used to synchronize the multiple videos (i.e. from the two ego-centric cameras, the two Kinect depth sensors, and the 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 duration. Each clip comprises the robot asking a question to one of the participants and the target participant responding accordingly. This yielded 456 clips where each clip has a duration ranging from 20 to 120 seconds.

In this chapter, we only took into account the recordings from the ego-centric cameras. First-person vision has been shown to be advantageous in analyzing social interactions [78] as it provides the most relevant part of the data. For instance, the people who the camera wearer interacts with tend to be centered 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. Fig. 10.4 illustrates simultaneous snapshots from the ego-centric clips.

The participants were asked to complete two different questionnaires, one before the interaction session (pre-study questionnaire) and the other after the interaction session (post-study questionnaire). All measures were on a 10-point Likert scale (from very low to very high). For the pre-study questionnaire, we used the BFI-10 [79] to measure the Big Five personality traits, which is the short version of the Big Five Inventory, and has been used in other studies, e.g. [44]. Each item contributes to the score of a particular trait. The post-study questionnaire consisted of five items (see Table 10.2) that evaluate the participants' engagement with the robot and measure their impressions about the robot's behaviors and abilities.

Feature Extraction. We used simple and computationally efficient low-level features to describe motion and changes from the first-person perspective [80]. As mentioned in Section 10.2.2, nonverbal cues conveyed through gaze direction, attention, and head movement carry important information regarding the individual's personality and internal states. These

Table 10.2 Post-study questionnaire to evaluate the interaction experience with the robot

Question	Interaction measure
I enjoyed the interaction with the robot.	Engagement
I thought the robot was being supportive.	Empathy
I thought the robot was assertive and social.	Extroversion
I thought the robot was being positive.	Positivity
I found the robot's behavior realistic.	Realism

behaviors might lead to significant motion in the first-person videos, which can be characterized by optical flow and motion blur. Attention shifts and rapid scene changes may also cause drastic illumination changes.

Blur values were computed based on the no-reference blur estimation algorithm of [81]. Given a frame, this algorithm yielded two values, vertical (BLUR-Ver) and horizontal blur (BLUR-Hor), ranging from 0 to 1 (the best and 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.

For optical flow, we used the SIFT flow algorithm proposed in [82]. We computed a dense optical flow estimate for each frame, where we set the grid size to 4. We converted the x and y flow estimate of a pixel into magnitude and angle, and then quantized the angles into eight orientation bins. We calculated the mean (MAG-Mean) and the median (MAG-Med) of the magnitude values per frame. For the angle values, two types of features were computed over a frame: (i) the number of times the angle bin i contained the most motion energy in a frame (ANG-Nrg-i) and (ii) the total number of pixels belonging to the angle bin i (ANG-Count-i). These features were normalized such that the sum over all eight bins was 1.

Since the frame rate of the ego-centric cameras was high (60 frames per second), all features were extracted from frames sampled every 200 milliseconds instead of at adjacent time instants. A clip was summarized by computing a total of 40 features over the frames. Each feature was computed by performing a series of operations over the blur, illumination, and optical flow features. These operations calculated the mean (Mean), median (Med), and standard deviation (Std) over all frames in a video, calculating the absolute mean (Abs-Mean) over all frames, applying z-score normalization (z) across all frames and taking the first (d1) and the second (d2) temporal derivatives.

Table 10.3 Significant correlations between the Big Five personality traits of the participants and their interaction experience measures (at a significance level of $p < 0.05$, $^{*}p < 0.01$). EXT: extroversion, AGR: agreeableness, CON: conscientiousness, NEU: neuroticism, OPE: openness

Trait	Extroverted robot condition	Introverted robot condition
EXT	Engagement (0.85*) Empathy (0.58)	–
AGR	Engagement (0.62)	–
CON	Positivity (0.71*)	Positivity (0.71)
NEU	Realism (0.60)	–
OPE	–	Positivity (0.70) Realism (0.67)

10.3.2.2 Experimental Results

This section presents the correlation analysis between the Big Five personality traits and the interaction experience, and also examines how personality is linked to the automatically extracted first-person vision features. We tested the statistical significance of the correlations (against the null hypothesis of no correlation) using a t-distribution test.

