Automatic Prediction of Impressions in Time and across Varying Context: Personality, Attractiveness and Likeability

Oya Çeliktutan and Hatice Gunes

Abstract—In this paper, we propose a novel multimodal framework for automatically predicting the impressions of *extroversion*, *agreeableness*, *conscientiousness*, *neuroticism*, *openness*, *attractiveness* and *likeability* continuously in time and across varying situational contexts. Differently from the existing works, we obtain visual-only and audio-only annotations continuously in time for the same set of subjects, for the first time in the literature, and compare them to their audio-visual annotations. We propose a time-continuous prediction approach that learns the temporal relationships rather than treating each time instant separately. Our experiments show that the best prediction results are obtained when regression models are learned from audio-visual annotations and visual cues, and from audio-visual annotations and visual cues combined with audio cues at the decision level. Continuously generated annotations have the potential to provide insight into better understanding which impressions can be formed and predicted more dynamically, varying with situational context, and which ones appear to be more static and stable over time.

Index Terms—Interpersonal perception, personality, attractiveness, likeability, time-continuous prediction

1 Introduction

This paper focuses on automatic prediction of impressions, namely, inferences about traits and characteristics of people based on their observable behaviours. Impressions are an integral part of our lives - we constantly make everyday decisions and long-term plans ranging from *whom we will sit next to on a bus journey* to *whom we are going to be friends with*, based on our judgements arising from social interactions.

Interpersonal perception have been widely investigated along various aspects over the past decades. Kenny [1] conceptualised the process of forming an impression of another as integration of separate information sources (e.g., physical appearance, behaviour) and personal interpretations into an unitary judgement, and proposed a mathematical model called PERSON. There has been a general agreement that while in the initial phase of impression formation the physical appearance (e.g., stereotype) is the primary source of information, the target's behaviours (e.g., personality) become more salient with information gathered over time [1], [2]. Researchers have examined the differences in how people form impressions in person versus by just watching someone as a passive observer, and reported that the passive means of making impressions were as accurate as meeting someone in person [2].

The target person is viewed differently when evaluated in different contexts, i.e., if the perceiver observes the target in a new context, there might be a change in the perceiver's impression [1]. It was also found that the correlation between impressions and self-assessments increases with the number and variety of targets' behavioural contexts observed by the perceiver [3]. Research in [1], [3], [4], [5] suggested that even thin slices (short durations) lead to consensus among different observers, and even complete strangers can make valid personality judgements after watching a short video of a person. Carney et al. [5] investigated the minimum sufficient conditions under which people make a trait inference, which was reported to be as small as 5 s for *neuroticism* and *openness*.

While impression formation has been a hot area of research in psychology, recent years have brought interest in computational models for perception of personality [6], and perception of human beauty, attractiveness and likeability [7]. Understanding these perception mechanisms is useful in many applications such as recruiter and candidate matching, person and romantic partner matching, adapting marketing messages based on the users' profiles, and is also essential in improving user experience and engagement in human-computer interaction [8].

This paper focuses on human-virtual agent interactions from the SEMAINE corpus [9] and presents an automatic personality prediction approach by assessing and mathematically modelling how impressions fluctuate with time. We ask external observers to make personality judgements while simultaneously watching/listening to a clip of a participant. Participants interact with three distinct virtual agents, each enforcing a different situational context. Differently from the existing works, we obtain visual-only and audio-only annotations continuously in time and across these varying situational contexts, for the same set of subjects, for the first time in the computer science literature, and compare them to the audio-visual annotations. We

• *The authors are with the Computer Laboratory, University of Cambridge, Cambridge, United Kingdom.*
E-mail: oya.celiktutan@gmail.com, haticeg@ieee.org.

Manuscript received 1 Nov. 2014; revised 27 Aug. 2015; accepted 14 Dec. 2015. Date of publication 0 . 0000; date of current version 0 . 0000.
Recommended for acceptance by S. D'Mello.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2015.2513401

employ a time-series regression method in conjunction with multimodal features for automatically predicting the impressions of *agreeableness*, *openness*, *neuroticism*, *conscientiousness*, *extroversion*, *engagement*, *facial attractiveness*, *vocal attractiveness* and *likeability* continuously in time. The proposed time-continuous prediction approach yields superior prediction results when trained with audio-visual annotations as compared to when trained with visual-only/audio-only annotations, which indicates personality perception is modelled better in the presence of more information. Our results also show that situational context is important for personality prediction, i.e., overall, better results are obtained for cheerful and friendly agent context.

Although modelling the dynamics of expressions and affect has been extensively studied in the literature [10], [11], to the best of our knowledge, time-continuous prediction of impressions has not been addressed yet. The SEMAINE system [9] is a representative system that analyses the nonverbal behaviours and affective states of the users interacting with a virtual agent and allows the virtual agent to react accordingly for maintaining the flow of the conversation. However, our aim is to understand the personality of the user in the course of the interaction. The proposed time-continuous approach enables automatic personality prediction in real-time as demonstrated in [12], [13], which is able to publish/send messages to a synthesis module for system adaptation.

2 Psychology Background

Personality is crucial to understanding human behaviours. Therefore, there exists a significant body of psychology literature on personality research. The traditional approach to describe personality is the trait theory that focuses on the measurement of general patterns of behaviours, thoughts and emotions, which are relatively stable over time and across situational contexts [14]. 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* (having tendency to negative emotions such as anxiety, depression or anger), *openness* (having tendency to changing experience, adventure, new ideas), *agreeableness* (cooperative, compliant, trustworthy) and *conscientiousness* (self-disciplined, organized, reliable, consistent). Although the general agreement has been that people show behavioural stability, a number of studies [15], [16] have demonstrated that there exists a substantial intra-person variability over short periods of time. This dynamic perspective has motivated researchers to develop the concept of *personality states* [15], [16] that can be regarded as short-term manifestations of traits. States represent how a person deviates from her or his typical way of acting (i.e., stable traits) at a given moment.

Research focusing on the impression formation has predominantly focused on the Big Five personality traits and examined each trait separately. Kenny [17] proposed a mathematical model, the so-called Weighted Average Model (WAM), and examined the impact of different factors in the level of consensus among multiple observers. In [1], Kenny reparametrised WAM into PERSON model, which comprises six factors of Personality, Error, Residual, Stereotype, Opinion and Norm. For example, stereotype is associated with the shared assumptions based on physical appearance.

Kenny [1] indicated that the external observers' impressions become more reliable when each observes a series of acts from the same target. In other words, personality impressions can change from one single act to another, but the accuracy increases with the number of observed acts (context). Borkenau et al. [3] also found that observers' accuracy in judging targets' personality increased with the variety of behavioural contexts. They recorded and judged each target across 15 behavioural contexts ranging from introducing oneself to telling a joke, from talking about hobbies to singing a song. While all traits seem to be inferred well from various behavioural contexts, inference of *openness* relates to more ability-demanding behaviours such as pantomime task.

Many works reported that impressions can be formed very quickly based on very little information (a few seconds only). In [18], Borkenau and Liebler asked external observers to view a 90 s-length-video of a target reading a text. They compared the agreement among the observers in two conditions: audio-visual video and visual-only (muted) video. While, for the Big Five personality traits, no significant difference has been found, audio information, especially verbal content, has been found to be more prominent in judging the target's intelligence. They found that the correlation between the impressions regarding different personality traits were higher in the presence of less information (i.e., visual-only video).

Carney et al. [5] examined the accuracy of personality judgements in varying exposure times (5, 20, 45, 60 and 300 s-slices of video) and temporal location of the slice within the video. In particular, they recorded 5 minutes-length videos of dyadic interactions and extracted slices (ranging from 5 s to 60 s) from three different temporal locations, i.e., beginning, middle and end of the video. Each target, engaged in an unstructured conversation, was assessed in 13 conditions (4 exposure times $\times$ 3 slice locations $+$ 300 s). The experimental results showed that, for *extroversion* and *agreeableness*, the exposure time and accuracy were found to be positively correlated, however there was no statically significant correlation found for *neuroticism*, *openness* and *conscientiousness*. The accuracy was also observed to be lower when the slices were extracted from the beginning of the video.

Ambady et al. [4] examined the role of personality, gender and nonverbal skills in zero-acquaintance situations from the perspective of both the observer and the target. They confirmed that extroverted people provide the others with the necessary cues for accurate interpretation, while less sociable people tend to be more accurate judges. Lorenzo et al. [19] also reported that physically attractive people tend to be more accurately judged by others. Physically attractive individuals are expected to be more sociable, friendly and intelligent than less attractive individuals. This renders them more desirable to be judged and easy to understand, which increases the positivity and the accuracy of personality impressions.

Willis and Todorov [20] also investigated impressions regarding *attractiveness* and *likeability* as well as *trustworthiness*, *competence* and *aggressiveness*. The minimum sufficient condition under which people make a trait inference based on facial appearance was reported to be as small as a tenth of a second (0.1 s). They reported that increasing

the exposure time from 0.1 s to 0.5 s yielded more subjectively satisfying impressions and more confidence in judgements, while increasing from 0.5 s to 1 s only enabled more confidence as the observers' impressions were already anchored on the initial inference. Moreover, increasing exposure time also provided relatively differentiated impressions, i.e., the impressions regarding different traits were found to be less correlated.

3 Engineering Background

Although making accurate personality judgements requires socio-cognitive skills, recently developed computational models can also make valid judgements. Youyou et al. [8] showed that computers' judgements of people's personalities based on their Facebook profiles are more accurate and valid than judgements made by their close friends or families.

There are two strategies coupled with two main problems in automatic personality analysis [6], which are personality recognition (prediction of actual personality) and personality perception (prediction of personality impressions). This paper focuses on personality perception, but we briefly mention the personality recognition trends at the end of this section.

In automatic personality perception, most of the existing methods focused on a subset or all dimensions of the Big Five Model [21], [22], [23]. There are also a number of studies that took into account the dimensions of likeability [24], [25], [26], and physical attractiveness [24], [26] and correlation between these dimensions and the Big Five [26].

