Multimodal Self-Assessed Personality Estimation during Crowded Mingle Scenarios Using Wearables Devices and Cameras

Cabrera-Quiros, Laura; Gedik, Ekin; Hung, Hayley

DOI
10.1109/TAFFC.2019.2930605

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Citation (APA)
Cabrera-Quiros, L., Gedik, E., & Hung, H. (2022). Multimodal Self-Assessed Personality Estimation during Crowded Mingle Scenarios Using Wearables Devices and Cameras. *IEEE Transactions on Affective Computing*, *13*(1), 46-59. Article 8769877. https://doi.org/10.1109/TAFFC.2019.2930605

Multimodal Self-Assessed Personality Estimation During Crowded Mingle Scenarios Using Wearables Devices and Cameras

Laura Cabrera-Quiros, Ekin Gedik, and Hayley Hung, *Member, IEEE*

Abstract—This paper focuses on the automatic classification of self-assessed personality traits from the HEXACO inventory during crowded mingle scenarios. These scenarios provide rich study cases for social behavior analysis but are also challenging to analyze automatically as people in them interact dynamically and freely in an *in-the-wild* face-to-face setting. To do so, we leverage the use of wearable sensors recording acceleration and proximity, and video from overhead cameras. We use 3 different behavioral modality types (movement, speech and proximity) coming from 2 sensors (wearable and camera). Unlike other works, we extract an individual's speaking status from a single body worn triaxial accelerometer instead of audio, which scales easily to large populations. Additionally, we study the effect of different combinations of modality types on the personality estimation, and how this relates to the nature of each trait. We also include an analysis of feature complementarity and an evaluation of feature importance for the classification, showing that combining complementary modality types further improves the classification performance. We estimate the self-assessed personality traits both using a binary classification (community's standard) and as a regression over the trait scores. Finally, we analyze the impact of the accuracy of the speech detection on the overall performance of the personality estimation.

Index Terms—Personality, Wearable acceleration, proximity, video, speaking turn, HEXACO

1 Introduction

The automatic detection and recognition of displayed personality traits, either perceived by oneself or by others, has received considerable interest in the affective computing community for the past 20 years, among other fields. Such interest generates as part of the endeavors to either 1) adapt the interaction of a system or virtual agent to each specific person's needs, or 2) use automatic systems to analyze social human behavior.

Different modalities have been used to analyze and estimate personality traits, with audio-visual approaches being predominant among the works [38]. In addition, the estimation of such traits has been addressed in different types of scenarios including meetings ([15], [34]), video logs (VLOGS) or self-presentations ([5], [6], [35]), radio broadcasts ([31], [32]) or social media ([12]), among other situations.

Nonetheless, most of the aforementioned efforts tend to share the same characteristics: 1) data of a single person can be easily differentiated from the rest, and 2) they do not have much missing data. For example, works using VLOGS (such as the Chalearn challenge [35]) have a clear view of the participant's faces and unique speech. In contrast, other scenarios do not allow the acquisition of clear, personalized and high quality data without specialized equipment.

One of such scenarios are mingle events, such as parties or networking events, where people are inherently encouraged to interact. These are intriguing scenarios from the social signal analysis perspective [39] due to their dynamic nature and also comprise a wide range of social interactions and the formation of free-standing conversational groups which also triggers research in group dynamics [1], [27].

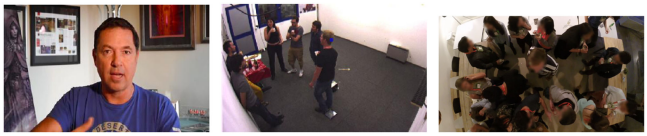
In this paper, we focus in the estimation of *self-assessed* personality traits from the HEXACO inventory [2] during crowded mingling events using wearable sensors and video cameras in a noninvasive manner.

Compared to other scenarios where the estimation of personality has been addressed, crowded mingle events are harder to analyze using audio-visual modalities. For example, during meetings or VLOGS settings the audio and frontal video for each participant is generally recorded separately, as can be seen in Fig. 1a. Hence, for these scenarios the camera has a clear view of a single participants' face and its speech is unique or can be robustly separated, providing rather clean data from these 2 modalities.

In contrast, mingle scenarios are crowded events where obtaining clean data from computer vision techniques is hard due to occlusion problems, changing light conditions and challenges with people re-identification. In addition, mingle scenarios present ambient noise due to the event itself that makes harder to record good quality audio for each person without customized equipment (eg. microphones).

- *L. Cabrera-Quiros is with the Department of Intelligent Systems at TU Delft, Delft 2628, CD, The Netherlands, and also with the Escuela de Ingeniería Electrónica at the Instituto Tecnolgico de Costa Rica, Cartago 30101, Costa Rica. E-mail: l.c.cabreraquiros@tudelft.nl.*
- *E. Gedik and H. Hung are with the Department of Intelligent Systems at TU Delft, Delft 2628 CD, The Netherlands. E-mail: {e.gedik, h.hung}@tudelft.nl.*

Manuscript received 8 Mar. 2018; revised 26 June 2019; accepted 18 July 2019. Date of publication 23 July 2019; date of current version 1 Mar. 2022. (Corresponding author: Laura Cabrera-Quiros.)
Recommended for acceptance by S. Scherer.
Digital Object Identifier no. 10.1109/TAFFC.2019.2930605



(a) (b) (c)

Fig. 1. Example snapshots of typical scenarios for personality estimation. (a) VLOG taken from the Chalearn Challenge 2016 [35], (b) mingle event taken from [40], (c) a more crowded mingle event (our data).

In this work we focus on the estimation of personality traits during mingle scenarios, leveraging wearable devices, sensing acceleration and proximity, and video cameras recording the event from above (see Fig. 1c as an example). Using these types of sensors also allows our method to be unobtrusive and to scale rather easily to a higher number of people. Furthermore, we focus on a crowded scenario, including up to 56 people freely interacting for 30 minutes.

Our main contributions are: 1) we leverage the use of wearable devices and overhead video cameras to estimate self-assessed personality traits during a crowded mingle event, 2) we estimate the self-assesed personality traits both using a binary classification (standard in the computing community) and as a regression over the trait level, 3) we compare the impact of different modality types on the estimation of the different personality traits as we hypothesize that each modality captures the event differently, 4) we study the impact of fusing different modality types in the estimation performance of each trait and, 5) we analyze the impact of the speaker detection in the overall performance of personality estimation.

This paper is an extension of our conference paper presented in the International Conference of Multimodal Interaction (ICMI) [8]. There, we introduce the use of speaking status detection from wearable acceleration (proposed in [16]) to create a third behavioral modality, and the use of global features from movement, speech and proximity. Here, we modify our method to add the use of the video modality, and focus on research questions related to comparisons against and the complementarity of this additional modality to those from the wearable (eg. acceleration, proximity and speaking status). In addition, we analyze the correlation between feature types, and the impact of a speech detection stage and the visibility of the participants in the cameras (missing data) on the classification performance.

The rest of the paper is structured as follows. Section 2 gives an overview of related work. In Section 3 the mingle data collected and used for our experiments is explained in detail. Section 4 describes our method while in Section 5 are summarize our methodology and results. We discuss our findings in Section 6. Finally, we conclude our work in Section 7.

2 Related Work

As the amount of works on personality analysis and estimation is extensive, we only focus on works estimating *self-assessed* personality, meaning that the participants filled an inventory/survey to score in their own personality traits.

Nonetheles, many efforts have been made in automated third-party attribution-based personality recognition (or personality impression) [6], [38], or focused on personality estimation in social media [12], [37], which are beyond the scope of this paper. A comprehensive review of the related personality computing literature can be found in [38]. Also, specific workshops/challenges (also using impressions) such as the MAPTRAITS [11] or the Chalearn Looking at People Challenge [35] have encouraged researchers to automatically analyze personality.

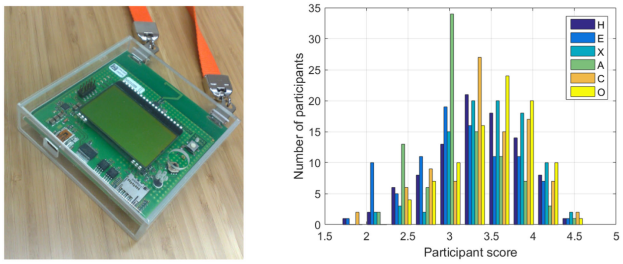
Within the domain of automated self-assessed personality estimation during face-to-face interactions, works can be grouped mainly in small meetings and mingle scenarios. The meeting setting generally involves a fixed number of people interacting, normally sitting around a table. In constrast, the mingle scenarios involve 4 or more people (56 in our case) freely interacting in standing groups. The conversational groups for this settings can form, merge and split, following the desire of the participants.

For the meeting setup, Pianesi et al. [34] proposed a method to recognize Extraversion and the Locus of Control during multi-party meetings of 4 people. The setting in this study has a pre-defined task and a controlled environment, where cameras and microphones were recording every participant individually. This work was extended by Lepri et al. [24]. Both works used the corpus which was first introduced by Mana et al. [26]. Another work on extraversion estimation was presented by Lepri et al. [25]. In this work, the authors estimate this trait (from the Big Five inventory) using meeting behavior such as speaking time and attention given/received during a meeting. They show that these two behaviors are a suitable indicator to detect extraverts during a meeting setting.