Relationship Between Personality and Interaction Experience. We investigated the possible links between the Big Five personality traits of the participants, the *extroversion/introversion* trait of the robot, and the participants' interaction experience with the robot. In Table 10.3, the significant results are given with their respective correlation values in parentheses.

For the extroverted robot condition, the perceived *engagement* with the robot is found to be significantly correlated with participants' *extroversion* trait, which validates the similarity rule [60,42]. We observe that the robot's perceived *empathy* positively correlates with the participants' *extroversion* trait. This might be due to the fact that extroverted people feel more control over their interactions and judge them as more intimate and less incompatible [83,84]. A study of *agreeableness* reported that more agreeable people showed strong self-reported *rapport* when they interacted with a virtual agent [85]. Cuperman and Ickes [41] also indicated that more agreeable people reported having more enjoyable interactions. Similarly, we observe that perceived *engagement* with the robot is highly correlated with the *agreeableness* trait of the participants. A significant relationship is also established between the robot's perceived *realism* and the *neuroticism* trait of the participants. People who score high on *neuroticism* tend to perceive their

Table 10.4 Selected statistically significant correlations between the participants' personality traits and first-person vision features (at a significance level of $p < 0.01$). BLUR: blur, ILLU: illumination, MAG: optical flow magnitude, ANG: optical flow angle, EXT: extroversion, AGR: agreeableness, CON: conscientiousness, NEU: neuroticism, OPE: openness

Trait	Extroverted robot condition	Introverted robot condition

EXT	–	BLUR-Ver-Mean(−0.55);
		BLUR-Ratio-Med(−0.49)

AGR	BLUR-Ver-Mean(0.36)	BLUR-Max-Med(0.35)

CON	BLUR-Ver-Mean(0.34);
	ILLU-Mean-Std(−0.33)
		BLUR-Ver-Med(−0.53);
		BLUR-Ratio-Med(−0.48);
		ANG-Nrg-1(0.35)

NEU	BLUR-Ver-Mean(−0.40);
	BLUR-Ratio-Med(−0.36);
	ILLU-Med-Std(−0.38);
	ANG-Nrg-1(0.41);
	ANG-Count-2(−0.42)
		BLUR-Ver-Mean(0.68);
		BLUR-Max-Std(0.40);
		BLUR-Ratio-Med(0.61);
		MAG-Mean-Mean(0.35);
		MAG-Mean-d1-Abs-Mean(0.38)

OPE	BLUR-Max-Mean(0.34);
	ILLU-Med-Std(0.39);
	ANG-Count-3(0.33)
		BLUR-Hor-Mean(0.47);
		BLUR-Ver-Mean(−0.43);
		MAG-Mean-Mean(-0.35);
		ANG-Count-1(0.35)

interactions as being forced and strained [41] and therefore the artificial behaviors of the robot might appear to them as realistic.

For the introverted robot condition, no significant correlations are obtained with participants' *extroversion*, *agreeableness*, and *neuroticism* traits. People who score high on *conscientiousness* tend to interact with others by showing greater attentiveness and responsiveness [41]. This might cause significant correlations with the interaction measure of *positivity* regardless of the robot's personality as the robot always provided feedback to the participant in the course of interaction.

Relationship Between Personality and First-Person Vision Features. The goal of this analysis was to study the one-to-one relationships between the Big Five personality traits of the participants and the automatically extracted first-person features. Table 10.4 shows the prominent features and the significant correlations.

In general the introverted robot condition provides a larger number of significant correlations with the extracted features. This can be due to the participants' attention being shifted more when interacting with the introverted robot. For the extroverted robot condition, the *neuroticism* trait of the participants shows significant relationships with all three feature types (blur, illumination, and optical flow), in particular with blur features. No significant correlations are found between participants' *extroversion* trait and the first-person features. For the introverted robot condition, the personality traits of *conscientiousness*, *neuroticism*, and *openness* of the participants show significant relationships with the blur and optical flow features. However, no correlations are found with the illumination features.