When developing automatic analysers, a key challenge is how to generate reliable annotation that is also referred to as ground truth. Similar to psychology, external observers are asked to view a video of the person and rate the person along the Big Five personality dimensions based on thin slices of behaviour ranging from 10 s to several minutes. The rating is usually scaled in seven levels between "strongly disagree" and "strongly agree" (i.e., 7-point Likert scale). However, employing observers to carry out this tedious task is a problem per se. A number of researchers [27], [28] obtained manual annotations through the Amazon Mechanical Turk (MTurk) service. Typically, several folds of independent ratings are run since there is rarely a full agreement between the raters.

In the engineering domain, unlike the psychology domain, little attention has been paid to the impression changes in time and across different contexts. More recently, methods that focus on temporal variability [29] and different situational contexts [26] have emerged. A number of works have also investigated situational factors and time course in the context of personality recognition. Batrinca et al. [30] studied personality recognition across collaborative tasks. Participants instructed by an agent was asked to perform a task on the computer screen where alternately the agent had four different levels of collaboration, from agreeable, stable to less likely to compromise, neurotic. In [31], Pianesi discussed the need for exploring personality states and building computational models that treat the stable traits as a combination of states changing in time. A number of works [29], [32] adopted the concept of personality states and investigated how to model and classify them automatically.

3.1 Related Work

Increasing interest in personality computing has brought about various approaches for automatic analysis. These approaches have been extensively reviewed in a recent survey paper [6]. Here we present an overview of the existing personality perception methods based on the input feature modality utilised, focusing particularly on the audio and vision modalities.

3.1.1 Unimodal Methods

Vision-based methods. Researchers have extensively exploited visual and vocal cues to extract both low-level and high-level features. Among these, [24], [26], [33] focused on the facial cues. Rojas et al. [24] modelled how one is perceived as extroverted, attractive, likeable, dominant, trustworthy, etc. based on still face images. Two low-level features were proposed to represent the face: holistic and structural. Holistic features were extracted from appearance information such as eigenfaces and Histogram of Gradient (HoG), while geometric features were extracted based on the spatial locations of the fiducial facial points (e.g., pairwise distance between points, the spatial relationship between each point and the mean face). Experimental results showed that a reliable prediction (e.g., extroverted versus introverted) was achieved by the holistic representation, in particular HoG, for the traits of dominance, threatening and mean.

Joshi et al. [26] investigated varied situational contexts using human-virtual agent interaction videos from the SEMAINE corpus [9]. External observers assessed the personality of a subject in each interaction that lasted for 14 seconds by providing a score between 1 and 10 for the whole clip. The raters were asked to consider the Big Five traits as well as participants' likeability, facial and vocal attractiveness, and engagement within the interaction based on visual-only displays. Only facial cues were extracted using the pyramids of HoG that counts the gradient orientations both in the whole face and in the localized portions. Mean and standard deviation of the histograms accumulated from all frames were fed into SVMs for regression. The prediction results showed that situational context strongly affects the raters' impressions along the Big Five dimensions, but the perception of attractiveness and likeability does not really change.

High-level features were taken into account by Biel et al. [33] on videos from Youtube, the so-called "video blogs", and annotations generated through a crowd-sourcing service similar to [22], [27]. They detected facial expression of emotions (e.g., anger, happiness, fear, sad) on a frame-by-frame basis and extracted emotion activity cues from sequences either by thresholding or by using a HMM-based method. These features were then fed into SVMs for predicting the five traits. Aran and Gaticia-Perez [23] used Motion Energy Images (MEIs) in a cross-domain learning framework. MEIs from Youtube video blogs [27] were employed to train Rigde Regression and SVR classifiers, and the trained classifiers were tested on small group meeting data for recognizing the trait of *extroversion*.

Audio-based methods. Speaking style (prosody, intonation, speaking rate) is widely represented by low-level features such as signal energy, Mel-frequency cepstral coefficients

(MFCC), pitch, and formants. Other commonly used features are the number of turn takings, speaking time and speaking length. In a prominent work, Mohammadi and Vinciarelli [34] utilised Praat tool to extract prosodic features (pitch, energy etc.) and the length of voiced/unvoiced segments as well as statistical features (maximum, minimum, mean, relative entropy). These features were used in conjunction with Logistic Regression and SVM to classify traits in speech clips from the SSPNet Speaker Personality Corpus. The experimental results demonstrated that *extroversion* and *conscientiousness* were best learned automatically using vocal cues.

3.1.2 Multimodal Methods

Recent methods are characterised by a wide range of multimodal features employed for automatic analysis. In a small group meeting scenario, Aran and Gaticia-Perez [35] used a set of multimodal features including speaking turn, pitch, energy, head and body activity, MEIs and social attention features. Although they obtained the ratings for one minute segments, namely for thin slices only, they extracted the cues both from the whole video and thin slices, and mapped these cues onto the ratings. While thin slices yielded the highest accuracy for *extroversion*, *openness* was better modelled by longer time scales. Similar features were used in [27] and [22] with Youtube video blogs. In [27], Biel et al. made use of speaking activity, prosodic features, looking activity (distance to camera, looking at the camera while speaking) and visual activity (MEIs). In their latter work [22], they exploited verbal content, both singly and jointly, with the same set of features as well as facial expression activity features [33]. On average, they achieved better results with verbal features.

Srivastava et al. [36] were interested in predicting personality of the movie characters. Clips extracted from movies (e.g., Titanic, The Prestige) were rated using a questionnaire. Each clip was represented by audio features (speaking activity, acoustic features), vision based semantic features (six basic emotions) and lexical features (number of words in the dialogue, content – negative or positive – of the dialogue). They proposed a two-tier approach for prediction by firstly mapping the extracted features onto the questionnaire responses using sparse and low-rank transformation (SLoT) and then computing the personality scores from predicted questionnaire responses. Better results were obtained with visual and audio features compared to lexical features. In both stages, fusing the three types of features provided an improvement over the prediction performance.

The studies presented in [29] and [32] are inspired by the work of Fleeson [15], [16]. Staiano et al. [29] addressed the prediction of personality states in four different meeting scenarios using the Mission Survival corpus [37]. Audio-visual recordings were divided into 5-minute long clips. Each clip was interpreted as a personality state annotated along the Big Five dimensions using a 10-item questionnaire. They asked external observers to rate the clips where only the participant under analysis was visible to raters, and the other participants were available through the audio channel. For predicting personality states, they used audio (pitch, amplitude, mean energy, spectral entropy etc.) and video features (social attention features, in particular, attention given, attention received based on the head pose and eye gaze). They modelled the transition from one state to another (e.g., low *extroversion* to high *extroversion* or vice versa) by Hidden Markov Models (HMM) and also classified personality state at a given time frame using Naive Bayes and SVMs. The comparative results showed that, to model *extroversion*, HMM would be a better choice compared to the non-sequential approaches, while for the remaining four dimensions Naive Bayes and SVMs worked better.

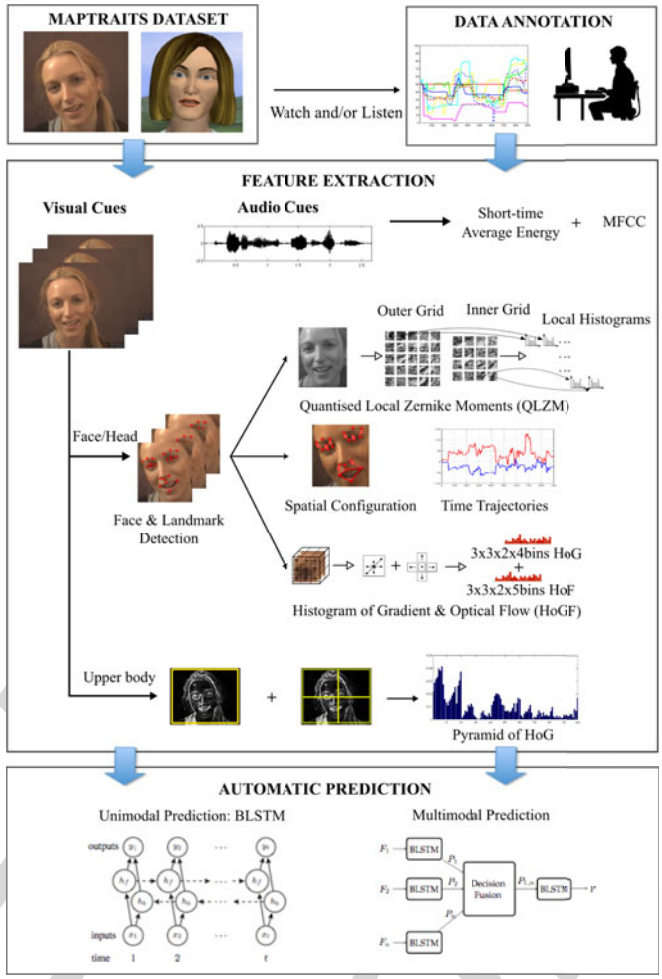

Fig. 1. The overview of the proposed approach for time-continuous prediction of impressions of personality, attractiveness and likeability.

3.2 Overview of Our Work

The block diagram of the proposed approach is shown in Fig. 1. In this paper, based on the findings in [1], [3], [5], [20], we hypothesise that impressions of personality, attractiveness and likeability exhibit variability across different situational contexts and over time. We create an interaction dataset from the available audio-visual recordings of the SEMAINE corpus [9]. We call this dataset the MAPTRAITS Dataset. This dataset consists of 30 clips of 10 subjects interacting with three SEMAINE agents. We propose a novel approach to personality perception modelling and collect a rich set of annotations in terms of personality, attractiveness and likeability as well as the modality of the observed data by asking the raters to provide their impressions continuously in time under

three conditions separately, i.e., visual-only, audio-only and audio-visual. We focus on the dimensions of *agreeableness* (AG), *openness* (OP), *neuroticism* (NE), *conscientiousness* (CO), *extroversion* (EX), *facial attractiveness* (FA), *vocal attractiveness* (VA), and *likeability* (LI). In addition to the Big Five personality traits, *facial attractiveness* describes how attractive the person appears based on the face, *vocal attractiveness* describes how attractive the person appears based on the voice and *likeability* describes how likeable one finds the person in the given context.