Batrinca et al. [5] presented a method to analyze interview style self-presentations performed during a video Skype call, which simulated an interview, to recognize all traits in the Big Five. Although they collected data for 89 people, they only interact with the interviewer for part of the call while the main segment for non-verbal cue extraction was a monologue.

Few works have addressed the free mingle setup, due mainly to its challenges regarding missing data, dynamic groups and visual obstacles (e.g., occlusions, changes in light conditions or appearances). For instance, Alameda-Pineda et al. [1] presented the Salsa dataset which consists of a mingle event and poster session with recordings from video cameras and proximity from IR from 18 people, making it similar to our dataset (see Section 3). Nonetheless, although they included the personality traits from the Big Five inventory for all their participants, they did not provide automatic estimation of these traits.

The closest work to our own was presented by Zen et al. [40]. They proposed a classification method to recognize Extraversion and Neuroticism (from the Big Five) using proximity related features extracted from multiple cameras in a considerably less crowded mingle event than ours (see Fig. 1b). These features were motivated by findings from social psychology about the relationship between proxemics and the 2 personality traits in question. Compared to this work, with a total of 7 participants, we present a significant increase with experiments evaluated on 56 people. Also, their proximity features are based on distances while ours rely on binary neighbor detection (see Section 3).



(a) (b)

Fig. 2. (a) Custom-made wearable devices, which record binary proximity and triaxial acceleration, (b) Proportion of participants with a similar score for the 6 different HEXACO traits.

To the best of our knowledge, we are the first to address the complexity of crowded mingle scenarios using a fusion approach of wearable devices and video cameras.

In contrast to our work, which tries to exploit the global behavior of people during an event to estimate their traits, other works have addressed the estimation of personality states[17] instead of personality traits. Thus, personality is treated as specific behavioral episodes that can change over time. The work by Kalimeri et al. [21] was the first to address this new approach to personality estimation. They did so in an office setting, following up to 54 people during 6 weeks. They used the Sociometric badge which incorporates accelerometer, IR, Bluetooth and audio, but they only use the speech and proximity modalities for their experiments.

3 Data

3.1 The MatchNMingle Dataset

This dataset,[1] which is publicly available [7], was collected during 3 separate social evenings in a public bar-restaurant. The participants were mostly students who signed up for a speed date event, each followed by a mingle session. For this study we only used the mingle section of MatchNMingle.

During each event, between 30 and 32 different people participated, with a total of 92 participants for the 3 events. From these, only 56 are used in this work (see Section 3.2.1 for a detailed explanation). Most participants were students between 18 and 30 years old (mean = 22.09, std = 2.34), who were recruited from a university campus. As this was part of a heterosexual speed date event, the number of participants per gender was balanced.

While the dataset provides up to 50 minutes of free mingling, a 30 minutes segment was selected where the number of people interacting was maximized. During this time, participants interacted freely in a space for that purpose (see Fig. 1c). They were allowed to leave the mingle area at will (eg. go to the bathroom), and request drinks/snacks.

3.2 Modalities

Wearable Devices. All participants were asked to use through the entire night a custom-made wearable device hung around their neck, like an ID badge (see Fig. 2a). This wearing method makes it perfect to replicate for other use-cases such as conferences, exhibitions, or business events. The wearable devices recorded triaxial acceleration at 20 Hz. Also, each device communicated with other devices using a radio-based beacon communication by emitting its own ID to all others. Thus, close devices in a 2-3 meter radius will detect each other as neighbors. These detections are considered as a binary proximity which updates every second. The communication also allows all devices to synchronize to a global timestamp. See [13] and [7] for more details.

Due to hardware malfunction, only 70 of the 92 devices recorded during the mingle segment.[2] From the functioning 70 devices, we eliminated 3 other devices due to incomplete data, leaving us with a total of 67 devices recording wearable acceleration and proximity. Although these missing devices could potentially affect the social interactions on terms of proximity, the quantitatively impact of it lies outside of the scope of this paper. Thus, we leave this for future work.

Finally, from these devices only *56 subjects* had both acceleration and video data available (see Section 3.2.1 for more).

Video. The mingle session was recorded by 5 GoPro Hero +3 cameras from above at 20 FPS, synchronized to the wearable devices using a global time. In addition, the MatchNMingle dataset provides full annotations of position and social actions (eg. speaking, gestures) for 2 of these 5 cameras, due to financial limitations (version V.1 of the dataset). The area for the mingle session was limited in space to $1m^2$ per person to ensure crowdedness. A snapshot of the event, for the mingle part, can be seen in Fig. 1c where we contrast the density of our event with that used by Zen et al. [40] (Fig. 1b).

3.2.1 Camera Low Visibility Subset

As stated before, only 2 of the 5 cameras have annotations (including positions) for the entire 30 minutes. Hence, those participants outside the field of view of the camera are treated as *not visible*. Also, due to the dynamic nature of the event itself, some participants are not captured by any of the cameras at some points (eg. going to the bathroom).

We analyzed the video data to extract a subset of participants that allows a fair comparison between modalities (eg. participants with video and wearable data). For this subset, all participants should be under the FoV of one of the cameras for at least half of the time (15 minutes). This time is not necessarily continuous. Hence, we ensured that there is a representative amount of data for each participants' video, even with missing data.

Thus, we are left with a final subset of *56 participants* that have both a working device and are visible at least 50 percent of the time.

3.3 Personality Questionnaires

Prior to the event, each participant filled in the HEXACO personality inventory [2], for which six dimensions are extracted: Honesty(H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O), by means of the HEXACO-PI-R survey [22]. In addition, each scale in HEXACO can be further separated into facet-level scales (e.g., Social Self-Esteem, Social Boldness, Sociability and Liveliness are part of the extraversion).

1. We used the version 1 of the dataset.

2. Both wearable acceleration and proximity were missing for these devices. These also did not send proximity information.

TABLE 1
Summary of Our Features Divided by Modality Type: W = Mov. from Wearable, S = Speaking, WS = Mov. from Wearable While Speaking, P = Proximity and V = Mov. from Video

	Feature	Type	Sensor
1	mean of accel. magnitude var. per window	W	Wearable
2	var. of accel. magnitude var. per window		
3	maximum length of S.T.	S	
4	mean length S.T.		
5	variance of length for S.T.		
6	maximum length of non-S.T.		
7	mean length non-S.T.		
8	variance of length for non-S.T.		
9	total lenght of S.T.		
10	mean of accel. magnitude var. per window for S.T.	WS	
11	var. of accel. magnitude var. per window for S.T.		
12	mean size of group interacted with	P	
13	largest size of group interacted with		
14	total number of people interacted with		
15	mean of total number of zeros of OF magnitude	V	Video
16	mean of mean OF magnitude var. per window		
17	var. of mean OF magnitude var. per window		
18	mean of mean OF magnitude from distribution (low)		
19	mean of mean OF magnitude from distribution (medium)		
20	mean of mean OF magnitude from distribution (high)		

(S.T. = Speaking turns.)

This survey consists of 100 questions[3] which are answered on a scale from 1 (strongly disagree) to 5 (strongly agree).

We chose the HEXACO rather than the more frequently used 5 factor models such as the Big-5 or the Five Factor Model (FFM). While the Big-5 and HEXACO are both derived from the same lexical studies (see [4] for review), the six-dimensional HEXACO model has been shown to more optimally capture the data in cross-cultural replications [2], and to outperform the Big-5 in both self-ratings (i.e., when participants complete the inventories about themselves) and in observer ratings (i.e., when participants complete the scale about another individual) [3].

Nevertheless, the HEXACO and five factor models are related in a number of ways: 1) extraversion and conscientiousness are the most similar among all the dimensions to their five factor counterparts, 2) agreeableness and emotionality in the HEXACO are rotated versions of their five factor counterparts, with traits related to anger loading on HEXACO Agreeableness instead of Big-5 Neuroticism, and traits relating to sentimentality loading on HEXACO Emotionality instead of Big-5 Agreeableness, and 3) terms such as honest, sincere, fair etc. that load on Big-5 Agreeableness are the separate dimension of HEXACO Honesty-Humility instead (see [4] for a review).

The distribution of trait scores over the participants for the 6 traits is presented in Fig. 2b. Also, the Cronbach's α coefficients were 0.81 for Honesty, 0.87 for Emotionality, 0.84 for Extraversion, 0.82 for Agreeableness, 0.83 for Conscientiousness and 0.77 for Openness to experience.[4]

3.4 Manual Annotations for Speaking Status

MatchNMingle also provides manual annotations for the position of the participants in the video, and 8 different social actions in the mingle session. These were annotated from video by multiple trained annotators [7].

3. http://hexaco.org/

4. The Cronbach's α coefficient is widely used to test the internal reliability of scales. By convention, 0.65 is considered sufficient, and 0.8 is considered good in terms of reliability.

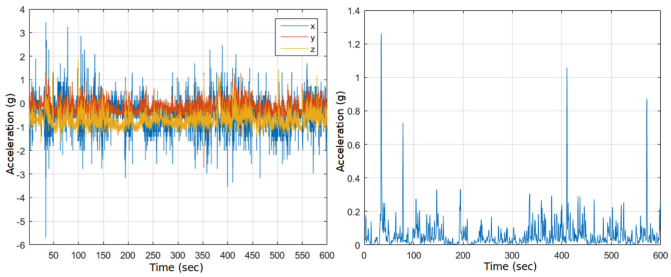


Fig. 3. (a) Original raw acceleration of a wearable device (after filtering effects of the gravity). (b) Body movement energy resulting from preprocessing (described in Section 4.1.1).