In Table 10.4, one significant relationship is seen between *agreeableness* and the vertical blur feature, which can be associated with head nodding and being positive and supportive. We observe that extroverted people tend to enjoy the interaction with the extroverted robot more than the interaction with the introverted robot. Our experimental results further show that *extroversion* is negatively correlated with the blur (motion) features for the introverted robot. This result indicates that less energetic (introverted) people like the introverted robot more, and it is possible to deduce this from the first-person vision features extracted.

For automatic personality prediction, we employed the linear Support Vector Regression method with nested leave-one-subject-out cross-validation. Optical flow-angle features (ANG-Nrg and ANG-Count) yielded the best prediction results in terms of coefficient of determination (R^2) and root-mean-square error ($RMSE$), where we obtained $\mu_{R^2} = 0.19$ and $\mu_{RMSE} = 1.63$ over all traits. The method successfully modeled the relationship between the first-person vision features and the traits of *agreeableness* ($R^2 = 0.48$, $RMSE = 1.37$), *conscientiousness* ($R^2 = 0.27$, $RMSE = 1.55$), and *extroversion* ($R^2 = 0.20$, $RMSE = 1.72$). Similarly, the study in [53] applied Ridge regression to predict the *extroversion* trait. Although the database, Likert scale, and visual feature set used were completely different, they also obtained the best results with motion-based features ($R^2 = 0.31$). Referring to this result as a baseline, our results for *agreeableness*, *conscientiousness*, and *extroversion* show that prediction of personality traits from first-person vision in the scope of HRI is a promising research direction.

10.4 CONCLUSION AND DISCUSSION

Robotics as a field is continuously evolving to address the ever-changing needs of humans in society. Today the potential of affective and social robotics is enormous, including but not limited to promoting the health and well-being of the elderly living at home [86], improving the quality of life of individuals via physical recovery and rehabilitation [87], assisting the caregivers of children with cognitive and social disabilities [5], assisting children with special medical needs such as diabetes [88], providing personalized education for children [89], and facilitating engagement in group interactions for improving team performance [6]. To deploy social robots in such naturalistic human–robot interaction settings, user modeling and personalization through automatic analysis of expressions, emotions, personality, and engagement is key.

In the light of the survey of the recent research trends and techniques used by social robots, we would like to conclude this chapter by highlighting three open problems in the field, together with a number of pathways that can be used to address these problems.

Cross-Fertilization Between Affective Computing and Social Robotics Fields. In recent years significant progress has been achieved in automatic analysis of affective and social signals, particularly of emotions and affective states; even so, computational social robotics has not yet incorporated these latest developments. There is an apparent lack of cross-fertilization between these fields and the field of social robotics. In the fields of affective computing and social signal processing, the current computational techniques integrate multimodal features from visual, audio, and physiological cues over time and utilize models trained with deep learning. However, to date, there has been virtually no effort to integrate these latest trends into social robots and test their viability in the context of human–robot interaction. This is mainly due to the need of real-time processing and to the lack of computational power available on the current robotic platforms. One possible solution to this issue is attaching external cameras onto the robots and performing the real-time processing on an external computer, as described in Section 10.3.1.2. However, this solution does not hold for mobile robots. Another promising direction is cloud robotics, where the captured data is directly streamed to a server via the network for effective and efficient computing (e.g. [90]). This brings additional challenges into play, including the analysis of affect and social signals using live-streamed data that has low spatial and temporal resolution.

Analysis Under Realistic and Adverse Conditions. For emotion recognition, most of the successful methods in computer science have focused on facial cues, and have been characterized by multimodal features, in particular combining facial cues with audio cues and bio-signals such as Electrodermal Activity (EDA). Bio-signals are useful when facial cues cannot be observed reliably. However, in real-life applications, it might not be always possible to attach sensors onto the participants to measure their physiological responses. Reducing the cost, the size, and the invasiveness of the physiological sensors that can work robustly under adverse conditions is expected to resolve many of these challenges. Body postures and hand gestures are important sources of information for the analysis of affective and social signals. Therefore, a promising direction is to use deep learning approaches that combine multiple visual cues, such as facial and bodily cues. However, fusing multiple cues in an effective and efficient manner still remains an open challenge in the field. Learning what to fuse and when as suggested in [26,91] will also help deal with missing data, i.e. the cases where one of the cues is not available or is not reliably detected.