The data used in this paper is similar to the work in [26] that also uses the human-virtual agent interactions from the SEMAINE corpus [9]. Joshi et al. [26] is simillar to [29] in that multiple clips of the same target person were considered, but each clip was annotated using a Likert scale. However, in this paper, we examine the temporal variability of personality impressions by developing time-continuous assessment. Rather than obtaining a single rating for the whole clip, raters continuously record their annotations for the aforementioned dimensions as the clip of the target subject plays.

For feature extraction, we take into account a multitude of features from visual and audio cues. We then utilise a time-series regression approach to model the temporal relationships between the continuously generated annotations and extracted features. We further apply decision-level fusion to combine the outputs of the audio and the visual regression models and compare the prediction results when regression models are trained using different modality labels, i.e., labels generated from visual-only, audio-only and audio-visual annotations. We also use the continuously generated annotations to examine which dimensions can be perceived and predicted more dynamically, varying with situational context, and which ones appear to be more static and stable over time.

4 Data, Annotation and Analysis

This section presents the process of creating clips, collecting annotations, generating ground-truth data and statistical analyses of the annotations.

4.1 Data

SEMAINE Corpus [9] provides a rich collection of people interacting with virtual agents in a naturalistic scenario. We took into account 10 different subjects. Each subject interacts with three Sensitive Artificial Listener (SAL) agents, namely, *Poppy*, *Obadiah* and *Spike*, resulting in 30 video recordings. To reduce the burden on the raters, we shortened and segmented each recording (approximately 5 minutes-long) into a 60 s clip containing several instances of turn taking. The 60 s length was found to be sufficiently long to capture personality impression changes and was reasonable for obtaining effective annotations. In [5], it was also indicated that 60 s yielded the optimal ratio between obtaining accurate impressions and slice length.

Each SAL agent has a specific character and accordingly exhibits stable behaviours driven by emotions. *Poppy* is always cheerful and positive, *Spike* is always angry and aggressive and *Obadiah* is always sad and miserable. The situational context created by each agent brings about a different behavioural act that may change over time.

4.2 Annotation

We conducted the annotation by designing and using an in-house tool [38] that functions similarly to GTrace [39]. The annotation tool requires the rater to scroll a bar between a range of values from 1 to 100 as the recording plays, but without pressing constantly, and stores the rating values at every pre-set time interval (please refer to Section 2.2 of [40] for more information). Each annotation resulted in a 60 s-rating-trajectory that is a sequence of values (between 1 and 100) showing how the impressions change as a function of time.

The clips were rated with respect to the Big Five personality traits, attractiveness and likeability by 21 paid raters - aged between 23 and 53 years ($mean = 29$) from different ethnic backgrounds. For each of the Big Five personality traits, the raters judged the target with respect to four adjectives selected from [41]. We asked the raters to indicate how much they agree or disagree with the provided adjectives regarding the target person, and for each clip a rater annotated one dimension at a time. Additionally, we asked the raters to scroll the bar as the clip plays, when they think that their impressions change. Apart from these, we did not give any particular instructions to the raters.

To investigate what kind of information source (e.g., voice, speaking tone, appearance, gestures) plays an important role in the perception, the annotations were collected under three conditions: visual-only, audio-only and audio-visual.

Visual-Only Annotation (VO). In visual-only annotation, only the visual channel was available to the raters (audio tracks were removed). Since annotation along one dimension (30 video clips at once) takes approximately 45 minutes per rater, we also divided the dimensions into two separate groups: i) *agreeableness*, *openness*, *likeability*, *neuroticism*, and *conscientiousness*; and ii) *extroversion*, *facial attractiveness*, *neuroticism*, and *conscientiousness*. We nevertheless asked both of the groups to annotate *conscientiousness* and *neuroticism* as these have been found to be the most challenging dimensions to understand and annotate by the raters. A total of 16 raters (nine females, seven males) annotated 30 video clips with respect to the seven dimensions (Big Five+2), which resulted in 7-10 annotations per clip, per dimension.

Audio-Only Annotation (AO). Contrary to visual-only annotation, the focus of this annotation task was only the audio channel (human subjects were not visible to the raters). A total of six raters (two females and four males) were divided into two groups and each group was asked to annotate the speech clips with respect to a subset of dimensions among the Big Five traits, *vocal attractiveness* and *likeability*. This yielded three annotations for each clip, for each dimension.

Audio-Visual Annotation (AV). Audio-visual annotation complements the annotation conditions mentioned above in that raters annotated the video clips taking into account both visual and vocal cues for all eight dimensions (Big Five+3). To obtain unbiased assessments, we employed five raters different from the visual-only and the audio-only rater pool and selected four raters out of the visual-only rater pool, provided that they performed annotation along the unseen portion of the dimensions. A total of nine raters (four females, five males) assessed all clips, resulting in five annotations per clip, per dimension.

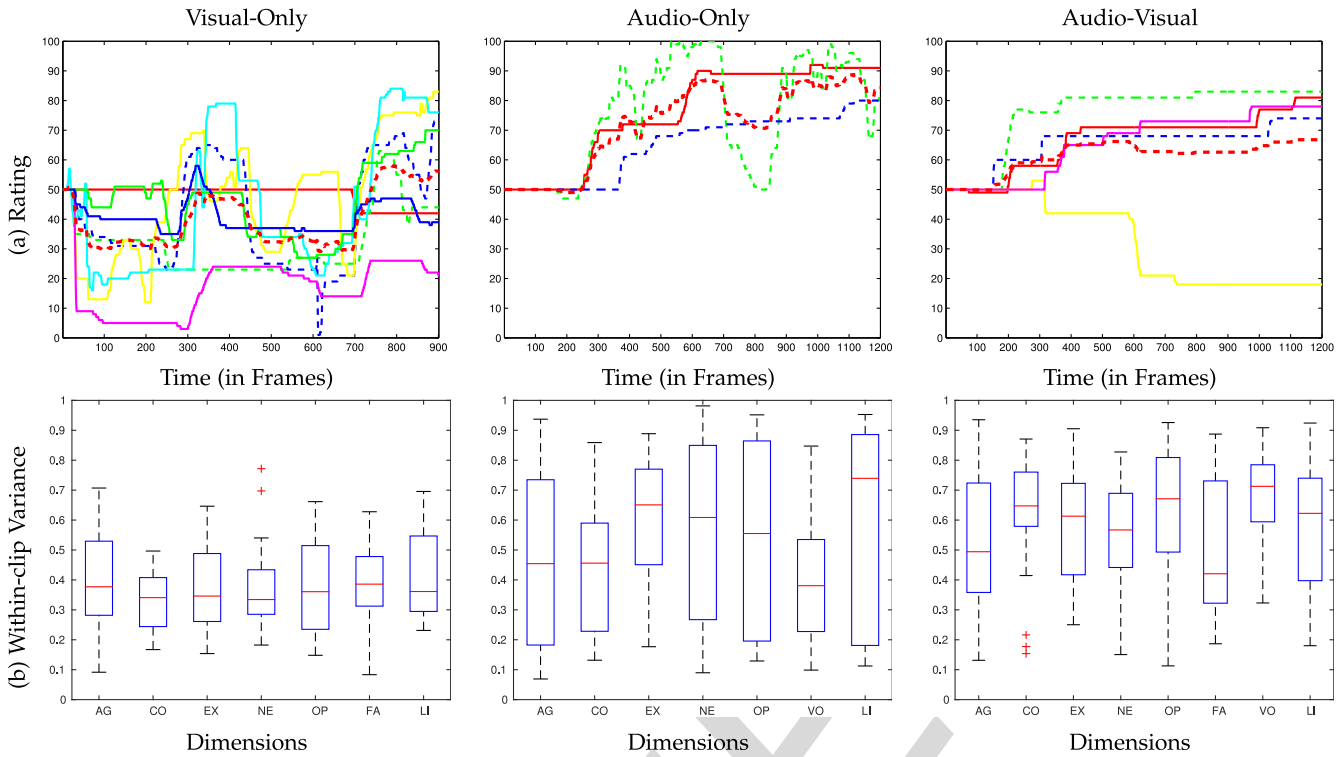

Fig. 2. (a) Continuous *agreeableness* annotations in time provided for one clip under three different annotation conditions: visual-only (left), audio-only (middle) and audio-visual (right). Red dashed line illustrates the mean trajectory of the time-continuous annotations (i.e., ground truth). (b) Distribution of within-class variance values per dimension for three annotation conditions as a box-and-whisker plot: visual-only (left), audio-only (middle) and audio-visual (right).

In our experiments, we set the time interval smaller than 100 ms to capture the slightest changes in the impressions [20]. We defined the intervals as 65 ms and 50 ms for visual-only and audio-visual/audio-only annotations, respectively, where the different time intervals were indeed the requirement of the annotation tool. Representative rating trajectories are illustrated in Fig. 2a for the *agreeableness* dimension, for one clip in three annotation conditions. One can observe that audio-visual/audio-only annotations yielded smooth trajectories over time, hence consensus among many raters is more obvious. However visual-only raters seem to hardly agree, yet some of the judgements show similar trends.

4.3 Analysis of Annotations

This section provides detailed analyses of the time-continuous annotations in terms of consensus among the raters, changes over time, impact of situational context and correlation between the dimensions.