Manual annotators 1) manually track all the people in the video and 2) annotate the 8 selected social actions for each of them. More than one action could be selected in parallel. The 30 minutes corresponding to the mingle were divided into segments of 2 minutes and annotated by 7 different coders. For the speaking status, the mean inter-annotator agreement between subjects using Fleiss-Kappa coefficient was 0.55, which corresponds to a moderate agreement.

4 Non-Verbal Cues

First, a summary of all our non-verbal cues is shown in Table 1, separated by the modality type and the sensor they come from (wearable or camera). Thus, from the 3 digital modalities at our disposal (wearable acceleration, proximity and video) we extracted 5 behavioral modality types: 1) Speech (S), 2) Movement from wearable (W), 3) Movement from wearable while Speaking (WS), 4) Proximity (P) and 5) Movement from video (V).

In the next subsections, a detailed description of the preprocessing on each sensor and the extraction of each global feature is presented.

4.1 Wearable Devices

For the wearable devices we grouped our cues, which originated from 2 different sensors or *digital* modalities (triaxial acceleration and proximity), in 3 *behavioral* modality categories and their combination: body movement energy (W), speaking turns (S), body movement during speaking turns (WS) and proximity (P). Each behavioral modality, is detailed below.

4.1.1 Body Movement Energy (W)

For each wearable device, a single acceleration magnitude from the 3 axes is computed. Next, we apply a sliding window calculating the variance over the magnitude of the acceleration, using a 3s window with a 2s shift. Previous works have shown this window size and shift to be the best for stream association and analysis using wearables [8], [9], [28]. A graphical representation of this process is presented in Fig. 3. This give us a better representation of *movement energy* over time than the raw acceleration magnitude, as can be seen in Fig. 3b.

To obtain a single value for the 30 minute segment, we calculate 2 features to represent the movement energy: the mean and variance of the energy values over all windows. These features are 1 and 2 in Table 1.

4.1.2 Speaking Turns (S)

Building on prior findings that people's speaking status is representative of their personality [5], [34], [38], we extracted them from each individual's accelerometer signal. The use of this non-traditional modality to detect speech is motivated by the well-studied relationship between bodily gestures and speaking [29]. To do so, we have used Transductive Parameter Transfer (TPT) [41]. In previous work, this method has shown experimentally to perform significantly better than a traditional machine learning approach [16]. We hypothesize that TPT is much better in capturing the person specific nature of the connection between body movements and speech. Speaking turns are then used to extract high-level features representing the interaction characteristics of a participant.

Transductive Parameter Transfer (TPT). For a feature space X and label space Y, N source datasets with label information $D_i^s = \left\{x_j^s, y_j^s\right\}_{j=1}^{n_i^s}$ and an unlabeled target dataset $X^t = \{x_j^t\}_{j=1}^{n_t}$ are defined. It is assumed that samples $X_i^s = \{x_j^s\}_{j=1}^{n_s}$ and X^t are generated by marginal distributions P_i^s and P^t, where $P^t \neq P_i^s$ and $P_i^s \neq P_j^s$. In the notation used, s always corresponds to source datasets while t corresponds to the target one. This approach aims to find the parameters of the classifier for the target dataset X^t by learning a mapping between the marginal distribution of the datasets and the parameter vectors of the classifier in the three following steps:

(1) *Train source specific classifiers on each source set* $\mathbf{D_i^s}$*:* Instead of using a Linear SVM as in [41], we have selected a L2 penalized logistic regressor as our classifier which is experimentally shown to perform better with our data. Chosen classifier minimizes

$$\min_{(w,c)} \frac{1}{2} w^T w + C \sum_{i=1}^{n} log(exp(-y_i(X_i^T w + c)) + 1). \quad (1)$$

Thus, for every source dataset D_i^s, parameters $\theta_i = (w, c)_i$ are computed.

(2) *Learn the relation between the marginal distributions* $\mathbf{P_i^s}$ *and the parameter vectors* $\theta_\mathbf{i}$ *using a regression algorithm:* Training set $T = \{X_i^s, \theta_i\}_{i=1}^{N}$ is formed by samples X_i^s and parameters θ_i obtained from each source dataset. A mapping $\hat{f} : 2^x \rightarrow \theta$, which takes a set of samples and returns the parameter vector θ needs to be learned. Assuming that elements in θ may be correlated, we have employed Kernel Ridge Regression [33], instead of the independent Support Vector Regressors used in [41]. Since we need to define the similarities between distributions X_i^s instead of independent samples, we employ an Earth Mover's Distance [36] based kernel. EMD kernel is computed as:

$$\kappa_{EMD} = e^{-\gamma EMD(X_i, X_j)}. \quad (2)$$

In Eq. (2), $EMD(X_i, X_j)$ corresponds to the minimum cost needed to transform X_i into X_j. The user defined parameter γ is set to be the average distance between all pairs of datasets.

(3) *Use* $\hat{f}$ *to obtain the classifier parameters on the target distribution:* After computing $\hat{f}(.)$, we directly apply this mapping to target data X^t to obtain θ^t. With θ^t known, we can infer the labels for the target dataset.

Extracting Speaking Turns. For detecting speaking turns with TPT, we selected simple statistical (mean and variance) and spectral features (power spectral density, using 8 bins with logarithmic spacing from 0-8 Hz as presented in [18]) that are expected to be representative of speech related body movements. These features were extracted from each axis of the raw acceleration, the absolute values from each axis of the acceleration, and magnitude of the acceleration using 3s windows with a 2s shift. Using the labeled data of 18 participants as sources, we obtained speaking turns for all 56 participants during 30 minutes. The time interval used for these 18 participants is not the same as the 30 minutes used for our experiments. As stated in Section 3, the labels for the speaking status of these 18 participants (sources) are obtained by manual annotation using the video.

Finally, derived features were extracted from the speaking turns (see Table 1). We create 7 global features from the speaking turns, which have the reference numbers 3 to 9 in Table 1. These features are the maximum, mean, variance and total of length of speaking turns, and the same for non speaking turns. In addition, we create 2 additional multimodal behavioral features (WS), which combines the movement and the speaking turns (reference numbers 10 and 11 in Table 1). These are the mean and variance of the movement energy only in those windows with detected speaking.

4.1.3 Proximity (P)

As stated before, each wearable device has a binary proximity detector based on beacon communication with other devices. So, each device emits its own ID to all other devices and a detection of a particular ID is treated as a neighbor. From these binary detections, a dynamic (in time) binary proximity graph can be generated for each participant. To eliminate false neighbor detections, the method proposed by Martella et al. [27] was applied.

Then, 3 features (ref. numbers 12, 13 and 14 in Table 1) were calculated for each participant from the proximity graphs: mean size and largest size of group participated in, and the total number of people interacted with during the event. Since we do not have actual distances, these features allow us to represent statistics related to the number of people's interactions during the event. To consider stable interactions in our proximity features, 2 nodes are only accounted as neighbors if they detect each other for more than one minute in the graphs.

4.2 Video Cameras

4.2.1 Movement from Camera (V)

First, we extract the dense optical flow of the entire video frame using the Farneback's algorithm [14]. Then, we obtain the position of each participant in each frame using a bounding box and extract the magnitudes of the flow vectors within this box, as seen in Figs. 4b and 4c. Then, a single movement value per participant per frame is calculated using the mean value of the magnitudes within the bounding box. This is done for all frames in the video, which are later concatenated. Hence, we can represent the movement of the participant in the video with a single time series (Fig. 4d). Notice that we use the magnitude of the flow

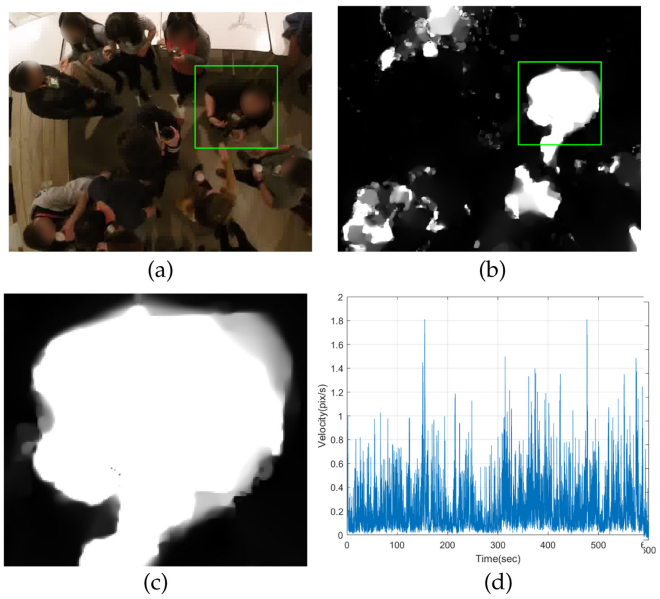


Fig. 4. Extraction of speed time series from video. (a) Participant's location in video. (b) Magnitude of dense optical flow for an entire frame. (c) Magnitude of dense optical flow within a person's bounding box. (d) Speed time series for one participant extracted from the mean magnitude of its optical flow.

vectors instead of the Cartesian values, as the participants always have a relative frame of reference (eg. their orientation changes with respect to the camera).

Next, we apply a sliding window calculating the variance over the magnitude of the acceleration in video, similarly as we do for the time-series from the wearable. The size od the windows are also 3 seconds with a 2 second shift.