Datasets and Ground Truth. Most of the available datasets in social robotics have relied on self-reported assessments, in particular, for assessing personality. However, in real-life applications, self-reported assessments might not be available for evaluating the performance of the automatic analyzers. Online crowd-sourcing platforms (e.g. MTurk) have recently gained popularity, due to their efficiency and practicality for collecting responses from crowds for large sets of data within a short period of time. Such efforts have clearly been proven to be efficient at predicting personality [92]. However, exploring novel ways to incorporate annotation disagreements into the analyzers, similarly to [93], is an avenue that needs to be explored further.

In summary, the review provided in this chapter illustrates that the capabilities of current social robots are quite limited. There is a clear need for incorporating the automatic affect analysis and social processing methods into real-life human–robot interaction applications and for improving these techniques to address the challenges of varying environmental lighting, user distance to camera, camera view, and real-time computational requirements. The availability of commercial robotic platforms such as iCat [35], iCub [59], and Nao [40], and developments in collaborative academic research such as the Frontiers Research Topic on *Affective and Social Signals for HRI*[4] provide us with a positive outlook. However, to truly address the existing challenges, researchers from the relevant fields, including but not limited to psychology, nonverbal behavior, vision, social signal processing, affective computing, and HRI, need to constantly interact with one another.

[4] http://journal.frontiersin.org/researchtopic/5162/affective-and-social-signals-for-hri.

ACKNOWLEDGMENTS

This work was funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (Grant Ref.: EP/L00416X/1).

REFERENCES

[1] Robot receptionists introduced at hospitals in Belgium, https://www.theguardian.com/technology/2016/jun/14/robot-receptionists-hospitals-belgium-pepper-humanoid (Accessed May 2017).

[2] Japanese bank introduces robot workers to deal with customers in branches, https://www.theguardian.com/world/2015/feb/04/japanese-bank-introduces-robot-workers-to-deal-with-customers-in-branches (Accessed May 2017).

[3] SoftBank's Robot 'Pepper' Flogs Beer and Burgers at Airport, https://www.bloomberg.com/news/articles/2017-02-10/softbank-s-robot-pepper-flogs-beer-and-burgers-at-airport-iyz2t9hb (Accessed May 2017).

[4] K. Wada, T. Shibata, T. Saito, K. Tanie, Psychological and social effects of robot assisted activity to elderly people who stay at a health service facility for the aged, in: IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422), vol. 3, 2003, pp. 3996–4001.

[5] M. Blow, K. Dautenhahn, A. Appleby, C.L. Nehaniv, D. Lee, Perception of robot smiles and dimensions for human–robot interaction design, in: The 15th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2006, Hatfield, Herthfordshire, UK, September 6–8, 2006, 2006, pp. 469–474.

[6] E. Short, K. Sittig-Boyd, M.J. Mataric, Modeling moderation for multi-party socially assistive robotics, in: IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016, IEEE, New York, NY, USA, 2016.

[7] S. Atmatzidou, S. Demetriadis, Advancing students' computational thinking skills through educational robotics: a study on age and gender relevant differences, Robotics and Autonomous Systems 75 (Part B) (2016) 661–670, https://doi.org/10.1016/j.robot.2015.10.008, http://www.sciencedirect.com/science/article/pii/S0921889015002420.

[8] S. Spaulding, G. Gordon, C. Breazeal, Affect-aware student models for robot tutors, in: Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems, AAMAS '16, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, USA, ISBN 978-1-4503-4239-1, 2016, pp. 864–872, http://dl.acm.org/citation.cfm?id=2937029.2937050.

[9] D. Leyzberg, S. Spaulding, B. Scassellati, Personalizing robot tutors to individuals' learning differences, in: Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction, HRI '14, ACM, New York, NY, USA, ISBN 978-1-4503-2658-2, 2014, pp. 423–430, http://doi.acm.org/10.1145/2559636.2559671.

[10] A. Ramachandran, C.M. Huang, B. Scassellati, Give me a break!: personalized timing strategies to promote learning in robot–child tutoring, in: Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction, HRI '17, ACM, New York, NY, USA, ISBN 978-1-4503-4336-7, 2017, pp. 146–155, http://doi.acm.org/10.1145/2909824.3020209.