4.3.1 Consensus Among the Annotators

A key challenge in designing intelligent user interfaces is establishing a reliable ground truth from multiple raters. Especially, in the case of continuous ratings, this has proven to be extremely difficult due to missing data, and variations in the speed and style of the raters, e.g., time lags may occur in responding to the conveyed cues or internal rating scales can drastically differ among the raters. In the literature, a common approach is to extract raters' trends, in other words, to compare two ratings in relative terms rather than in absolute terms, e.g., whether there has been a rise, fall or level stretch [42].

Prior to any analysis, we apply z-score normalization to mitigate the effects of different internal rating scales where each annotation is normalized with respect to its mean and standard deviation. Correlation-based approaches such as Cronbach's α coefficient have been widely used to measure the degree of agreement among multiple raters (i.e., inter-rater agreement or consensus) in the literature. However, it is not straightforward to use these approaches in the case of time-varying data. Dynamic Time Warping (DTW) not only permits comparison by a shifting operation, but also incorporates warping operations such as insertion and deletion. Therefore we apply DTW to align two rating trajectories. The DTW algorithm searches for the best correspondence between two trajectories that minimizes the sum of cumulative distances. In our experiments, we set the locality constraint to 2 s.

After each annotation pair is aligned using DTW, we measure the agreement in terms of Pearson's correlation and Cronbach's alpha. We also eliminate the outliers by using the following strategy. Assume we have K annotations per clip, i. e, $\{y_1, \ldots, y_K\}$. We first compute the pairwise correlations between each annotation y_i and the remaining $K-1$ annotations, $\{y_j\}_{j \neq i}$. If only the mean of its pairwise correlations is greater than a threshold, we take into account y_i when computing the ground-truth for the corresponding clip. We set the threshold such that at least three reliable raters are considered per video clip. The average number of *selected* raters is 7.5 (std = 1.4, min = 3, max = 11) and 4.4 (std = 0.6, min = 3, max = 5) per clip per dimension for the visual-only condition and for the audio-visual condition, respectively. Note that there are only three raters per speech clip in the audio-only condition, we therefore took into account all of the annotations.

TABLE 1
Measure of Agreement among the Selected Raters in Terms of Mean Pearson's Correlation (ρ) and Mean Cronbach's Alpha (α) across Different Modalities

	Visual-only		Audio-only		Audio-visual	
	ρ	α	ρ	α	ρ	α
AG	0.47(0.40)	0.85(0.81)*	0.27	0.01	0.47(0.30)	0.70(0.48)†
CO	0.38(0.17)	0.79(0.61)*	0.24	0.22	0.56(0.35)	0.77(0.47)*
EX	0.46(0.39)	0.85(0.81)*	0.47	0.64†	0.53(0.43)	0.78(0.62)*
NE	0.44(0.35)	0.87(0.82)*	0.41	0.30	0.49(0.18)	0.75(0.21)*
OP	0.42(0.27)	0.80(0.70)*	0.35	0.12	0.56(0.22)	0.71(0.10)*
FA	0.45(0.28)	0.82(0.70)*	-	-	0.43(0.26)	0.68(0.42)†
VA	-	-	0.20	0.17	0.63(0.36)	0.85(0.56)*
LI	0.46(0.36)	0.84(0.70)*	0.05	0.05	0.50(0.19)	0.72(0.31)*

*The values in the parentheses indicate the level of consensus before eliminating the outliers. While * indicates good internal consistency ($0.7 \leq \alpha < 0.9$), † indicates acceptable values ($0.6 \leq \alpha < 0.7$). AG: Agreeableness, CO: Conscientiousness, EX: Extroversion, NE: Neuroticism, OP: Openness, FA: Facial Attractiveness, VA: Vocal Attractiveness, LI: Likeability.*

In Table 1, we tabulated the degree of agreement among the *selected* raters with respect to each dimension under three conditions. Each value in the parenthesis indicates the inter-agreement before eliminating the outliers. One can observe that this approach yielded a significant increase in the level of consensus both in visual-only and audio-visual conditions. After the annotation task took place, we asked each rater which trait was difficult to judge. The visual-only raters mostly agreed that *conscientiousness* was the most difficult one. Our analysis also validates this comment as we obtained the lowest consensus for *conscientiousness* in visual-only modality. While the visual-only raters also found *agreeableness* and *openness* challenging, the audio-only and audio-visual raters reported that they generally felt confident in their observations.

Once the reliable raters were determined, we generated the ground truth by evaluating the mean of the selected rating trajectories per video/speech clip. The mean trajectory (the red dashed line) amounts to the ground-truth as illustrated in Fig. 2a.

4.3.2 Variation in the Impressions

In order to examine the variation in the impressions formed by the observers in time, we considered the generated ground-truths (mean trajectories) and presented the within-clip variance (i.e., their variances over time per clip) across different annotation conditions for each dimension in Fig. 2b. Average of the within-class variances over all dimensions are found to be higher for audio-only annotation ($\sigma^2_{av} = 0.50$) and for audio-visual annotation ($\sigma^2_{av} = 0.58$) as compared to visual-only annotation ($\sigma^2_{av} = 0.37$).

The raters of the audio-visual condition mostly agreed that *conscientiousness* and *openness* have a static characteristic as they claimed that, once they made their decision, their impressions hardly changed for the rest of the clip. On the other hand, *extroversion* is found to be dynamic by all raters. Fig. 2b is in line with the raters' feedback because within-class variance for *conscientiousness* is lower and more compact in visual-only annotation and in audio-only condition as compared to the other dimensions. The other dimensions that have low within-class variation are *neuroticism* in visual-only annotation and *vocal attractiveness* in audio-only annotation, which are also less likely to vary in absolute values over time.

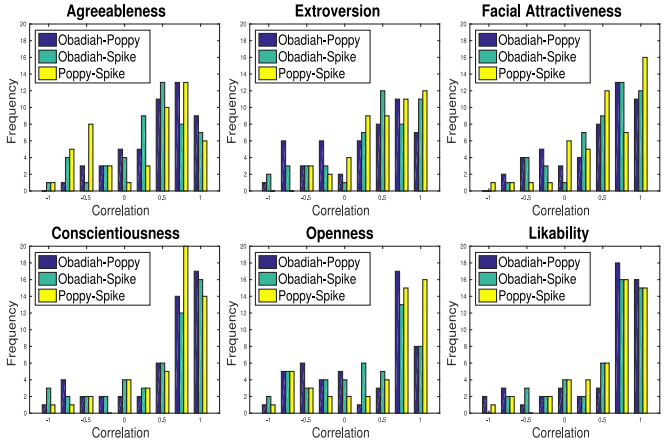

Fig. 3. Mean correlations for all subject's annotations per dimension (audio-visual condition). Correlations between multiple annotations differ across different subjects and different agents.

4.3.3 The Effect of Situational Context

We also examined the effect of different situational context, namely, interaction with each virtual agent (*Poppy*, *Obadiah* and *Spike*), on the raters' impressions. Fig. 3 shows a histogram of the correlations between each rater's annotations for three context (agents) for all target subjects with respect to *agreeableness*, *extroversion*, *facial attractiveness*, *conscientiousness*, *openness* and *likeability*. We computed the correlations between the annotations of the same subject's three clips separately for each rater, where in each clip the subject interacts with a different virtual agent. Namely, we presented the correlations between the annotations for *Obadiah* and *Poppy*, *Obadiah* and *Spike*, and *Poppy* and *Spike*, and observed how each rater's impressions change from one interaction clip to another (e.g., from *Obadiah* to *Poppy*). For *conscientiousness*, *openness* and *likeability* the correlations are centred around larger values and their extent is small. On the other hand, they are centred at smaller values and spanned over a larger range of values for *agreeableness*, *extroversion* and *facial attractiveness*. This confirms that the raters' impressions differ depending on the human-virtual agent interaction context and the dimension assessed. We observed similar trends in visual-only and audio-only annotations as well, especially, for the impressions along *openness* and *likeability* in the visual-only annotation.

4.3.4 Correlation between the Dimensions

In this experiment, we investigated whether there are any relationships between the different dimensions. We measured the correlation between the annotations along pairs of dimensions using the same approach introduced in Section 4.3.1. We present the significant correlations in Table 2. One can observe that *facial attractiveness* and *likeability* are highly correlated with each other as well as with *agreeableness*, *openness* and *extroversion*. This can be explained by the "Halo Effect" [43], i.e., the raters tend to assign good attributes to the person they like or find attractive. Especially, *likeability* shows high positive correlation with *agreeableness*. Among the Big Five dimensions, *agreeableness*, *extroversion* and *openness* are

TABLE 2
Significant Correlations among the Dimensions in the Visual-Only Condition (VO), in the Audio-Only Condition (AO) and in the Audio-Visual Condition (AV)

Visual-Only	Agreeableness-Openness	0.42
	Agreeableness-Likeability	0.44
	Facial Attractiveness-Likeability	0.42
Audio-Only	Agreeableness-Likeability	0.41
	Extroversion-Likeability	0.43
	Extroversion-Openness	0.40
	Openness-Likeability	0.46
Audio-Visual	Agreeableness-Openness	0.44
	Conscientiousness-Vocal Attractiveness	**0.57**
	Facial Attractiveness-Likeability	0.40

All correlations found to be significant ($p < 0.05$). The value in bold represents strong positive relationship ($^*p < 0.01$).

the ones that are highly correlated with each other. Unlike what was reported in [18], we could not observe any significant differences in the correlations between the dimensions across visual-only and audio-only/audio-visual conditions. Similar patterns of correlation seem to occur regardless of the annotation condition.

5 Feature Extraction

In the literature, a multitude of features have been proposed and used for describing and measuring human behaviour. For visual cues, we are motivated by approaches for recognising face/head gestures [44] and for predicting affective states [45] in video sequences. In particular, we captured the face/head and body movements considering both spatial and spatio-temporal appearance features (e.g., Zernike moments, gradient and optical flow) and geometric features (e.g., spatio-temporal configuration of facial landmark points). In addition to visual cues, we represented the audio cues using the well-known features such as short-term average energy and Mel-Frequency Cepstral Coefficients.