In addition to the above, we also extract 3 additional time series per participant for different levels of movement intensity. To do so, we separate the flow magnitudes of each frame within each person's bounding box in 3 bins (low, medium and high), using the third percentile. Then, we calculate the mean movement for each frame for these 3 separate bins. Thus, we obtain 3 additional time series per participant, as seen in Fig. 5. We do this to further analyze the impact of the type of event (high versus low acceleration variation) on the detection of personality.

Finally, to obtain our global features we calculate the mean (in time) total number of zeros in the optical flow vectors, and the mean variance of all time series (entire and 3 separated by bins). We also include the variance for the entire time series. This give us a total of 6 global features for the video modality (V).

Compensating Video Complexity. As can be seen in Fig. 6a, sometimes the bounding box with the participant's location captures movement that does not corresponds to the participant itself. Hence, instead of using the raw optical flow, we apply a multivariate Gaussian function centered at the bounding box location as a weighting factor to compensate for the extracted flow vectors that possibly do not belong to the participant (eg. borders):

$$f(X,\mu,\Sigma) = \frac{1}{\sqrt{|\Sigma|(2\pi)^2}} e^{(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}. \quad (3)$$

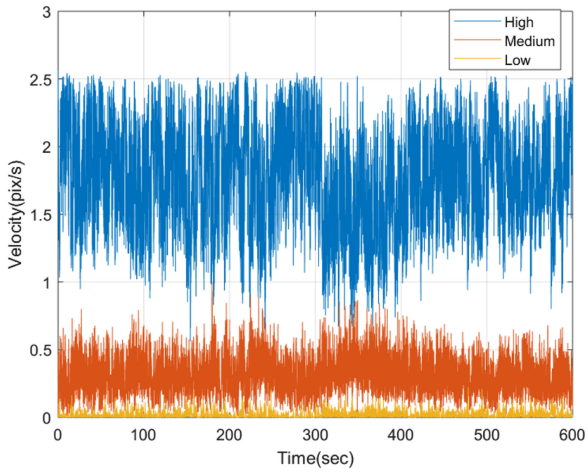


Fig. 5. Binned velocity signals from video for one participant.

Where $\mu = [\mu_x, \mu_y]$, μ_x and μ_y are the center of the bounding box, and Σ is the covariance matrix.

The aim of Eq. (3) is to adapt to the form of the person given its position in the image plane, and give it a higher weight to the flow vectors located in the center of the bounding box where, we hypothesize, the person is truly located. Thus, μ_x and μ_y control the position of the box and the covariance matrix Σ its inclination. More specifically, given:

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \quad (4)$$

We define Σ_{XX} as a quadratic function of the position for the bounding box with respect to the image plane, or $f(x) = a * \mu_x^2 + b * \mu_x + c$. The same applies for Σ_{YY}, using μ_y. Finally, Σ_{XY} and Σ_{YX} are define as a function of 2 variables given by $f(x,y) = d * (\mu_y - w/2) * -(\mu_y - h/2)$, were w and h represent the width and height of the image. Here, a, b, c and d are constants that depend of the resolution of the image.

Fig. 6c shows how the form of the Gaussian given by Eq. (3) changes depending on the position in the image plane of the participant with respect to that of the camera. Thus, if the person is directly under the camera Eq. (3) produces a more symmetric distribution, whereas Fig. 6b shows the distribution required for the weighting of a person located in the top right position of the image plane.

4.3 Motivation for Our Features

Previous work ([10], [18], [19], [20]) has shown that the movement of the people while they interact is a good indicative of their levels of arousal or emotional state. These haven then been used for estimating different components in the interactions such as cohesion or dominance. Our movement features (1-2, 10-11, and 15-20 in Table 1) are based on this premise. Thus, the main hypothesis is that the amount, variance and intensity in which a person moves during a conversation will be indicative of their personality.

Similarly, several works in both automatic computing [21], [25], [34] and social science [30] have addressed the predictive power of speaking status for most personality traits, with an emphasis on extraversion and neuroticism along the dimensions of the Big Five. In our case, we rely on

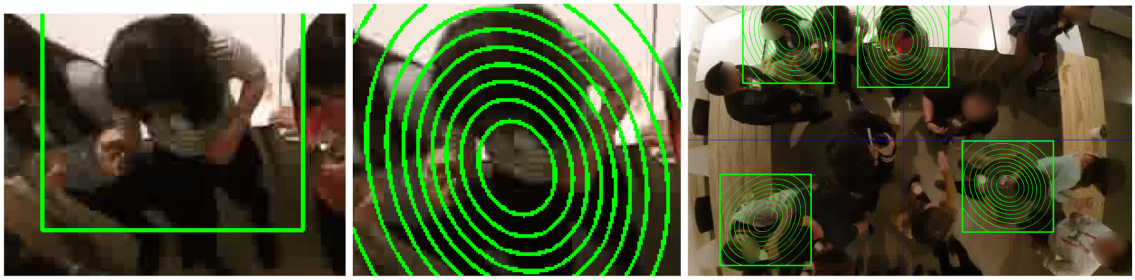


Fig. 6. Correction of outside flow vectors. (a) Example of third-party movement capture by the bounding box. (b) Weight correction using a multivariate Gaussian function. (c) Examples of multivariate Gaussian functions given the position w.r.t the camera

a method for the detection of speaking turns from wearable devices [8], [16], [18].

Regarding our proximity features, we gathered past features used in similar setups such as that presented by Zen et al. [40]. Nonetheless, as these works generally have a nominal distance between people, we had to modify the proximity features to use binary proximity information.

5 Experiments

As summarized in Table 1, we divided our set of features in 5 behavioral modality types: 1) Speech (S), 2) Movement from wearable (W), 3) Movement from wearable while Speaking (WS), 4) Proximity (P) and 5) Movement from video (V). In the next subsections we compare and analyze the complementarity of these feature types, both with a correlation analysis and during classification.

In addition, we analyze the impact of the speech detection in the overall performance of the personality estimation by comparing it to the speech annotations provided by the MatchNMingle dataset.

5.1 Feature Correlation Analysis

Fig. 7 shows the correlations for our final 20 features (summarized in Table 1). First, we can see 5 clusters in this figure that correspond to each modality type: 1-2 for W, 3-9 for S, 10-11 for WS, 12-14 for P and 15-20 for V.

Some of the correlations results are as expected. For example, we can see how the features related to speaking turns from the set S (3 to 5) are inversely correlated to those related to non-speaking turns (6 to 8).

Nonetheless, there are some interesting results. The feature for low distribution values of OF magnitude (18) does not correlate strongly with any of the other features, not even those in the same modality set V. This might be due to remaining noise in the video (after our filtering described in Section 4.2) most likely captured by this feature.

Another interesting result in this figure are the correlations between the features of mean and variance movement from the wearable W (1 and 2) and the video V (16 and 17). The absolute values for these correlations are low (around 0-0.2). An explanation for such low values can be that, as we hypothesized, each modality might be providing complementary information about the person's personality.

5.2 Comparison of Behavioral Modality Types

Once we have seen the correlation between all the features, we proceed to analyze the impact of each modality type separately. To do so, we trained 5 different binary classifiers using only those features for the given set (W, S, WS, P or V). We used a L1 penalized logistic regressor (to reduce possible overfitting) and applied a 10-fold cross-validation. Note that, as we only have one sample per subject (i.e., a 20-dimensional vector), using a 10-fold crossvalidation is valid in our case without contamination between the train and test set. To create binary labels from our trait scores, we used the median value for each trait as threshold with the higher values in the positive class. We do this for each fold, so the median is

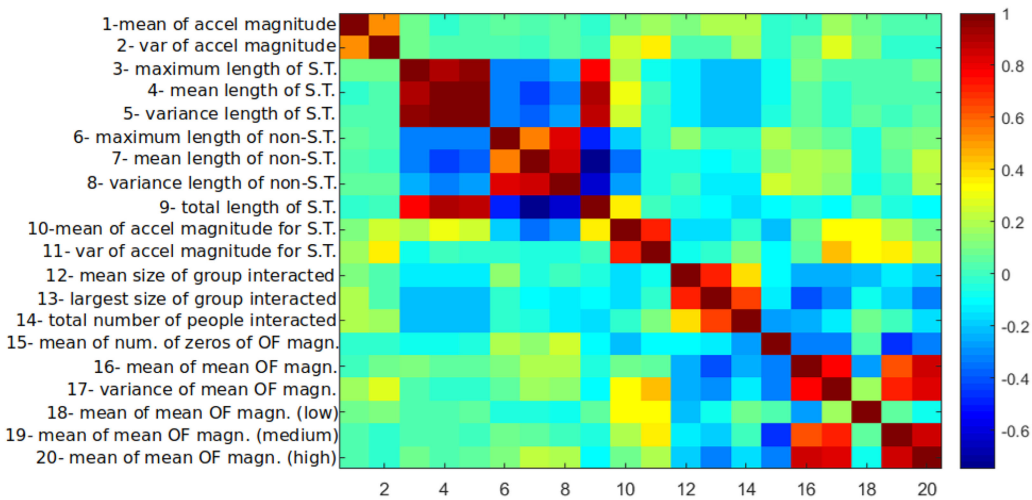


Fig. 7. Correlation between all features in Table 1 (better seen in color).