[11] T. Schodde, K. Bergmann, S. Kopp, Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making, in: Proceedings of the 2017 ACM/IEEE International Conference on Human–Robot Interaction, HRI ’17, ACM, New York, NY, USA, ISBN 978-1-4503-4336-7, 2017, pp. 128–136, http://doi.acm.org/10.1145/2909824.3020222.

[12] H. Yan, M.H. Ang, A.N. Poo, A survey on perception methods for human–robot interaction in social robots, International Journal of Social Robotics 6 (1) (2014) 85–119.

[13] J.K. Burgoon, L.K. Guerrero, K. Floyd, Nonverbal Communication, Allyn and Bacon, Boston, MA, USA, 2009.

[14] M. Pantic, Automatic analysis of facial expressions, in: Encyclopedia of Biometrics, 2015, pp. 128–134.

[15] E. Sariyanidi, H. Gunes, A. Cavallaro, Automatic analysis of facial affect: a survey of registration, representation, and recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (6) (2015) 1113–1133.

[16] H. Gunes, C. Shan, S. Chen, Y. Tian, Bodily Expression for Automatic Affect Recognition, John Wiley & Sons, Inc., 2015, pp. 343–377.

[17] P. Ekman, W. Friesen, J. Hager, The Facial Action Coding System, 2nd ed., Weidenfeld and Nicolson, London, UK, 2002.

[18] P. Ekman, J. Campos, R. Davidson, F.D. Waals, Emotions inside out, Annals of the New York Academy of Sciences (2003) 1000.

[19] H. Gunes, B. Schuller, Categorical and dimensional affect analysis in continuous input: current trends and future directions, Image and Vision Computing 31 (2) (2013) 120–136, https://doi.org/10.1016/j.imavis.2012.06.016, http://www.sciencedirect.com/science/article/pii/S0262885612001084.

[20] M.F. Valstar, M. Mehu, B. Jiang, M. Pantic, K. Scherer, Meta-analysis of the first facial expression recognition challenge, IEEE Transactions on Systems, Man and Cybernetics. Part B. Cybernetics 42 (4) (2012) 966–979, https://doi.org/10.1109/TSMCB.2012.2200675.

[21] M.F. Valstar, T. Almaev, J.M. Girard, G. McKeown, M. Mehu, L. Yin, et al., FERA 2015 – second facial expression recognition and analysis challenge, in: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, 2015, pp. 1–8.

[22] B. Schuller, M. Valstar, R. Cowie, M. Pantic, Avec 2012: the continuous audio/visual emotion challenge – an introduction, in: Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI ’12, ACM, New York, NY, USA, ISBN 978-1-4503-1467-1, 2012, pp. 361–362, http://doi.acm.org/10.1145/2388676.2388758.

[23] F. De la Torre, W.S. Chu, X. Xiong, F. Vicente, X. Ding, J.F. Cohn, Intraface, in: IEEE International Conference on Automatic Face and Gesture Recognition (FG), 2015.

[24] S. Jaiswal, M. Valstar, Deep learning the dynamic appearance and shape of facial action units, in: IEEE Winter Conference on Application of Computer Vision, 2016.

[25] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, H. Sahli, Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks, in: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC ’15, ACM, New York, NY, USA, ISBN 978-1-4503-3743-4, 2015, pp. 73–80, http://doi.acm.org/10.1145/2808196.2811641.

[26] S. Chen, Q. Jin, Multi-modal conditional attention fusion for dimensional emotion prediction, in: Proceedings of the 2016 ACM on Multimedia Conference, MM

'16, ACM, New York, NY, USA, ISBN 978-1-4503-3603-1, 2016, pp. 571–575, http://doi.acm.org/10.1145/2964284.2967286.

[27] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, et al., AV+EC 2015: the first affect recognition challenge bridging across audio, video, and physiological data, in: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, AVEC '15, ACM, New York, NY, USA, ISBN 978-1-4503-3743-4, 2015, pp. 3–8, http://doi.acm.org/10.1145/2808196.2811642, 2015.