5.1 Visual Features

We first detected and tracked 49 landmark points per frame using the face landmarking tool developed by Xiong and De la Torre [46]. It applies Supervised Descent Method for nonlinear least squares problems and Scale Invariant Feature Transform (SIFT) features for face alignment. For feature extraction, we only considered a subset of landmarks that play a prominent role in identifying face gestures. This subset consists of 21 landmark points including eye corners, eyebrow corners, eye lids, nostril and mouth corners. We further used the tracked landmark points to capture face, head and body movements. In the following subsections, we describe the details of the visual features extracted under three categories: appearance, geometric and hybrid features.

5.1.1 Appearance Features

We considered two types of appearance features, for describing the face activity and the body activity.

Face Activity. We used the tracked landmark points to determine a rectangle enclosing the face, to crop faces based on these rectangles, and to align the faces based on the coordinates of the eye centers using affine transformation. The cropped and aligned faces were resized such that each face has the size of 128×128. For each frame, we computed the histograms of Quantised Local Zernike Moments (QLZM) [47]. We first calculated and quantised a set of Zernike Moments in the neighbourhood of each pixel of an input face image where each ZM describes local appearance variation at a unique scale and orientation, and formed a QLZM image. The QLZM image was then divided into subregions with respect to two grids, an inner partition and an outer partition. The double partition aims to mitigate the errors due to face alignment. A position-dependent histogram was computed for each subregion, and each face was represented by concatenating these local histograms. In our experiments, we partitioned the face by applying a 5×5 outer grid and a 4×4 inner grid, and considered two ZMs that yield a 16-bin histogram. This resulted in a 656-length feature vector per face (or frame) and a $656 \times T$ feature matrix per clip, where T is the number of frames in a clip.

Body Activity. The coordinated movement between head and shoulders, and postural changes form a rich source of information for understanding human behaviour. To capture these bodily cues, we detected and tracked the box enclosing the upper body over T frames. We used the off-the-shelf Calvin upper detector [48] to determine the candidate boxes enclosing the upper body. Calvin upper body detector [48] searches for upper bodies within an image by using a sliding window based on the deformable part based models [49], where each part is described by HoG and classified using SVMs. As in [48], we refined the upper body boxes by combining them with the face detection using the landmark locations. We encoded the dynamics of the upper body by extracting and collecting pyramids of HoGs [50] over T frames. The pyramid of HoGs [50] extends the classical HoG [51] by a hierarchical spatial representation. First, a HoG is computed for the whole image which is then divided into four non-overlapping blocks. At each level, it recursively divides the blocks into sub-blocks, each arranged in a 2×2 grid, and computes HoG for each block. The final representation is obtained by concatenating the position-dependent HoGs from different levels. In our experiments, we considered 8-bin orientation histograms in a one-level pyramid. This resulted in a 40-length feature vector per frame and a $40 \times T$ feature matrix per clip.

5.1.2 Geometric Features

We extracted two types of geometric features based on the time *trajectory* and the spatial *configuration* of the landmark points. Each landmark point generates a motion pattern in space and in time that can be used to simultaneously capture eye/eyebrow/mouth movements (e.g., eye blinking, eye raising, smiling) and head movements (e.g., head nodding, shaking). To model these motion patterns, we used the spatial and temporal relative distances between the landmark points as proposed in [11], [44]. First, the 21 landmarks points tracked over T frames were stored into a $42 \times T$ trajectory matrix where each column corresponds to the x and y coordinates of the landmark points. To render the landmark point trajectories independent from their initial position, we took into account the relative displacements of the landmark points with respect to the first frame by subtracting the first column of the trajectory matrix from every column. Secondly, we computed 11 distance

measures between pairs and groups of landmark points and accumulated these distances over T frames into a $11 \times T$ distance matrix in order to capture the eye/eyebrow, mouth and head movements from one frame to another. For example, eye/eyebrow actions such as eye blinking, eyebrow raising and head movements such as forward/backward movement, and head yaw can be described in terms of the relationships between eyelid centres, eye/eyebrow corners, and mouth shape. Speaking activity can be inferred by the configuration of the mouth landmark points. The pairwise distances between the landmark points are not scale invariant, therefore we normalized the distances with respect to the inter-ocular distance (distance between the inner eye corners) in each frame.

5.1.3 Hybrid Features

As an alternative descriptor for face activity, we combined local appearance and motion information around the facial landmark points. More explicitly, we computed Histogram of Gradient and Histogram of Optical Flow (HoF) in the spatio-temporal neighbourhood of the landmark points and concatenated these histograms into a single feature vector (HoGF). Extension of HoG and HoF to the temporal domain results in a position dependent histogram [52]. The local neighbourhood of a detected point is divided into a grid with $M \times M \times N$ (i.e., $3 \times 3 \times 2$) spatio-temporal blocks. For each block, 4-bin gradient and 5-bin optical flow histograms are computed and concatenated into a 162-length feature vector. We calculated HoGF values for left/right eye centers and mouth centers at two spatial levels, i.e., we considered two spatial scale parameters, $\sigma^2 = 4, 8$, and set the temporal scale parameter to 2, $\tau^2 = 2$. This resulted in 972-length feature vector per frame and, by accumulating over T frames, a $972 \times T$ feature matrix per clip. Rather than a hand-crafted representation, these features provide a unified representation for the local information of the facial parts (e.g., eyes, mouth) both in the space and the time domain.

5.2 Audio Features

Mel Frequency Cepstral Coefficents (MFCCs) and short-time average energy (STAE) are essential features in automatic speech and speaker recognition, and their viability has been frequently demonstrated for affect analysis [11] and personality trait analysis [6]. MFCCs can be interpreted as a speech signature. We extracted the MFCC features by using the Praat tool [53] that has been widely used in automatic affect recognition. In our study, we applied a 40 ms-long window with a time step of 20 ms. We selected 12 MFCC features and, based on the selected MFCC features, computed delta MFCC and autocorrelation MFCC features as follows. Let $MFCC_\nu(i)$ be the ν^{th} MFCC coefficient of the time segment i, delta MFCC features are calculated as $\Delta MFCC_\nu(i) = \Delta MFCC_\nu(i) - \Delta MFCC_\nu(i+1)$ and autocorrelation MFCC features as $ACMFCC^l_\nu(i) = \frac{1}{L}\sum_{j+i}^{i+L}(MFCC_\nu(j) \cdot MFCC_\nu(j+l))$ where L and l are the correlation window length and the correlation lag, respectively. In addition to MFCC features, we computed the short-time average energy. We set the length of time window to 40 ms in our experiments. In total, this resulted in 37 audio features per frame and, for T frames, a $37 \times T$ feature matrix per audio clip.

6 Automatic Prediction

We employed the extracted visual and audio features (Section 5) to train a separate regression model for each dimension. We modelled the time-continuous nature of the audio-visual behavioural data and inferred the rating trajectory using Bidirectional Long Short-Term Memory Networks (BLSTM) [54] that have been widely used for time-series prediction in automatic speech recognition [55], emotion classification [56] and continuous affect prediction [11].

6.1 Unimodal Prediction

We employed Bidirectional Long Short-Term Memory Networks (BLSTM) [54] to establish a relationship between a target rating trajectory (e.g., each has a size of $1 \times T$) and the input features extracted from the whole clip (e.g., each has a size of $d \times T$). Long Short-Term Memory Networks (LSTM) [57], [58] are an extension of Recurrent Neural Networks (RNN) that are well able to deal with sequential patterns. Unlike traditional artificial neural networks, RNNs include self-connected hidden layers that allow mapping from the current time instant to the output by taking into account preceding inputs (past) and also the succeeding inputs (future). The term bidirectional indicates reaching both directions (past and future) that is achieved by two separate hidden layers scanning the input sequences forward and backward. One drawback of RNNs is the vanishing gradient problem [59], [60] - the back-propagated error decays or blows up exponentially in time, which restricts the extent of access to past and future inputs. LSTM was specifically developed to remedy such shortcomings of RNNs.

An LSTM hidden layer (memory blocks), consists of one or more recurrently connected cells whose activation is controlled by three multiplicative gates, i.e., the input gate, forget gate and output gate. These gates enable the cells to store and access the past and future inputs over longer time scales. While the forget gate controls the recurrent connection of the cell, the input and output gates control the input and output of the cell. As long as the input gate is closed, the new inputs will not be transferred to the activation of the cell. Similarly, the activation of the cell can be made available to the rest of the network by opening the output gate. The network structure representing Bidirectional LSTM (BLSTM) employed in this work is provided in Fig. 1 (see Unimodal prediction: BLSTM). The advantage of BLSTM is that its hidden layers are designed to encode the sequential relationship over longer time scales and its structure allows to take into preceding inputs (past) and the succeeding inputs (future) when predicting the current output [61].