TABLE 2
Complementarity of Behavioral Modality Types from All Sources (Wearable + Video))

Modality set Combination	Performance per trait					
	H	E	X	A	C	O
W	0.47 ± 0.10	0.46 ± 0.05	0.54 ± 0.13	0.35 ± 0.21	0.45 ± 0.11	0.58 ± 0.11
S	0.59 ± 0.23	0.37 ± 0.19	0.36 ± 0.13	0.50 ± 0.10	0.58 ± 0.12*	0.54 ± 0.17
WS	0.64 ± 0.20*	0.43 ± 0.10	0.53 ± 0.13*	0.46 ± 0.12	0.60 ± 0.17*	0.44 ± 0.05
P	0.48 ± 0.12	0.39 ± 0.16	0.55 ± 0.22*	0.63 ± 0.25	0.48 ± 0.15	0.63 ± 0.20**
V	0.55 ± 0.21	0.43 ± 0.11	0.42 ± 0.16	0.45 ± 0.10	0.54 ± 0.20*	0.56 ± 0.19*
W-S	0.51 ± 0.14	0.39 ± 0.16	0.38 ± 0.13	0.36 ± 0.19	0.58 ± 0.14	0.44 ± 0.17
W-WS	0.59 ± 0.15	0.44 ± 0.09	0.47 ± 0.14	0.44 ± 0.21	0.53 ± 0.13**	0.50 ± 0.11
W-P	0.47 ± 0.10	**0.46 ± 0.05**	**0.56 ± 0.25***	0.49 ± 0.29	0.42 ± 0.13	**0.69 ± 0.17***
S-WS	0.49 ± 0.20	0.37 ± 0.17	0.37 ± 0.13	0.41 ± 0.10	**0.65 ± 0.11****	0.47 ± 0.15
S-P	0.63 ± 0.23	0.41 ± 0.10	0.46 ± 0.17	0.53 ± 0.27	0.58 ± 0.14	0.51 ± 0.14
WS-P	0.57 ± 0.13	0.44 ± 0.16	0.51 ± 0.22	**0.64 ± 0.16***	0.56 ± 0.20	0.62 ± 0.22
W-V	0.56 ± 0.12	0.36 ± 0.13	0.43 ± 0.14	0.41 ± 0.11	0.56 ± 0.19	0.62 ± 0.22
S-V	0.53 ± 0.20	0.39 ± 0.14	0.39 ± 0.15	0.40 ± 0.10	0.59 ± 0.25	0.53 ± 0.22
WS-V	0.63 ± 0.24	0.39 ± 0.12	0.46 ± 0.14	0.41 ± 0.09	0.51 ± 0.19	0.51 ± 0.19
P-V	0.56 ± 0.18	0.41 ± 0.09	0.48 ± 0.19	0.52 ± 0.17	0.59 ± 0.14	0.60 ± 0.17
W-S-WS	0.58 ± 0.19*	0.39 ± 0.16	0.38 ± 0.19	0.37 ± 0.17	0.57 ± 0.17	0.45 ± 0.16
W-S-P	0.49 ± 0.15	0.37 ± 0.16	0.43 ± 0.18	0.41 ± 0.24	0.47 ± 0.10	0.52 ± 0.10
W-WS-P	0.53 ± 0.17	**0.46 ± 0.05**	0.55 ± 0.26	0.59 ± 0.20	0.48 ± 0.13	0.65 ± 0.13
S-WS-P	0.53 ± 0.19	0.40 ± 0.15	0.42 ± 0.20	0.56 ± 0.24	0.63 ± 0.15	0.50 ± 0.11
W-S-V	0.53 ± 0.13	0.44 ± 0.06	0.39 ± 0.15	0.43 ± 0.10	0.58 ± 0.22*	0.63 ± 0.25
W-WS-V	**0.71± 0.15****	**0.46 ± 0.05**	0.39 ± 0.09	0.39 ± 0.13	0.55 ± 0.21	0.57 ± 0.18
W-P-V	0.55 ± 0.19	0.39 ± 0.16	0.49 ± 0.22	0.48 ± 0.22	0.55 ± 0.19	0.68 ± 0.20*
S-WS-V	0.54 ± 0.26	0.37 ± 0.17	0.37 ± 0.18	0.39 ± 0.11	0.60 ± 0.17**	0.49 ± 0.24
S-P-V	0.50 ± 0.13	0.33 ± 0.16	0.46 ± 0.19	0.51 ± 0.16	0.58 ± 0.19	0.50 ± 0.09
WS-P-V	0.61 ± 0.20	0.41 ± 0.09	0.42 ± 0.17	0.58 ± 0.19*	0.57 ± 0.15	0.61 ± 0.19
W-S-WS-P	0.56 ± 0.14	0.34 ± 0.16	0.47 ± 0.21	0.43 ± 0.21	0.61 ± 0.12**	0.52 ± 0.10
W-S-WS-V	0.63 ± 0.09*	0.37 ± 0.16	0.36 ± 0.10	0.41 ± 0.13	0.57 ± 0.17*	0.54 ± 0.21
W-S-P-V	0.43 ± 0.09	0.41 ± 0.10	0.46 ± 0.18	0.43 ± 0.16	0.50 ± 0.15	0.65 ± 0.16
W-WS-P-V	0.66 ± 0.12	0.43 ± 0.10	0.52 ± 0.26	0.50 ± 0.23	0.57 ± 0.15*	0.61 ± 0.19*
S-WS-P-V	0.50 ± 0.21	0.37 ± 0.16	0.50 ± 0.19	0.53 ± 0.18	0.56 ± 0.18	0.54 ± 0.15*
All	0.56 ± 0.09	0.39 ± 0.12	0.46 ± 0.18	0.41 ± 0.15	0.58 ± 0.14*	0.60 ± 0.18

*Mean accuracy (± deviation per fold) of classification for different combinations of modality types (Table 1). Bold = best result for the trait. Underline = best result while including video.(**$p < 0.01$ and *$p < 0.05$ for t-test comparison with classifier assigning labels at random.*

calculated only for the training set. The mean (± deviation per fold) median values were 3.40±0.05 for H, 3.18 ± 0.02 for E, 3.56 ± 0.01 for X, 3.12 ± 0.03 for A, 3.34 ± 0.06 for C and 3.5 ± 0.01 for O. This procedure resulted in balanced class distributions.

The first rows of Table 2 summarizes the mean accuracy and standard deviation within the folds for each trait and modality type set. Most of these results are equal or below the random baseline, with some exceptions.

Notice also that the accuracy values differ given the trait and the modality set. This shows that there is not a general modality set that would work for all traits and that each feature type has a different impact given the personality trait. Furthermore, one of the few accuracies over the random baseline is the Openness to experience (O) using the Proximity (P) features. This correlate with what has been found in previous research [8], [40], which supports that proximity features are a good indicator for this trait.

5.3 Modality Complementarity

We now proceed to evaluate the complementarity of our 5 different modality types for the binary classification of personality traits. For this purpose, we trained different classifiers with the different combinations of the modality types using early fusion. Similar to Section 5.2, we selected a L1 penalized logistic regressor with a 10-fold cross-validation, and use the median of the training set per fold to create binary labels from the personality scores.

Table 2 presents the mean accuracy and deviation for selected combinations. In addition, the significance for the results is included. This was calculated using a t-test, comparing against a classifier assigning labels at random which becomes the baseline for our experiments (50 percent chance). For each trait, the best result is in bold and the best result when the video type (V) is used is underlined, for further comparison between sources. The latter is done to better compared what was obtained in previous work [8]. This table is also separated given the number of modality types combined (double line), and given the presence of the video modality (bottom of each sub-block).

The Best and Worst Performing Traits. As seen in Table 2, our best results corresponds to the traits of Honesty (H) and Openness to experiences (O). For the trait of Openness to Experiences (O) we already obtained an acceptable result

(0.63 ± 0.2) using only proximity-based features, but with a high variance. In addition, we can see across Table 2 that combinations of this modality (P) with the types movement (W) and video (V) tend to give the better results, even when only combining 2 types. For instance, combining these 3 modality types gives an accuracy of 0.68 ± 0.2. However, the best result for this trait is obtained when combining only proximity and movement (0.69 ± 0.17). Thus, it appears that the modality of movement (either video or wearable) added complementary information to that in proximity for this trait. Furthermore, one should notice that the movement from wearables and movement from video are, from the technical perspective, recording similar features (eg. mean movement). Nonetheless, the results of each combined separately with proximity differed, with a 0.69 against a 0.60 respectively.

For the trait of Honesty (which is our best result overall) it appears that the type of movement while speaking (WS) gives the most information, complemented by the proximity and movement from video types, as can be seen in the different combinations of these in Table 2. Furthermore, for this trait we obtained (0.71 ± 0.15) when combining the modality types of movement (W), movement while speaking (WS) and movement from video (V).

Notice also that the experiment using only the modality sets from the wearable devices are similar to those presented in [8]. Nonetheless, here we used a different subset of participants (eg. only those with at least a 50 percent of time under the cameras) for a fair comparison with video. Thus, the results might vary with respect to our previous work but the general insights maintained. For example, similarly to [8], here we also found that the proximity modality is a good indicator for the trait of Openness to experience (O) or that movement with speaking give good insights about the trait of Honesty (H).

Except for the trait of Honesty, all the best results per trait are obtained when combining only 2 modalities (second sub-block of Table 2). Also, these combinations do not include video. Nonetheless, they all included a type of movement, with M-P having the best result for 3 traits (Emotionality still performs under the random baseline). Thus, these results suggest that movement features are a feasible indicator to assess personality of people during crowded and in-the-wild scenarios.