[28] D. McColl, A. Hong, N. Hatakeyama, G. Nejat, B. Benhabib, A survey of autonomous human affect detection methods for social robots engaged in natural HRI, Journal of Intelligent & Robotic Systems 82 (1) (2016) 101–133.

[29] F. Cid, J.A. Prado, P. Bustos, P. Nunez, A real time and robust facial expression recognition and imitation approach for affective human–robot interaction using Gabor filtering, in: IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 2188–2193.

[30] F. Cid, J. Moreno, P. Bustos, P. Núnez, Muecas: a multi-sensor robotic head for affective human robot interaction and imitation, Sensors 14 (5) (2014) 7711–7737.

[31] S. Boucenna, P. Gaussier, P. Andry, L. Hafemeister, A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game, International Journal of Social Robotics 6 (4) (2014) 633–652.

[32] M. Leo, M.D. Coco, P. Carcagnì, C. Distante, M. Bernava, G. Pioggia, G. Palestra, Automatic emotion recognition in robot–children interaction for ASD treatment, in: IEEE International Conference on Computer Vision Workshop (ICCVW), 2015, pp. 537–545.

[33] Robokind Robots Advanced Social Robotics, http://robokind.com/ (Accessed May 2017).

[34] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, P. McOwan, Multimodal affect modelling and recognition for empathic robot companions, International Journal of Humanoid Robotics 10 (1) (2013).

[35] A. van Breemen, X. Yan, B. Meerbeek, iCat: an animated user-interface robot with personality, in: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems, ACM, 2005, pp. 143–144.

[36] D. Schacter, C. Wang, G. Nejat, B. Benhabib, A two-dimensional facial-affect estimation system for human–robot interaction using facial expression parameters, Advanced Robotics 27 (4) (2013) 259–273.

[37] S. Lucey, Y. Wang, M. Cox, S. Sridharan, J.F. Cohn, Efficient constrained local model fitting for non-rigid face alignment, Image and Vision Computing 27 (12) (2009) 1804–1813.

[38] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, et al., Real-time human pose recognition in parts from single depth images, Communications of the ACM 56 (1) (2013) 116–124, https://doi.org/10.1145/2398356.2398381, http://doi.acm.org/10.1145/2398356.2398381.

[39] W. Wang, G. Athanasopoulos, G. Patsis, V. Enescu, H. Sahli, Real-Time Emotion Recognition from Natural Bodily Expressions in Child–Robot Interaction, Springer International Publishing, Cham, ISBN 978-3-319-16199-0, 2015, pp. 424–435.

[40] Aldebaran Softbank Group, Who is Nao?, https://www.ald.softbankrobotics.com/en/cool-robots/nao (Accessed 22 May 2017).

[41] R. Cuperman, 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 97 (4) (2009) 667–684.

[42] A. Aly, A. Tapus, A model for synthesizing a combined verbal and nonverbal behavior based on personality traits in human–robot interaction, in: Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction, 2013.

[43] P.J. Corr, G. Matthews, The Cambridge Handbook of Personality Psychology, Cambridge University Press, 2009.

[44] A. Vinciarelli, G. Mohammadi, A survey of personality computing, IEEE Transactions on Affective Computing 5 (3) (2014) 273–291.

[45] O.P. John, S. Srivastava, Big five inventory (BFI), in: Handbook of Personality: Theory and Research, vol. 2, 1999, pp. 102–138.

[46] L. Teijeiro-Mosquera, J.I. Biel, J.L. Alba-Castro, D. Gatica-Perez, What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube, IEEE Transactions on Affective Computing 6 (2) (2015) 193–205.

[47] 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 (2016) 705–721, https://doi.org/10.1109/ACCESS.2016.2614525.

[48] R.J. Larsen, T.K. Shackelford, Gaze avoidance: personality and social judgments of people who avoid direct face-to-face contact, Personality and Individual Differences 21 (6) (1996) 907–917.

[49] R.E. Riggio, H. Friedman, Impression formation: the role of expressive behavior, Journal of Personality and Social Psychology 50 (2) (1986) 421–427.

[50] R. Lippa, The nonverbal display and judgment of extraversion, masculinity, femininity, and gender diagnosticity: a lens model analysis, Journal of Research in Personality 32 (1) (1998) 80–107.