In Section 5, we introduced six feature types, namely, visual QLZM, HoGF, configuration, trajectory, PHoG and MFCC+STAE. In our experiments, we used these features to train unimodal regression models using both unimodal labels (visual-only/audio-only) and multimodal labels (audio-visual). More explicitly, *unimodal prediction with unimodal labels* learned a separate regression model for unimodal features and unimodal labels by modelling the relationship between visual features and visual-only labels, and between audio features and audio-only labels. *Unimodal prediction with multimodal labels* learned a separate regression model for unimodal features and multimodal labels by modelling the relationship between visual features and

TABLE 3
Unimodal Prediction Results

			Unimodal Prediction with Unimodal Labels								Unimodal Prediction with Multimodal Labels							
			AG	CO	EX	NE	OP	AT	LI	av	AG	CO	EX	NE	OP	AT	LI	av
Face	QLZM	R^2	0.05	0.08	0.04	0.05	0.05	0.08	0.06	0.06	0.18	0.33	0.06	0.23	0.11	0.08	0.18	0.17
		MSE	0.48	0.42	0.47	0.41	0.45	0.43	0.47	0.45	0.51	0.49	0.68	0.49	0.75	0.54	0.55	0.57
	HoGF	R^2	0.08	0.07	**0.13**	0.13	0.05	0.13	0.13	0.10	0.22	**0.41**	0.07	0.20	0.23	0.14	0.19	0.21
		MSE	0.48	0.38	**0.44**	0.34	0.47	0.37	0.45	0.42	0.53	**0.38**	0.62	0.57	0.75	0.45	0.60	0.56
Face & Head	Trajec.	R^2	0.04	0.02	0.04	0.12	0.04	0.08	0.12	0.06	0.21	0.30	0.08	0.20	0.23	0.08	0.17	0.18
		MSE	0.44	0.31	0.40	0.41	0.45	0.36	0.39	0.39	0.50	0.43	0.65	0.47	0.61	0.45	0.62	0.53
	Config.	R^2	**0.10**	0.05	0.09	0.09	0.02	0.06	0.10	0.07	**0.23**	0.33	0.07	0.20	0.13	**0.16**	**0.25**	0.20
		MSE	**0.38**	0.31	0.36	0.40	0.39	0.38	0.41	0.38	**0.47**	0.44	0.55	0.45	0.71	**0.41**	**0.45**	0.50
Body	PHoG	R^2	0.07	0.07	0.02	**0.17**	0.07	**0.16**	**0.14**	**0.10**	0.19	0.39	**0.10**	**0.26**	**0.25**	0.14	0.22	**0.22**
		MSE	0.46	0.35	0.44	**0.30**	0.39	**0.32**	**0.39**	**0.38**	0.60	0.39	**0.69**	**0.45**	**0.59**	0.43	0.50	**0.52**
Audio	MFCC+ STAE	R^2	0.09	0.06	0.12	0.05	0.09	0.07	**0.14**	0.09	0.10	0.27	0.05	0.17	0.19	**0.31**	0.21	0.18
		MSE	0.44	0.49	0.59	0.67	0.63	0.43	**0.54**	0.54	0.51	0.47	0.61	0.45	0.56	**0.47**	0.52	0.51

The best prediction results per dimension are highlighted in bold for unimodal prediction with unimodal labels and unimodal prediction with multimodal labels. AG: Agreeableness, CO: Conscientiousness, EX: Extroversion, NE: Neuroticism, OP: Openness, AT (visual-only): Facial Attractiveness, AT (audio-only): Vocal Attractiveness, LI: Likeability, av: average over all dimensions.

audio-visual labels, and between audio features and audio-visual labels.

6.2 Multimodal Prediction

Multimodal prediction, in our case, is based on decision-level fusion, and it combines visual features and audio features at the decision level. Most of the methods in the literature (e.g., [27], [35], [62]) pooled visual and audio features and fed them into one classifier or regressor (i.e., feature-level fusion). We opted for decision level fusion as the features from different modalities have different representations yet decisions all have similar representation. This renders the decision-level fusion more applicable and straightforward than feature-level fusion in our case. As shown in Fig. 1 (see Multimodal Prediction), we combined the predictions from each unimodal regression model into a matrix which was then fed into the BLSTM for the final prediction. Our approach can also be interpreted as a hierarchical regression in that, at the first step, each feature type is treated separately, and at the second step, the individual predictions from different models are fused.

In our experiments, we combined each type of the visual features with the audio features at the decision level (e.g., HoGF and MFCC+STAE, PHoG and MFCC+STAE, etc.). More explicitly, we fused the unimodal prediction outputs and mapped them onto the multimodal labels.

7 Experiments and Results

In this section, we examined prediction results with respect to the role of annotation conditions, features, decision-level fusion and situational contexts for automatic prediction.

7.1 Experimental Setup and Evaluation Metrics

Prior to any analysis, we applied feature normalisation so that the range of feature values were rescaled to $[-1, 1]$. We learned the optimum parameters for BLSTM by using the *leave-one-subject-out* cross validation strategy, where in each fold we used 27 clips for training and validation, and the remaining three clips for testing. We used the same training parameters as proposed in [11], i.e., we used a network with one hidden layer and set the learning rate to 10^{-4}. The optimum momentum parameter for each dimension was selected from the range of values ($[0.5, 0.9]$).

We used two metrics for experimental evaluation and performance comparison of the methods introduced in Section 6, namely, coefficient of determination (R^2) and mean square error (MSE). R^2 measures how well the learned model fits the unseen samples and yields a value between 0 and 1 where larger values indicate better fit. MSE gives the average of the squared errors. Since we applied z-score normalization when generating the ground-truth, MSE values can vary between 0 and 4. These metrics are widely used for prediction and are described in detail in [40].

7.2 Prediction Results

Experimental results for the proposed regression approaches described in Section 6 are given in Tables 3 and 4. We considered the best result to be the prediction yielding the maximum coefficient of determination (R^2) and the minimum error (MSE). Looking at Tables 3 and 4, all dimensions have been successfully predicted using the proposed time-series regression approach ($R^2 > 2$ and $MSE < 0.6$) with the exception of *extroversion* and *facial attractiveness*. This result is especially surprising for *extroversion* as this trait has been the easiest trait to recognise/predict in the literature. We further examined the prediction results with respect to the role of different labels (visual/audio-only versus audio-visual), features in predicting each dimension and the contribution of the decision-level fusion.

7.2.1 The Role of Different Labels and Features

Unimodal labels versus Multimodal labels. Table 3 compares two unimodal prediction approaches, namely, unimodal prediction with unimodal labels and unimodal prediction with multimodal labels, with respect to each feature type. One can observe that the proposed time-series regression approach yielded superior prediction results when trained with audio-visual labels ($R^2_{av} = 0.22$ with PHoG features) as compared to when trained with visual-only/audio-only

labels ($R^2_{av} = 0.10$ with PHoG features). Personality perception was modelled better in the presence of more information. Learning with audio-visual labels especially benefited the prediction of *conscientiousness* and *openness*. Raters emphasized that these dimensions were very difficult to annotate without audio. As expected, the prediction results significantly improved for *conscientiousness* and *openness* up to $R^2_{CO} = 0.41$ and $R^2_{OP} = 0.25$, respectively, while we obtained R^2 values lower than 0.1 with visual-only labels. On the other hand, for *extroversion*, mapping visual features onto visual-only labels yielded slightly better results ($R^2_{EX} = 0.13$ with HoGF features) as compared to audio-visual labels ($R^2_{EX} = 0.10$ with PHoG features). None of the dimensions were successfully modelled using audio features (MFCC+STAE) and audio-only labels except for *extroversion* and *likeability* ($R^2_{EX} = 0.12$ and $R^2_{LI} = 0.14$). This might be due to the fact that we had less number of raters in the audio-only annotation than in the visual-only and the audio-visual annotations and therefore lower level agreement was obtained between raters. Table 1 also validates this - all Crobach's alpha values are found to be lower than 0.6 in the audio-only annotation with the exception of *extroversion* dimension.

Features. Taking into consideration the results of unimodal prediction with multimodal labels in Table 3, PHoG body features ($R^2_{av} = 0.22$ and $MSE_{av} = 0.52$) work best in general for time-continuous prediction where HoGF ($R^2_{av} = 0.21$ and $MSE_{av} = 0.56$) and configuration ($R^2_{av} = 0.20$ and $MSE_{av} = 0.50$) are among the face features that can be considered as a runner-up. QLZM face features yielded the worst prediction results ($R^2_{av} = 0.17$ and $MSE_{av} = 0.57$), which might be due to the fact that QLZM requires a preprocessing stage where the faces are aligned and cropped based on the location of the eye centres. This method is simple but prone to localisation errors. Any misalignment might cause deteriorations in prediction accuracy for the time-series regression which relies on learning the temporal dependencies between frames rather than treating each frame separately.

While PHoG body features worked better for the prediction of *neuroticism*, *openness* and *extroversion*, the best prediction for *facial attractiveness* and *likeability* was achieved using the configuration of landmark points. This finding confirms what has been reported in the related literature on facial beauty analysis [63], [64] and *facial attractiveness* analysis [65] - facial proportions and features extracted based on geometry play an important role in assessing attractiveness. *Conscientiousness* was best predicted by facial cues (HoGF) where $R^2_{CO} = 0.41$ and $MSE_{CO} = 0.38$. This result may be due to the fact that the raters might have focused on the face activity changes rather than focusing on the global appearance changes (body cues). Another important relationship was observed between audio features and audio-visual labels for *vocal attractiveness* where $R^2_{AT} = 0.31$ and $MSE_{AT} = 0.47$.

7.2.2 Impact of Decision-Level Fusion

In Table 4, we only present three pairwise combinations that yielded the maximum coefficient of determination (R^2) and the minimum error (MSE) among all possible combinations of features. We compare them with the best results of unimodal prediction with multimodal labels in Table 3, which are given in Table 4 as well. Combining configuration or PHoG features with audio features is found to be the best solution for predicting *conscientiousness*, *neuroticism* and *likeability*. In general, prediction results are slightly improved when configuration features are combined with audio features at the decision level ($R^2_{av} = 0.24$ and $MSE_{av} = 0.51$). However, only for *openness* and *extroversion*, unimodal prediction approach works better as compared to multimodal prediction. This might be due to the fact that visual cues that are conveyed and perceived in the course of impression formation play a more dominant role in predicting *extroversion*.

TABLE 4
Multimodal Prediction Results

		AG	CO	EX	NE	OP	LI	av
Unimodal Prediction	R^2	**0.23**	0.41	**0.10**	0.26	**0.25**	0.25	0.22
	MSE	**0.47**	0.38	**0.69**	0.45	**0.59**	0.45	0.52
HoGF + MFCC+STAE	R^2	0.17	0.44	0.05	0.16	0.18	0.19	0.20
	MSE	0.54	0.35	0.76	0.49	0.64	0.55	0.55
Config. + MFCC+STAE	R^2	**0.24**	0.39	0.04	**0.32**	0.15	**0.28**	**0.24**
	MSE	**0.49**	0.42	0.69	**0.39**	0.61	**0.44**	**0.51**
PHoG + MFCC+STAE	R^2	0.17	**0.50**	0.06	0.26	0.23	0.23	0.24
	MSE	0.49	**0.34**	0.76	0.48	0.57	0.51	0.52

The best prediction results per dimension are highlighted in bold. AG: Agreeableness, CO: Conscientiousness, EX: Extroversion, NE: Neuroticism, OP: Openness, LI: Likeability, av: average over all dimensions.