The trait of Extraversion (X) shows some of the lower performance, with the best result been 0.55 ± 0.22 accuracy while using only the set for the proximity modality (Table 2). This modality has also proved in previous efforts to be a good indication for extraversion [40]. Nonetheless, one should wonder about the low results for this trait in Tables 2 and 2, which are in most cases not even over the random baseline. We first hypothesized that the modalities of Speech (S), Proximity (P) and movement from video (V) could be a good indicator for this trait, as they record elements that will help to detect an extrovert person: speaking, proximity to others and hand gestures [2], [8], [40]. However, apart from proximity, these modalities provided performances below the random baseline when evaluated independently. This was also the case when combined. We can only imply that the hand-crafted features described in Section 4 are not representative enough to distinguish between extroverts and introverts in such a complicated setup as are the mingle scenarios. Previous work [8] showed us that fusing the information of movement while speaking and proximity increases these results. Also, we hypothesized that a better description of gestures status and if these are related to speech might be a better indicator for extraversion. Nonetheless, the detection of gestures from video in-the-wild is by itself a challenging problem that is beyond the scope of this paper.

Another interesting insight comes with the fact that fusing all modalities does not guarantees the best result. In fact, as can be seen in Table 2, none of the traits has a better performance when combining all modality types (last column). Thus, each trait is reflected differently in the modality types and this is in consequence relevant for the classification. For example, we have already discussed that features from the proximity set are a good indicator for Openness to experience (O) and that this increases when combined with the set of features from video. Nonetheless, when also combining these 2 sets with the modality sets of movement (W) or Speech (S) decreases the performance, even when compared to the result for the Proximity set only (Table 2). This shows that the type of modality itself has an important role in the estimation of the different traits and that this is not only related to more information.

The Impact of Video. With respect to video, we first must emphasize that all the results including video only have partial information, as the subjects selected were chosen to account only for those participants that were at least 50 percent of the time under the cameras. This means that, although some participants have complete video data for the entire 30 minutes, some participants have the worst case of only 15 minutes of video.

Nonetheless, Table 2 shows that movement from video (V) alone gives results over the random baseline for the traits of Honesty (H), Conscientiousness (C) and Openness to Experiences (O). Furthermore, Table 2 now shows that adding the modality of video (V) to any of the other modalities improves the overall results for these 3 traits. This suggests that the features selected for video can deal with missing data in video (until a certain point) as they are meant to be accumulative over time and still give complementary information for the other modalities. Also, when adding movement from video we also obtained acceptable results (underline in Table 2) even with missing data for some participants. Thus, we hypothesize that obtaining more data from video might further improve these results. However, when intended to increase the visibility threshold for the participants from 50 percent to 80 percent to at least, we reduce the number of participants from 56 to 22, which will not be representative enough to generalize.

5.4 Regression Experiments

In addition to the classification experiments, we also report the results for our regression experiments in this section, using a simple least square linear regressor. For validation, we also applied a 10-fold crossvalidation as in the past subsections, and report the mean value of the Mean Square Error (MSE) per fold.

Fig. 8 summarizes the results for the 6 traits using the same modality combinations as shown in Table 2. The minimum error per trait was 0.32 for H (WS-V), 0.39 for E (W), 0.23 for X (WS), 0.27 for A (WS), 0.31 for C (V) and 0.25 for O (W-P-V). These graphs are not normalized over all traits,

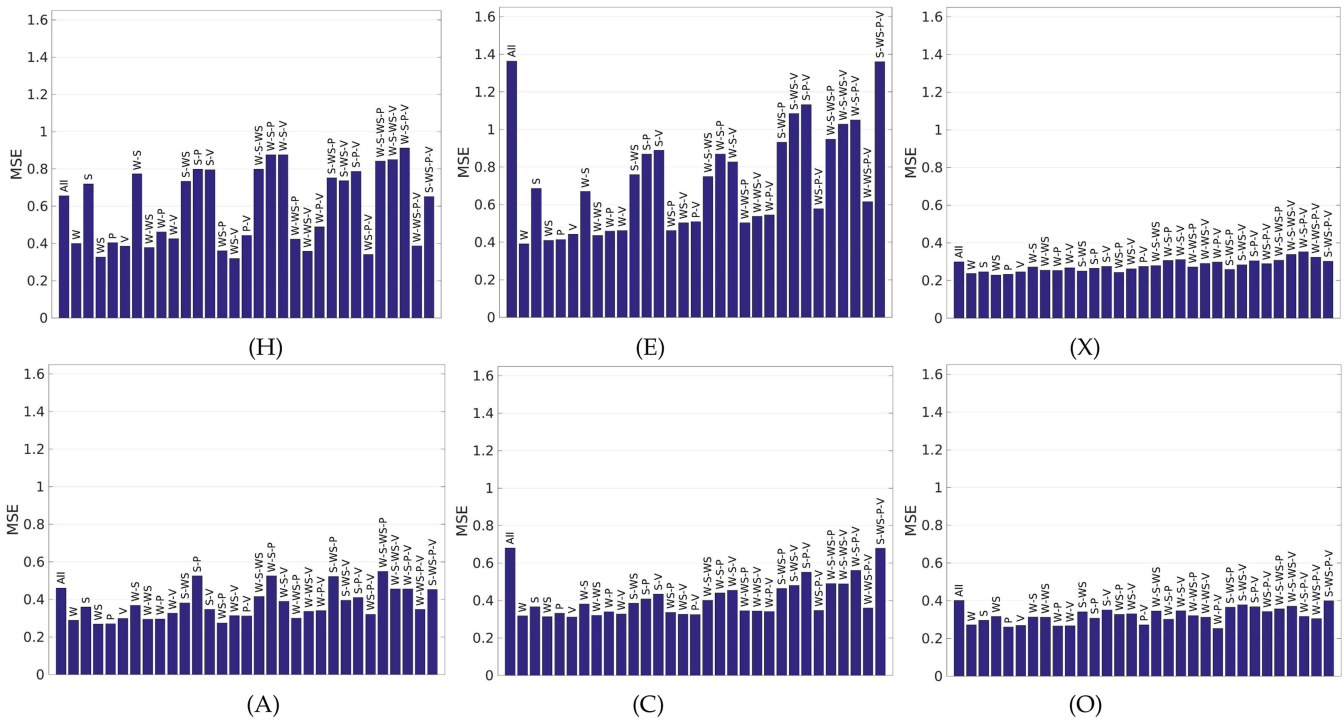


Fig. 8. MSE for regression experiments for different traits.

as the variance in scores differs greatly between them (see in Fig. 2b). Using a normalization over traits could potentially biased the results.

We can see from these results that, similar to those results in Table 2, the trait of Emotionality performed the worst of all traits. Nonetheless, all trait features reported rather acceptable results with a mean amoung the feature combinations of H = 0.60 ± 0.21, E = 0.73 ± 0.28, X = 0.28 ± 0.03, A = 0.37 ± 0.08, C = 0.41 ± 0.10 and O = 0.32 ± 0.04.

One can notice that the results reported as best performing for the classification setup (see Table 2) also have low error values in the regression. For example, the combination of W-WS-V for H, which was our best performance in the classification setting, has a MSE of 0.36 for the regression. Similarly, the best MSE result for O was obtained with the combination W-P-V. This same combination was the best performing feature combination in classification, when involving video. This reflects that our features are suitable to be used in both the regression and the classification setting.

Nonetheless, we should notice the difference between the regression and the classification results for the trait of extraversion (X). Although the best result for classification for this trait (0.56 for W-P) still is one of the lowest in the regression error, the general classification performances were rather low, mostly below the random baseline (see Table 2). In contrast, the overall MSE in the regression are rather acceptable as seen in Fig. 8(X). One possible explanation is that this difference depends on the distribution of the trait scores and the process in which the scores are converted into binary labels. We will discuss this issue further in Section 6.

5.5 Feature Importance

In this section we address the importance of each feature and feature type for the classification of the six different traits. Thus, in parallel to training different classifiers with different modality types, we also analyze the importance of each feature for the classifiers' decision.

To do so, we applied a random forest classification. Fig. 9 shows the importance of the each feature for the classifier that involves all modality types. Notice that the features with the highest importance are different for each trait. Now, let us discuss the relationship of each feature with the trait.

First, we can see from Fig. 9 that the feature for total length of speaking turns (9) is present among the top 3 more important features for H and X. This supports what has been found in previous work for extraversion [25], [34], showing that knowledge about the speaking time of the people is a good description [25]. Similarly, one or more features associated with proximity (12-14) are included in the top 3 for H, X, A and O. Once again, this aligns with the findings in previous work [21], [40] which have shown, both through analysis of social science literature [30] and empirically, that these feature types are good indicators, specially X and O.

Finally, one should notice that the features corresponding to movement from the wearables (W) and the video (V) generally appeared combined within the top three more important features. For example, for O we see that feature 18 (mean OF in the lowest bin from the video) and 2 (variance in acceleration magnitude from the wearables) are within the top 3. This is yet another insight that supports the complementarity between wearable and video movement features, as discussed in Section 5.1.

We will discuss more about the relationship between this feature importance analysis and the results obtained in Section 5.3 in Section 6.