[51] J. Joshi, H. Gunes, R. Goecke, Automatic prediction of perceived traits using visual cues under varied situational context, in: IEEE International Conference on Pattern Recognition (ICPR), 2014, pp. 2855–2860.

[52] M. Schroder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, et al., Building autonomous sensitive artificial listeners, IEEE Transactions on Affective Computing 3 (2) (2012) 165–183, https://doi.org/10.1109/T-AFFC.2011.34.

[53] O. Aran, D. Gatica-Perez, One of a kind: inferring personality impressions in meetings, in: Proceedings of the ACM International Conference on Multimodal Interaction, 2013.

[54] D. Sanchez-Cortes, O. Aran, M.M. Schmid, D. Gatica-Perez, A nonverbal behavior approach to identify emergent leaders in small groups, IEEE Transactions on Multimedia 14 (3) (2012) 816–832.

[55] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, M. Zancanaro, Multimodal recognition of personality traits in social interactions, in: Proceedings of the 10th International Conference on Multimodal Interfaces, ICMI ’08, 2008, pp. 53–60.

[56] O. Celiktutan, H. Gunes, Automatic prediction of impressions in time and across varying context: personality, attractiveness and likeability, IEEE Transactions on Affective Computing 8 (1) (2017) 29–42.

[57] O. Celiktutan, H. Gunes, Computational analysis of human–robot interactions through first-person vision: personality and interaction experience, in: 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2015, pp. 815–820.

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

[59] L. Natale, F. Nori, G. Metta, M. Fumagalli, S. Ivaldi, U. Pattacini, et al., The iCub platform: a tool for studying intrinsically motivated learning, in: Intrinsically Motivated Learning in Natural and Artificial Systems, Springer, 2013, pp. 433–458.

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

[61] O. Celiktutan, E. Sariyanidi, H. Gunes, Let me tell you about your personality!: real-time personality prediction from nonverbal behavioural cues, in: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 1, 2015, IEEE, 2015, p. 1.

[62] C.L. Sidner, M. Dzikovska, Human–robot interaction: engagement between humans and robots for hosting activities, in: Fourth IEEE International Conference on Multimodal Interfaces, 2002, pp. 123–137.

[63] A. Kapoor, R.W. Picard, Y. Ivanov, Probabilistic combination of multiple modalities to detect interest, in: Proceedings of the 17th International Conference on Pattern Recognition, vol. 3, ICPR 2004, 2004, pp. 969–972.

[64] C. Oertel, G. Salvi, A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, ICMI '13, ACM, New York, NY, USA, ISBN 978-1-4503-2129-7, 2013, pp. 99–106, http://doi.acm.org/10.1145/2522848.2522865.

[65] C. Peters, S. Asteriadis, K. Karpouzis, Investigating shared attention with a virtual agent using a gaze-based interface, Journal on Multimodal User Interfaces 3 (1) (2010) 119–130.

[66] A. Cerekovic, O. Aran, D. Gatica-Perez, Rapport with virtual agents: what do human social cues and personality explain?, IEEE Transactions on Affective Computing 8 (3) (2017) 382–395, https://doi.org/10.1109/TAFFC.2016.2545650.

[67] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P.W. McOwan, A. Paiva, Automatic analysis of affective postures and body motion to detect engagement with a game companion, in: Proceedings of the 6th International Conference on Human–Robot Interaction, HRI '11, ACM, New York, NY, USA, ISBN 978-1-4503-0561-7, 2011, pp. 305–312, http://doi.acm.org/10.1145/1957656.1957781.

[68] W. Benkaouar, D. Vaufreydaz, Multi-sensors engagement detection with a robot companion in a home environment, in: Workshop on Assistance and Service Robotics in a Human Environment at IEEE International Conference on Intelligent Robots and Systems (IROS2012), 2012, pp. 45–52.

[69] Kompai Robots, http://kompai.com/ (Accessed May 2017).

[70] H. Salam, M. Chetouani, Engagement detection based on mutli-party cues for human robot interaction, in: International Conference on Affective Computing and Intelligent Interaction (ACII), 2015, IEEE, 2015, pp. 341–347.

