Estimating Self-Assessed Personality from Body Movements and Proximity in Crowded Mingling Scenarios

Laura Cabrera-Quiros[1,2]* , Ekin Gedik[1]*, Hayley Hung[1]
[1]Delft University of Technology, Netherlands
[2]Instituto Tecnológico de Costa Rica, Costa Rica
{l.c.cabreraquiros,e.gedik, h.hung}@tudelft.nl

ABSTRACT

This paper focuses on the automatic classification of self-assessed personality traits from the HEXACO inventory during crowded mingle scenarios. We exploit acceleration and proximity data from a wearable device hung around the neck. Unlike most state-of-the-art studies, addressing personality estimation during mingle scenarios provides a challenging social context as people interact dynamically and freely in a face-to-face setting. While many former studies use audio to extract speech-related features, we present a novel method of extracting an individual's speaking status from a single body worn triaxial accelerometer which scales easily to large populations. Moreover, by fusing both speech and movement energy related cues from just acceleration, our experimental results show improvements on the estimation of Humility over features extracted from a single behavioral modality. We validated our method on 71 participants where we obtained an accuracy of 69% for Honesty, Conscientiousness and Openness to Experience. To our knowledge, this is the largest validation of personality estimation carried out in such a social context with simple wearable sensors.

CCS Concepts

•Computing methodologies → Supervised learning by classification; *Transfer learning;* **•Human-centered computing →** *Ubiquitous and mobile computing design and evaluation methods;*

Keywords

wearable acceleration; proximity; speaking turn; personality

1. INTRODUCTION

In the past 15 years, the automatic recognition of displayed personality has received increasing interest due to the pursuit of intelligent systems that can adapt to every individual [12]. In this paper, we focus on crowded mingling events such as cocktail parties (see Figure 1), which are intriguing scenarios due to their dynamic nature, the large number of simultaneous interactions, and the varied goals of each individual.

(a) (b)

Figure 1: Example snapshots of mingling events (a) a less crowded event taken from [13], (b) the more crowded mingle event from our secenario.

Specifically in the domain of self-assessed personality recognition during dynamic face-to-face social interactions, many previous applications focused on scenarios with a predefined task (eg. meetings). Also, the majority of prior work studied scenarios with a limited number of simultaneously occurring social interactions (generally just one such as meetings [10], and certainly lower than 5 [13]). In contrast, in this paper we investigate a scenario with on average over 15 simultaneous interactions occurring freely and dynamically.

Furthermore, audio-visual approaches are predominant in the field for predefined task scenarios due to the low number of people involved [12]. The characteristics of such scenarios enables them to set up several cameras, typically directed frontally or near frontally on participants, and microphones that capture relatively clean audio data. However, for crowded mingle events the visual boundaries between persons become harder to discriminate in the video and the noise of the event itself makes the extraction of speech features from audio more challenging. Also, recording each individual's voice could have higher perceptions of privacy invasion. Thus, wearable devices are appealing alternatives, as they inherently encapsulate the sensor data of a single individual, and are pervasive enough to avoid disturbing normal behavior and easily scalable to larger populations.

In this paper, we present an approach to automatically recognize the self-assessed personality traits from the HEXACO inventory using triaxial accelerometers and proximity sensors embedded in a wearable device hung around the neck. HEXACO is a personality inventory which includes analogous items to the well known Big-Five [1]. HEXACO also includes the trait for Honesty-Humility, which measures sincerity, fairness, greed avoidance, and modesty.

Our main contributions in this study are: (i) we address the problem of classifying self-assessed personality recogni-

*L. Cabrera-Quiros and E. Gedik contributed equally to this article

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2993170

tion in more complex and crowded mingle scenarios than previous work, where several social interactions are occurring dynamically; (ii) our approach is solely-based on sensors that can be embedded in a wearable device which makes it easily scalable, and (iii) we propose a reliable approximation of speaking status from acceleration using a transfer learning approach, resulting in improved recognition performance even when fusing cues from two *behavioural* modalities originating from a single *digital* modality.

2. RELATED WORK

We focus on works estimating *self-assessed* personality, although many efforts have been made in automated third-party attribution-based personality recognition [3]. There are also many works focused on personality estimation in social media, which are beyond the scope of this paper. A comprehensive review of the related personality computing literature can be found in [12]. Within the domain of automated self-assessed personality estimation, works can be grouped mainly into meetings and mingle scenarios.

As an example of the meeting setting, Pianesi et al. [10] 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. Batrinca et al. [2] presented a method to analyze self-presentations performed by participants in-front of a camera during a 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.

To the best of our knowledge, we are the first to address the complexity of crowded mingle scenarios using solely wearable devices. The closest work to our own was presented by Zen et al. [13]. In a considerably less crowded mingle event than ours, the authors proposed a classification method to recognize Extraversion and Neuroticism (from the Big-Five) using proximity related features extracted from multiple cameras. 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 71 people. Finally, their proximity features are based on distances while ours rely on binary neighbor detection (see Section 3).

3. OUR DATA

We collected data during three separate 2-hour social evenings in a public bar-restaurant. The participants were mostly students which signed up, for a small fee and drinks, for a speed date and a mingle event afterwards. During each event, between 30 and 32 different participants, with a total of 94 participants for the 3 events, were asked to use a custom-made wearable device hung around their neck, which recorded triaxial acceleration at 20Hz. This wearing method makes it perfect for other use-cases such as conferences, exhibitions, or business events.