7.2.3 Effect of Situational Context

We also investigated the effect of different situational context, namely, interaction with the three different virtual agents (*Poppy*, *Obadiah* and *Spike*), on the automatic prediction results. Fig. 4 illustrates the automatic prediction results with respect to each virtual agent in terms of R^2. For this analysis, we took into account the multimodal prediction results with configuration features and audio features in Table 4, which provided the highest $R^2_{av} = 0.24$ and the lowest $MSE_{av} = 0.51$. In general, a better relationship between the automatic prediction and the situational context was established for *Poppy*. This shows that people display more visible and observable cues when interacting with *Poppy* and then *Obadiah*, but are less expressive in their behaviours when interacting with *Spike*. Interactions with *Poppy* and *Obadiah* are more prominent especially for the *likeability* dimension.

8 Discussion and Conclusions

This paper proposed a novel multimodal framework for automatically predicting the impressions of *agreeableness*, *openness*, *neuroticism*, *conscientiousness*, *extroversion*, *facial attractiveness*, *vocal attractiveness* and *likeability* continuously in time. We aimed at exploring the variability in impressions across different communication channels (i.e., visual-only, audio-only and audio-visual) and across different situational contexts.

Conclusions. Our work contributes to the existing literature of personality computing in multiple ways. We obtained continuous annotations where external observers watched or listened to clips of an individual subject and provided their impressions of a given dimension as the clip

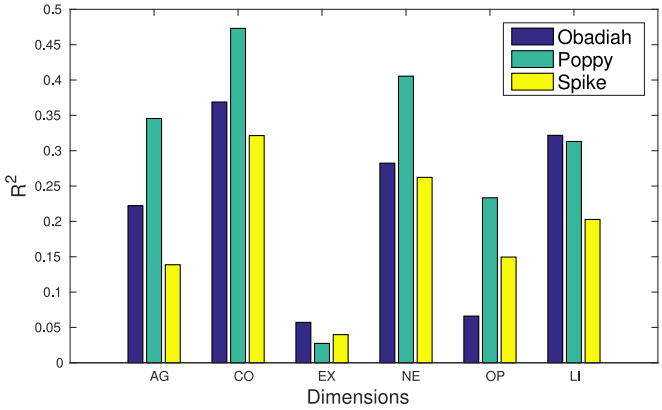

Fig. 4. Multimodal prediction results in terms of R^2 with respect to the three agents (Poppy, Obadiah and Spike). AG: Agreeableness, CO: Conscientiousness, EX: Extroversion, NE: Neuroticism, OP: Openness, LI: Likeability.

played. We collected multiple annotations under varying situational context for each subject separately as well as under different observed modalities. At the next level, we developed a dynamic framework for predicting the impressions. We first extracted a set of visual and audio features to represent each clip and then mapped these features onto the continuous annotations using a time-series regression method (BLSTM). Our experimental results demonstrated that multimodal regression is well capable of modelling the extracted features and the audio-visual labels for all dimensions except for *extroversion* and *facial attractiveness*. We also performed decision-level fusion by combining individual regression outputs obtained from visual and audio features and further improved the prediction of *conscientiousness*, *neuroticism* and *likeability*. We show that varying situational context causes the manifestation of different facets of people's personality. In order to obtain a complete assessment of the observed individual's behaviour and personality, one needs to have in hand multiple displays of the observed subject from audio and visual channels together.

Situational context affects both the raters' perceptions and the automatic predictions. The correlation between the raters' multiple annotations for the same subject in Fig. 3 revealed that the raters' impressions do not change as much as for *conscientiousness*, *openness* and *likeability* as compared to those for *agreeableness*, *extroversion* and *facial attractiveness* from one context to another. Our automatic prediction results also support this phenomenon as, in Table 3, we obtained lower prediction performance for *extroversion* and *facial attractiveness* compared to *conscientiousness*, *openness* and *likeability*. The analysis with respect to the virtual agents (Fig. 4) showed that a better relationship between automatic prediction and situational context was established for *Poppy* and for *Obadiah* with the audio-visual labels. This confirmed that subjects interacting with *Poppy* and *Obadiah* were perceived as more active and expressive, however their cues were more subtle when interacting with *Spike*. Similarly, Batrinca et al. in [30] obtained relatively better personality recognition performance in one of the collaborative contexts. Kalimeri et al. [32] also confirmed the role of social context (i.e., e-mail) in understanding *extroversion* and *conscientiousness*, while for the other dimensions (*agreeableness*, *neuroticism*, *openness*) such an evidence was not found.

In the literature, the only temporal modelling attempt was proposed by Staiano et al. [29]. Their experimental results showed that *extroversion* was better modelled using HMMs rather than a non-sequential technique. Our experimental results instead showed that, in general, time-continuous prediction is better suited for the dimensions of *agreeableness*, *conscientiousness*, *neuroticism*, *openness*, *vocal attractiveness* and *likeability*.

Note that most of the results published in the literature are not directly comparable to one another, as the annotation procedure, the data used and the performance evaluation metrics employed are all different. Although the data partition protocol is different, we compare our prediction results with the baseline results made available as part of the MAPTRAITS 2014 Challenge [40] ($COR_{av} = 0.17$ and $MSE_{av} = 0.59$) as well as the best 6-fold cross validation results on the training set provided by Kaya and Salah in [66] ($COR_{av} = 0.17$ and $MSE_{av} = 0.45$). In our work, we used BLSTM to map the visual and audio features onto the audio-visual labels and combined them at the decision-level. This seems to be a promising approach for time-continuous prediction of observers' impressions as we obtained $COR_{av} = 0.31$ and $MSE_{av} = 0.51$ over all dimensions.

The proposed approach predicted *conscientiousness* ($R^2_{CO} = 0.50$) best using decision-level fusion. The other dimensions predicted with high accuracy were *neuroticism* and *vocal attractiveness* ($R^2_{NE} = 0.32$ and $R^2_{VO} = 0.31$) and also large R^2 values were obtained for *agreeableness*, *openness* and *likeability* ($R^2_{AG} = 0.24$, $R^2_{OP} = 0.25$ and $R^2_{LI} = 0.28$). Biel and Gatica-Perez presented their personality prediction results in terms of R^2 in [27]. Although we calculated R^2 metric differently from [27], i.e., R^2 was calculated per clip and then an average was taken over all of the 30 clips, we compared our results with the results provided in [27] to get an idea on the overall performance of the proposed time-continuous personality prediction approach. In contrast to our results, in [27], fusing visual and audio cues was found to be the best approach in predicting *extroversion* ($R^2_{EX} = 0.41$). However, R^2 values obtained were lower than 0.20 for the rest of the dimensions.

Limitations. Despite the notable contribution of the proposed approach, collecting annotations in a time-continuous manner by employing raters is a very tedious task. We had a limited number of raters due to this reason, which resulted in unbalanced annotations across visual-only, audio-only and audio-visual annotations. Having less raters for audio-only annotation and for audio-visual annotation was due to the fact that the raters reported the annotation with audio to be much easier as compared to the visual-only annotation. In general, they felt more confident about their judgements both in the audio-only annotation and in the audio-visual annotation. However, Table 1 shows that the agreement among audio-only raters was found to be significantly lower. This might be the main reason that the proposed approach performed poorly in learning the relationships between audio features and audio-only labels (see unimodal prediction results in Table 3).

Our experimental data is rich in terms of annotation conditions, the number and type of dimensions, however there is a limited number of subjects and clips due to the

challenges in collecting time-continuous annotations. Therefore there might be a possibility of overfitting, especially, when high-dimensional features were used. Our expectation was to obtain better prediction results with QLZM and HoGF features, however, these features performed slightly poorly as compared to the features that have lower dimensionality.

Although *extroversion* has been frequently reported to be the easiest dimension to recognise or predict, the time-continuous prediction approach cannot model this dimension accurately. *Extroversion* has been found to be dynamic and fluctuating over time by all raters. Fig. 3 also shows that *extroversion*, *facial attractiveness* and *agreeableness* dimensions differ more from one context to another. Taking into consideration the results of unimodal prediction with multimodal labels in Table 3, the worst prediction results were obtained for *extroversion* ($R^2_{EX} = 0.10$) and *facial attractiveness* ($R^2_{FA} = 0.16$), respectively. Due to the limited number of clips, our approach might not be able to capture high-frequency changes over time as well as differences across varying contexts for *extroversion* and *facial attractiveness*.

Future Work. Crowd-sourcing platforms have been widely used to obtain large number of data annotated simultaneously in shorter durations of time. Therefore, as a next step, we plan to increase the number of clips and raters using crowd-sourcing techniques. Another promising research direction would be investigating the impact of different time slices (20 s, 30 s, 45 s etc.) on the prediction tasks, as proposed in [5].

Acknowledgments

This research work was supported by the EPSRC MAPTRAITS Project (Grant Ref: EP/K017500/1) and the EPSRC HARPS Project under its IDEAS Factory Sandpits call on Digital Personhood (Grant Ref: EP/L00416X/1). Much of this work was done while the authors were with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom.

References

[1] D. A. Kenny, "Person: A general model of interpersonal perception," *Personality Social Psychol. Rev.*, vol. 8, pp. 265–280, Aug. 2004.

[2] (2014). Society for Personality and Social Psychology. Even fact will not change first impressions [Online]. Available: http://www.sciencedaily.com/releases/2014/02/140214111207.htm

[3] P. Borkenau, N. Mauer, R. Riemann, F.M. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence," *J. Personality Social Psychol.*, vol. 86, no. 4, pp. 599–614, 2004.