[71] X. Xiong, F. De la Torre, Supervised descent method and its applications to face alignment, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 532–539.

[72] E. Sariyanidi, H. Gunes, M. Gökmen, A. Cavallaro, Local Zernike moment representations for facial affect recognition, in: Proceedings of the British Machine Vision Conference, 2013.

[73] E. Sariyanidi, H. Gunes, A. Cavallaro, Biologically-inspired motion encoding for robust global motion estimation, IEEE Transactions on Image Processing 26 (3) (2017) 1521–1535, https://doi.org/10.1109/TIP.2017.2651394.

[74] M. Pantic, M. Valstar, R. Rademaker, L. Maat, Web-based database for facial expression analysis, in: Proceedings of the IEEE International Conference on Multimedia and Expo, 2005, p. 5.

[75] C.C. Chang, C.J. Lin, LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1–27.

[76] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, M. Bartlett, The computer expression recognition toolbox (CERT), in: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition Workshops, 2011, pp. 298–305.

[77] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2001, p. I.

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

[79] B. Rammstedt, 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 41 (1) (2007) 203–212, https://doi.org/10.1016/j.jrp.2006.02.001.

[80] C. Tan, H. Goh, V. Chandrasekhar, L. Liyuan, J. Lim, Understanding the nature of first-person videos: characterization and classification using low-level features, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops, 2014.

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

[82] C. Liu, J. Yuen, A. Torralba, SIFT flow: dense correspondence across scenes and its applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 33 (5) (2011) 978–994.

[83] A.M. von der Putten, N.C. Kramer, J. Gratch, How our personality shapes our interactions with virtual characters – implications for research and development, in: Proceedings of the International Conference on Intelligent Virtual Agents, 2010.

[84] A.W. Heaton, A.W. Kruglanski, Person perception by introverts and extraverts under time pressure: effects of need for closure, Personality & Social Psychology Bulletin 17 (2) (1991) 161–165.

[85] S.H. Kang, J. Gratch, N. Wang, J.H. Watt, Agreeable people like agreeable virtual humans, in: Lecture Notes in Computer Science, 2008, pp. 253–261.

[86] J. Saunders, D.S. Syrdal, K.L. Koay, N. Burke, K. Dautenhahn, Teach me, show me: end-user personalization of a smart home and companion robot, IEEE Transactions on Human-Machine Systems 46 (1) (2016) 27–40, https://doi.org/10.1109/THMS.2015.2445105.

[87] W.G. Louie, S. Mohamed, G. Nejat, Human–Robot Interaction for Rehabilitation Robots: Principles and Practice, Taylor & Francis Group, Boca Raton, FL, USA, 2017, pp. 25–70.

[88] L. Cañamero, M. Lewis, Making new "New AI" friends: designing a social robot for diabetic children from an embodied AI perspective, International Journal of Social Robotics 8 (4) (2016) 523–537, https://doi.org/10.1007/s12369-016-0364-9.

[89] D. Leyzberg, S. Spaulding, B. Scassellati, Personalizing robot tutors to individuals' learning differences, in: Proceedings of the 2014 ACM/IEEE International Conference on Human–Robot Interaction, HRI '14, ACM, New York, NY, USA, ISBN 978-1-4503-2658-2, 2014, pp. 423–430, http://doi.acm.org/10.1145/2559636.2559671.

[90] Y. Yamauchi, Y. Kato, T. Yamashita, H. Fujiyoshil, Cloud robotics based on facial attribute image analysis for human–robot interaction, in: IEEE International Symposium on Robot and Human Interactive Communication, 2016.

[91] F. Li, N. Neverova, C. Wolf, G.W. Taylor, Modout: learning to fuse face and gesture modalities with stochastic regularization, in: International Conference on Automatic Face and Gesture Recognition, 2017.

[92] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, et al., ChaLearn LAP 2016: First Round Challenge on First Impressions – Dataset and Results, Springer International Publishing, Cham, ISBN 978-3-319-49409-8, 2016, pp. 400–418.

[93] V. Sharmanska, D. Hernández-Lobato, J.M. Hernandez-Lobato, N. Quadrianto, Ambiguity helps: classification with disagreements in crowdsourced annotations, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2194–2202.