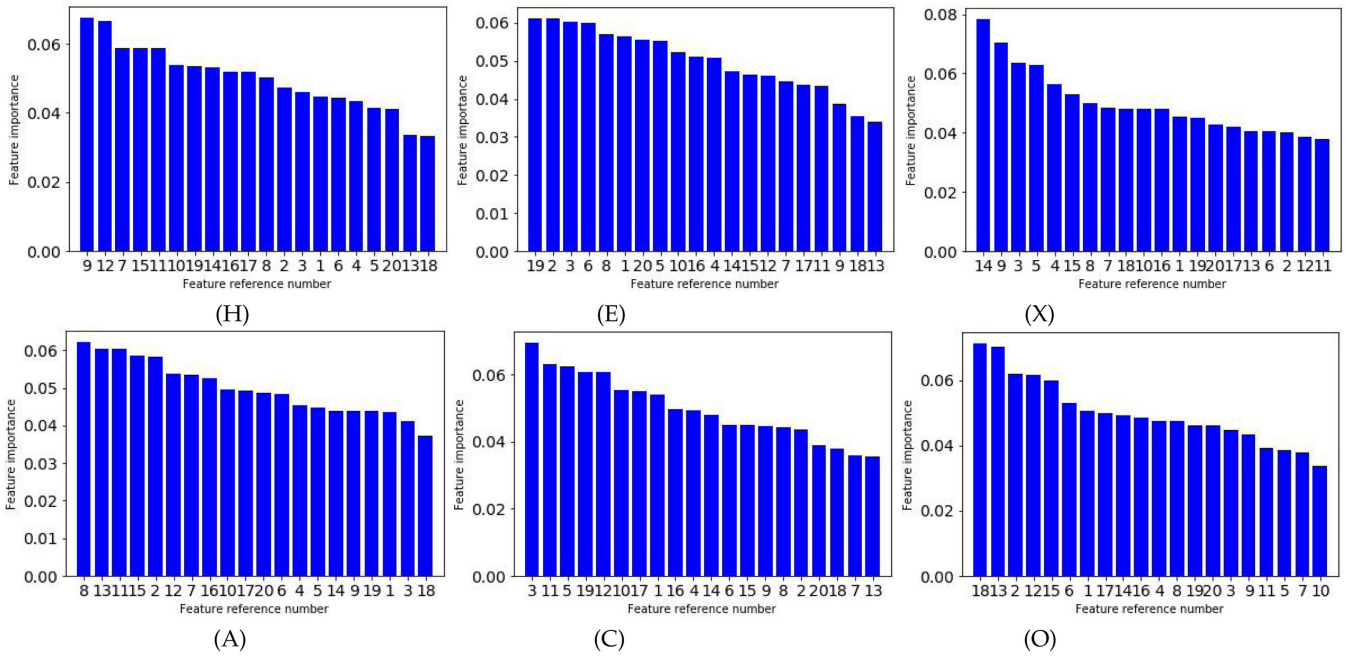


Fig. 9. Importance of all features for different traits unsing a random fores classifier.

5.6 Impact of Speech Detection on the Personality Estimation

Finally, we intended to assess the impact of the speech estimation status in the personality estimation performance. As recall, for obtaining our global speech features we relied in a transfer learning approach which extracted binary speech status from wearable acceleration (see Section 4.1.2). Now, we intend to compare the estimation of the personality traits using this speech estimation with the speech ground truth.

To do so, we used the annotated binary speech status provided by the MatchNMingle dataset [7]. These annotations were done manually by trained annotators every frame at 20 FPS. To performed a fair comparison against the speech estimation, we applied the same window to the annotations as we applied to the speech detection using TPT (see Section 4.1.2). Thus, we obtained binary time-series (speaking/no speaking) with the same number of samples that can be directly compared to those extracted from the TPT method.

From these ground truth time-series, we extracted our 7 global features for speaking (3 to 9 in Table 1) using the same process as described in Section 4.1.2. In addition, we use these streams to calculated the global features for movement while speaking (10 and 11 in in Table 1). The procedure was the same as in Section 4.1.2, when using the TPT estimation.

TABLE 3
Impact of Speech Detection in the Personality Estimation

Trait	Ground Truth		Estimated (TPT)	
	S	WS	S	WS
H	0.53 ± 0.24	0.43 ± 0.09	0.59 ± 0.23	$\mathbf{0.64 \pm 0.20}$
E	$\underline{0.40 \pm 0.11}$	0.42 ± 0.15	0.37 ± 0.19	$\mathbf{0.43 \pm 0.10}$
X	$\underline{0.49 \pm 0.17}$	0.42 ± 0.14	0.36 ± 0.13	$\mathbf{0.53 \pm 0.13}$
A	$\underline{0.53 \pm 0.20}$	$\underline{\mathbf{0.59 \pm 0.16}}$	0.50 ± 0.10	0.46 ± 0.12
C	$\underline{0.50 \pm 0.17}$	$\underline{0.40 \pm 0.17}$	0.58 ± 0.12	$\mathbf{0.60 \pm 0.17}$
O	0.53 ± 0.21	$\underline{\mathbf{0.60 \pm 0.13}}$	0.54 ± 0.17	0.44 ± 0.05

Mean accuracy ($\pm$ deviation per fold) of classification for the S and WS types using the Ground Truth and the estimated (TPT) speech. (Bold = best result for the trait in this table, underline = ground truth has better result).

Table 3 shows the comparison of the accuracy, for each trait, when using the ground truth speech and the speech estimated using our TPT. Similarly to previous sections, these results were obtained using a 10-fold cross-validation and a L1 penalized logistic regressor.

First, we can see that all the best results (for Table 3 only) correspond to the modality set of movement while speaking (WS). This is interesting as the set itself is a multimodal representation, taking into account the status in one modality (speech) to filter the information on the other (movement). Thus, also in this experiment we can see how the complementarity between modalities increases the performance of the estimation.

Furthermore, perhaps the most intriguing result in this table is that using the ground truth from the speech status does not necessarily implies a better performance in the estimation of personality traits. As can be seen in Table 3, only 5 results (those underlined in the table) are better when using the ground truth speech. One could hypothesized that improving the performance in an early stage (speech detection, in this case) will have a positive impact in the final estimation (personality). However, this experiment proves that this is not necessary the case.

For this case in particular, we hypothesize that the better performances while using the estimation for speech instead of the ground truth are due to the method that was used to detect the speech from the wearable acceleration. Our TPT method relies in the assumption that we move when we speak and hence we can use this to approximate the speaking status. Nonetheless, it is quite possible that this is not necessarily the case for all events. Thus, the speech estimation might also be taking into account movement from other components of the interaction (eg. gestures or fidgeting movements), which can be informative for the estimation of the personality traits.

6 Discussion

First, we will discuss the difference between our classification and regression results for some traits.

As previously mentioned in Section 5.3, the mean of each trait (per fold) was used as a threshold to separate the trait scores into high and low categories for a binary classification task. This is common practice in previous works on the automatic estimation of personality [34], [38].

Nonetheless, performing this separation could result in missclassifications for samples close to the median, as we are separating samples which are close together. Moreover, the non-verbal behavior between two close scores can be almost indistinguishable. Because of this issue, some prior work on personality classification only take the data from the scores below the first and above the third quartile, leaving out those samples which are around the median[37]. Their aim is to separate those people with high and low extraversion levels, and has reported better performances than the median thresholding.

The above analysis could explain why our results for the classification are not optimal for traits such as extraversion (X), but does not explain why the regression performs with acceptable errors for this trait in particular. We hypothesize that this is due to the distribution of the trait scores. As can be seen in Fig. 2b, the distribution for extraversion is somewhat uniform around its median (3.5), whereas for other better performing traits such as honesty (H) or openness (O) the distribution around each median is skew towards one class. Thus, even when both classes (high/low) have a balanced number of samples, it seems that for O, for example, these samples came from a particular bin in the scores. In contrast, for extraversion the samples for the same class are distributed across different bins. This could explain why the extraversion has so low accuracy values for classification, as the samples categorized with the same label (low/high) have different scores in reality and, quite possibly, different and more discriminative non-verbal behavior within each class.

Furthermore, we have seen that overall our findings aligned with those presented by previous work on personality estimation [8], [21], [25], [40]. Mainly, proximity feature have proven to be a good indication for the traits O or A, and speaking-based features (e.g., speaking time) are good when predicting extraversion (X).

Although must of the works cited also involved standing conversational groups, similar findings for feature types importance for given traits (e.g., speaking status or movement from video) have been made for efforts involving seated meetings or VLOGS [38]. Unfortunately, features such as proximity cannot be mapped from one setting (free/standing) to the other (seated), but we hypothesize that for setups similar to ours the features here presented could generalize rather acceptably.

We should also discuss the trait of Honesty, as this is not directly found in the Big Five inventory, and was our best performing feature when combining movement of wearable(W), movement while speaking(WS) and movement from video (V).

To properly link the modalities' impact with the trait, one must study the nature of the trait itself. Honesty is inversely equivalent to the common element shared by the Dark Triad variables [23]. Its sub-scales are sincerity, fairness, greed avoidance and modesty [4]. People with low Honesty scores *'will flatter others to get what they want, (...) and will feel a strong sense of self-importance*.[5] This might get reflected better in WS, P and V as all these modalities take into account an interaction with someone. Moreover, one can see in Fig. 9 for honesty, that within the top 3 of importance for features are those of total length of speaking and mean length of no speaking turns. This could reflect another part of the trait, as people with high honesty score tend to speak less but more truthfully [23].