[4] N. Ambady, M. Hallahan, and R. Rosenthal, "On judging and being judged accurately in zero-acquaintance situations," *J. Personality Social Psychol.*, vol. 69, no. 3, pp. 518–529, Sep. 1995.

[5] D. R. Carney, C. R. Colvin, and J. A. Hall, "A thin slice perspective on the accuracy of first impressions," *J. Res. Personality*, vol. 41, no. 5, pp. 1054–1072, 2007.

[6] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 273–291, Jul.–Sep. 2014.

[7] H. Gunes and B. Schuller, *Automatic Analysis of Aesthetics: Human Beauty, Attractiveness and Likability*. Cambridge, U.K.: Cambridge Univ. Press, 2015, pp. 84–93.

[8] W. Youyou, M. Kosinski, and D. Stillwell, "Computer-based personality judgments are more accurate than those made by humans," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 4, pp. 1036–1040, 2015.

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

[10] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *Int. J. Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, 2012.

[11] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Trans. Affective Comput.*, vol. 2, no. 2, pp. 92–105, Apr.-Jun. 2011.

[12] Oya Çeliktutan, Evangelos Sariyanidi, and Hatice Gunes, "Let me tell you about your personality! real-time personality prediction from nonverbal behavioural cues," in *Proc. 11th IEEE Int. Conf. Workshops Autom. Face Gesture Recog.*, 2015, p. 1.

[13] Hatice Gunes, Oya Celiktutan, Evangelos Sariyanidi, and Efstratios Skordos, "Real-time prediction of user personality for social human-robot interactions: Lessons learned from public demonstrations," in *Proc. Int. Workshop Des. Eval. Social Robots Public Settings IROS*, 2015, pp. 1–6.

[14] P. J. Corr and G. Matthews, *The Cambridge Handbook of Personality Psychology*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[15] W. Fleeson, "Towards a structure- and process-integrated view of personality: Traits as density distributions of states," *J. Personality Social Psychol.*, vol. 80, pp. 1011–1027, 2001.

[16] W. Fleeson, "Moving personality beyond the person-situation debate," *Current Directions Psychol. Sci.*, vol. 13, no. 2, pp. 83–87, 2004.

[17] D. A. Kenny, *Interpersonal Perception: A Social Relations Analysis*. New York, NY, USA: Guilford Press, 1994.

[18] P. Borkenau and A. Liebler, "Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence," *J. Personality Social Psychol.*, vol. 65, no. 3, pp. 546–553, 1993.

[19] G. L. Lorenzo, J. C. Biesanz, and L. J. Human, "What is beautiful is good and more accurately understood physical attractiveness and accuracy in first impressions of personality," *Psychol. Sci.*, vol. 21, no. 12, pp. 1777–1782, 2010.

[20] J. Willis and A. Todorov, "First impressions: Making up your mind after a 100-ms exposure to a face," *Psychol. Sci.*, vol. 17, no. 7, pp. 592–598, Aug. 2006.

[21] G. Mohammadi, A. Origlia, M. Filippone, and A. Vinciarelli, "From speech to personality: Mapping voice quality and intonation into personality differences," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 789–792.

[22] J. I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez, "Hi youtube!: Personality impressions and verbal content in social video," in *Proc. of ACM Int. Conf. Multimodal Interaction*, 2013, pp. 119–126.

[23] O. Aran and D. Gatica-Perez, "Cross-domain personality prediction: From blogs to small group meetings," in *Proc. of ACM Int. Conf. Multimodal Interaction*, 2013, pp. 127–130.

[24] M. Rojas, D. Masip, A. Todorov, and J. Vitria, "Automatic prediction of facial trait judgments: Appearance vs. structural models," *PloS One*, vol. 6, no. 8, p. e23323, 2011.

[25] F. Eyben, F. Weninger, E. Marchi, and B. Schuller, "Likability of human voices: A feature analysis and a neural network regression approach to automatic likability estimation," in *Proc. Int. Workshop Image Analysis Multimedia Interactive Serv.*, 2013, pp. 1–4.

[26] J. Joshi, H. Gunes, and R. Göcke, "Automatic prediction of perceived traits using visual cues under varied situational context," in *Proc. Int. Conf. Pattern Recog.*, 2014, pp. 2855–2860.

[27] J. Biel and D. Gatica-Perez, "The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of Vlogs," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 41–55, Jan. 2013.

[28] G. Mohammadi, S. Park, K. Sagae, A. Vinciarelli, and L. P. Morency, "Who is persuasive?: The role of perceived personality and communication modality in social multimedia," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 19–26.

[29] J. Staiano, B. Lepri, R. Subramanian, N. Sebe, and F. Pianesi, "Automatic modeling of personality states in small group interactions," in *Proc. 19th ACM Int. Conf. Multimedia*, 2011, pp. 989–992.

[30] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, "Multimodal recognition of personality traits in human-computer collaborative tasks," in *Proc. 14th ACM Int. Conf. Multimodal Interaction*, 2012, pp. 39–46.

[31] F. Pianesi, "Searching for personality [social sciences]," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 146–158, Jan. 2013.

[32] K. Kalimeri, B. Lepri, and F. Pianesi, "Going beyond traits: Multimodal classification of personality states in the wild," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 27–34.
[33] J. I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: Predicting personality from facial expressions of emotion in online conversational video.," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2012, pp. 53–56.
[34] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Trans. Affective Comput.*, vol. 3, no. 3, pp. 273–284, Jul.–Sep. 2012.
[35] O. Aran and D. Gatica-Perez, "One of a kind: Inferring personality impressions in meetings," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2013, pp. 11–18.
[36] R. Srivastava, J. Feng, S. Roy, S. Yan, and T. Sim, "Don't ask me what I'm like, just watch and listen," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 329–338.
[37] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proc. 10th Int. Conf. Multimodal Interfaces*, 2008, pp. 53–60.
[38] B. P. Motichande, "A graphical user interface for continuous annotation of non-verbal signals," Final Project, BSc FT Computer Science, Queen Mary University of London, London, U.K., 2013.
[39] R. Cowie and G. McKeown. (2010). Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme, *SEMAINE Rep. D6b*. Available: http://www.semaine-project.eu/
[40] O. Çeliktutan, F. Eyben, E. Sariyanidi, H. Gunes, and B. Schuller, "MAPTRAITS 2014: Introduction to the audio/visual mapping personality traits challenge," in *Proc. Int. Conf. Multimodal Interaction*, 2014, pp. 529–530.
[41] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of Personality: Theory and Research*, vol. 2. New York, NY, USA: Elsevier, 1999, pp. 102–138.
[42] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. Int. Workshop Emotion Representation, Anal. Synthesis Continuous Time Space*, 2013, pp. 1–8.
[43] (2016, Jan.). Wikipedia. Halo effect [Online]. Available: http://en.wikipedia.org/wiki/Halo_effect
[44] H. C. Akakin and B. Sankur, "Robust classification of face and head gestures in video," *Image Vis. Comput.*, vol. 29, no. 7, pp. 470–483, Jun. 2011.
[45] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.
[46] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 532–539.
[47] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro, "Local Zernike moment representations for facial affect recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 108.1–108.13.
[48] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images," *Int. J. Comput. Vis.*, vol. 99, no. 2, pp. 190–214, Sep. 2012.
[49] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[50] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image Video Retr.*, 2007, pp. 401–408.
[51] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.
[52] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[53] (2015, Oct.). Praat: doing phonetics by computer [Online]. Available: http://www.fon.hum.uva.nl/praat/
[54] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, vol. 385. New York, NY, USA: Springer, 2012.
[55] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5-6, pp. 602–610, 2005.
[56] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affective Comput.*, vol. 3, no. 2, pp. 184–198, Apr.-Jun. 2012.
[57] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[58] A. F. Gers, N. N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *J. Mach. Learn. Res.*, vol. 3, pp. 115–143, Mar. 2003.
[59] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
[60] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *Proc. A Field Guide to Dynamical Recurrent Neural Networks*, Kremer and Kolen, Eds. Piscataway, NJ, USA: IEEE Press, 2001.
[61] A. Graves. (2013, Aug.). Rnnlib: A recurrent neural network library for sequence learning problems [Online]. Available: http://sourceforge.net/projects/rnnl/
[62] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Connecting meeting behavior with extraversion: A systematic study," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 443–455, Oct.-Dec. 2012.
[63] H. Gunes and M. Piccardi, "Assessing facial beauty through proportion analysis by image processing and supervised learning," *Int. J. Human-Comput. Studies*, vol. 64, no. 12, pp. 1184–1199, 2006.
[64] H. Gunes, "A survey of perception and computation of human beauty," in *Proc. ACM Multimedia Int. Workshop Social Signal Process.*, Dec. 2011, pp. 19–24.
[65] S. Kalayci, H. K. Ekenel, and H. Gunes, "Automatic analysis of facial attractiveness from video," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 4191–4195.
[66] H. Kaya and A. A. Salah, "Continuous mapping of personality traits," in *Proc. Int. Conf. Multimodal Interaction, 1st Audio/Visual Mapping Personality Traits Challenge*, 2014, pp. 17–24.

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

Hatice Gunes is a University senior lecturer (associate professor) at the Computer Laboratory, University of Cambridge, United Kingdom. Her research interests are in the areas of affective computing, social signal processing, interactive computer vision, and applied machine learning. She has published more than 80 technical papers in these areas and was a recipient of a number of awards for the Outstanding Paper (IEEE FG'11), Quality Reviewer (IEEE ICME'11), Best Demo (IEEE ACII'09), and Best Student Paper (VisHCI06). She serves on the Management Board of the Association for the Advancement of Affective Computing (AAAC), and the Steering Committee and the Editorial Board of the *IEEE Transactions on Affective Computing*.