Finally, note that the feature combination for the classification task with the best performing results (see Table 2) have the same trend as the importance in the feature found in Section 5.5. For instance, the best performing result for A in classification came from the combination of the feature types WS-P. Similarly, in Fig. 9, we can see that the features 11 and 13, which are in the top three most important for this trait, are features corresponding to WS and P, respectively. The same can be seen for the best results for X,C, O. Moreover, in both the classification and regression task we can see that features from the wearables (W) and the video (V), which had a low correlation as was discussed in Section 5.1, tend to perform better when combined. This furhter implies the complementarity nature of these two feature types.

7 Conclusion

In this paper we have shown a novel approach to estimate self-assessed personality during crowded mingling events leveraging wearable accelerometers, proximity and video cameras. To the best of our knowledge, we are the first to address this complex problem for a mingle scenario with such a high number of subjects (56).

We compared 5 different sets given the modality type that generate them: movement (W) from the wearable acceleration, speech (S) status obtained from a novel transfer learning method, TPT [41] to extract reliable speech information from acceleration, movement while speaking (WS), proximity from wearables (P), and movement from video (V). This comparison was done in the classification performance and as feature correlation.

Our best performing trait were Honesty (H) with a 71 percent mean accuracy when using the modality type set M-MS-V, and Openness to Experiences (O) with 69 percent mean accuracy when using W-P. When estimating all other traits, except for Emotionality (E), our method performed significantly above a random baseline.

Finally, we analyze the impact of the speech estimation in the final performance for personality trait estimation. We found that having a better estimation for the speech detection, or even the ground truth, does not necessarily reflects in a better estimation of the personality trait.

Acknowledgments

The authors will like to thank Andrew Demetriou and Leander van der Meij for they support in this work. This publication was supported by the Dutch national program COMMIT and the Instituto Tecnológico de Costa Rica.

5. Taken textually from http://hexaco.org/

REFERENCES

[1] X. Alameda-Pineda, J. Staiano, R. Subramanian, L. Batrinca, E. Ricci, B. Lepri, O. Lanz, and N. Sebe, "SALSA: A novel dataset for multimodal group behavior analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2015.

[2] M. Ashton, K. Lee, M. Perugini, P. Szarota, R. De Vries, L. DiBlas, K. Boies, and B. De Raad, "A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages," *J. Personality Social Psychology*, vol. 86, pp. 356–366, 2004.

[3] M. C. Ashton and K. Lee, "The prediction of honesty – humility-related criteria by the HEXACO and five-factor models of personality," *J. Res. Personality*, vol. 42, pp. 1216–1228, 2008.

[4] M. C. Ashton, K. Lee, and R. E. D. Vries, "The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory," *Personality Social Psychology Rev.*, vol. 18, no. 2, pp. 139–152, 2014.

[5] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: Automatic personality assessment using short self-presentations," in *Proc. Int. Conf. Multimodal Interaction*, 2011, pp. 255–262.

[6] 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.

[7] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The matchnmingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Trans. Affect. Comput.*, 2018. doi: 10.1109/TAFFC.2018.2848914.

[8] L. Cabrera-Quiros, E. Gedik, and H. Hung, "Estimating self-assessed personality from body movements and proximity in crowded mingling scenarios," in *Proc. Int. Conf. Multimodal Interaction*, 2016, pp. 238–242.

[9] L. Cabrera-Quiros and H. Hung, "Who is where?: Matching people in video to wearable acceleration during crowded mingling events," in *Proc. Int. Conf. Multimedia*, 2016, pp. 267–271.

[10] G. Castellano, S. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2007, pp. 71–82.

[11] O. Celiktutan, F. Eyben, E. Sariyanidi, H. Gunes, and B. Schuller, "MAPTRAITS 2014: The first audio/visual mapping personality traits challenge," in *Proc. Int. Conf. Multimodal Interaction*, 2014, pp. 529–530.

[12] F. Celli, E. Bruni, and B. Lepri, "Automatic personality and interaction style recognition from facebook profile pictures," *ACM Int. Conf. Multimedia*, 2014, pp. 1101–1104.

[13] M. Dobson, "Low-power epidemic communication in Wireless Ad Hoc networks," PhD thesis, Vrije Universiteit, Amsterdam, Netherlands, 2013.

[14] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scandinavian Conf. Image Anal.*, 2003, pp. 363–370.

[15] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vis. Comput.*, vol. 27, pp. 1775–1787, 2009.

[16] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Pers. Ubiquitous Comput.*, vol. 21, pp. 723–737, 2016.

[17] D. Gundogdu, A. Finnerty, J. Staiano, S. Teso, A. Passerini, F. Pianesi, and B. Lepri, "Investigating the association between social interactions and personality states dynamics," *Roy. Soc. Open Sci.*, vol. 4, 2018, Art. no. 170194.

[18] H. Hung, G. Englebienne, and J. Kools, "Classifying social actions with a single accelerometer," in *Proc. Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, 207–210.

[19] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 563–575, Oct. 2010.

[20] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations from non-verbal activity cues," *IEEE Trans. Audio Speech Language Process.*, vol. 17, no. 3, pp. 501–513, Mar. 2009.

[21] K. Kalimeri, B. Lepri, and F. Pianesi, "Going beyond traits: Multimodal classification of personality states in the wild," in *Proc. Int. Conf. Multimodal Interaction*, 2013, pp. 27–34.

[22] K. Lee and M. Ashton, "Psychometric properties of the HEXACO personality inventory," *Multivariate Behavioral Res.*, vol. 39, pp. 329–358, 2004.

[23] K. Lee and M. Ashton, "The dark triad, the big five and the hexaco model," *Personality Individual Differences*, vol. 67, pp. 2–5, 2014.

[24] B. Lepri, A. Mana, N. an Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling the personality of participants during group interactions," in *Proc. Int. Conf. User Model. Adaptation Personalization*, 2009, pp. 114–125.

[25] 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.

[26] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection," in *Proc. Workshop Tagging Mining Retrieval Human Related Activity Inf.*, 2007, pp. 9–14.

[27] C. Martella, M. Dobson, A. van Halteren, and M. van Steen, "From proximity sensing to spatio-temporal social graphs," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2014, pp. 78–87.

[28] C. Martella, E. Gedik, L. Cabrera-Quiros, G. Englebienne, and H. Hung, "How was it?: exploiting smartphone sensing to measure implicit audience responses to live performances," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 201–210.

[29] McNeill, David, *Language and Gesture*, Cambridge, U.K.: Cambridge Univ. Press, 2000.

[30] M. Mehl, S. Gosling, and J. Pennebaker, "Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life," *J. Personality Social Psychology*, vol. 90, pp. 862–877, 2006.

[31] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 273–284, Jul-Sep. 2012.

[32] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proc. Int. Workshop Social Signal Processing*, 2010, pp. 17–20.

[33] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[34] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro, "Multimodal recognition of personality traits in social interactions," in *Proc. Int. Conf. Multimodal Interaction*, 2008, pp. 53–60.

[35] V. Ponce-Lpez, B. Chen, M. Oliu, C. Corneanu, A. Claps, I. Guyon, X. Bar, H. Escalante, and S. Escalera, "Chalearn lap 2016: First round challenge on first impressions-dataset and results," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 400–418.

[36] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[37] C. Segalin, F. Celli, L. Polonio, M. Kosinski, M. Stillwell, N. Sebe, M. Cristani, and B. Lepri, "What your facebook profile picture reveals about your personality," in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 460–468.

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

[39] M. Vinciarelli, A. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, pp. 1743–1759, 2009.

[40] G. Zen, B. Lepri, E. Ricci, and O. Lanz, "Space speaks-towards socially and personality aware visual surveillance," in *Proc. 1st ACM Int. Workshop Multimodal Pervasive Video Anal.*, 2010, pp. 37–42.

[41] G. Zen, E. Sangineto, E. Ricci, and N. Sebe, "Unsupervised domain adaptation for personalized facial emotion recognition," in *Proc. Int. Conf. Multimodal Interaction*, 2014, pp. 128–135.

Laura Cabrera-Quiros received the 'Licenciatura' and master's degrees from the Instituto Tecnólogico de Costa Rica. She is a guest postdoctoral researcher with the Pattern Recognition and BioInformatics Group from Delft University of Technology. In 2014, she received a full scholarship by the Costa Rican goverment to pursue her postgraduate studies. Her main interests include use and fusion of wearable sensing and computer vision for different applications, specifically those oriented to the analysis of social behavior and health monitoring.

Ekin Gedik received the bachelor's and master's degrees from Middle East Technical University, Turkey, in 2010 and 2013, respectively. He is a postdoctoral researcher with the Pattern Recognition and Bioin-formatics Group of Delft University of Technology. His research interests include but are not limited to social behaviour analysis, wearable sensing, affective computing and pattern recognition. He is currently focused on analysis and detection of social behaviours, interaction and their connection to various social phenomena.

Hayley Hung received the PhD degree in computer vision from the Queen Mary University of London, in 2007. She is an associate professor with the Pattern Recognition and Bioinformatics group (TU Delft), head of the Socially Perceptive Computing Lab. Between 2010-2013, she held a Marie Curie Fellowship at the Intelligent Systems Lab (University of Amsterdam). From 2007 to 2010, she was a post-doctoral researcher at Idiap Research Institute in Switzerland. Her research interests include social signal processing, computer vision, and machine learning. She is a member of the IEEE.