Each device communicated with other devices using a radio-based beacon communication by emitting its own ID to all other devices around it in a 2-3 meter radius, allowing them also to synchronize with each other every second. These detections are considered as a binary proximity.

A 30 minutes segment from the mingle was selected to maximize the number of people interacting. We used this for our experimental validation. Due to hardware malfunction, only 71 of the devices recorded data during this segment.

Finally, 5 GoPro Hero +3 cameras recorded the event from above. This video data was only used to label the speaking status (ground truth) of 18 participants for 10 minutes to train our speaking status detector. This 10 minute segment was extracted in a non-overlapping part of the mingle from the 30 minute segment we used for testing. A snapshot of the event can be seen in Figure 1, where we contrast the density of our event with that used by Zen et al. [13].

Prior to the event, each participant filled in the HEXACO personality inventory [1], 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 [6].

4. NON-VERBAL CUES

We can group our cues, originating from 2 *digital* modalities (wearable acceleration and proximity), into 3 *behavioural* modality categories: speaking turns, body movement energy, and proximity. Detailed descriptions of each set of cues are presented below. Table 1 summarizes our derived features per cue type with a reference number.

4.1 Speaking Turns

Building on prior findings that people's speaking status is representative of their personality [2, 10, 12], 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 [8]. We have used a novel transfer learning method, Transductive Parameter Transfer (TPT) [14], which experimentally shown to perform significantly better than a traditional machine learning approach. 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.

4.1.1 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 the Linear SVM presented in [14], we have selected a L2 penalized logistic regressor as our classifier which is experimentally shown to perform better with our data. Chosen classifier minimizes Equation (1).

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

   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 θ_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 [9], instead of the independent Support Vector Regressors used in [14]. Since we need to define the similarities between distributions X_i^s instead of independent samples, we employ an Earth Mover's Distance [11] based kernel. EMD kernel is computed as:

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

In Equation (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.

4.1.2 TPT for 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 [5]) 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 71 participants during 30 minutes. 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).

4.2 Body Movement Energy

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 (similar to Section 4.1). This gave us a better representation of *movement energy* over time than the acceleration magnitude. 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 in all windows. Finally, we create 2 multi-modal behavioral features from the mean and variance of the energy values in all windows during the detected speaking turns.

4.3 Proximity

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. [7] was applied. Then, 2 features were calculated for each participant from the proximity graphs: the 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.

Table 1: Summary of our features. S.T.= Speaking turns, E.T.=entire event

	Feature	**Modality**
1	mean of accel. magnitude var. per window during E.T	Movement **(M)**
2	var. of accel. magnitude var. per window during E.T	
3	maximum length of S.T.	Speaking turns **(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.	Movement + Speaking turns **(MS)**
11	var. of accel. magnitude var. per window for S.T.	
12	largest size of group interacted with	Proximity **(P)**
13	total number of people interacted with	

Table 2: Correlations between selected features and traits ($p < 0.05$ for all correlations)

Feature	7	8	9	12	13
H	-0.419	-0.235	0.261	x	x
X	x	x	x	0.254	0.307
O	x	x	x	-0.291	x

5. EXPERIMENTAL RESULTS

5.1 Performance of TPT on Detecting Speaking Turns

First, we tested the performance of the TPT method against traditional person independent machine learning approaches on the subset of 18 participants with labels for speaking turns. In this test, we used leave-one-out cross validation. With the TPT method, each participant acted as target and all others acted as sources, once. For the traditional approaches, the other participants' data was concatenated to form the training set for each participant. Different linear (logistic regression) and non-linear classifiers (Hidden Markov Models and random forests) were used in the comparison. Paired one-tailed t-tests between performances (Area under the curve (AUC)) of these methods (Mean AUC for LR:58%, HMM:59%, RF:56%) and TPT (%65) showed TPT significantly ($p < 0.01$) outperforms all of them. Compared to the implementation in [14], which yielded an average AUC of %60, our implementation provided significantly better results ($p < 0.05$). Detailed explanation of this experiment can be found in [4]. These tests show that using TPT to extract speaking turns provides more robust and reliable results, which will allow us to have a proxy for speaking status without needing audio.

5.2 Feature-trait Correlation

Table 2 shows the correlations of the features. In this Table, only the comparisons between features and traits with a significant value are summarized. For the trait of Honesty (H), the cues related with non-speaking turns tend to have an inverse correlation with the trait, suggesting that honest people may tend to be more vocal. Interestingly, all proximity features are directly correlated with the Extraversion (X) trait. This supports the impact of proxemics (management of spatial relationship) on this trait, as found in [13].

5.3 Classification of HEXACO Traits

We treated the personality detection as a binary classification problem, where each item of the HEXACO inventory yielded one label for each participant, as positive or negative.

Table 3: Mean accuracy (%) ± std. error. M:Movement; S:Speaking turns; MS: Movement+Speaking turns; P:Proximity. Statistical significance against a random baseline is indicated:- **(p<0.01), *(p<0.05).

	Concatenated Features Combinations														
	M	**S**	**MS**	**P**	**M+S**	**M+MS**	**M+P**	**S+MS**	**S+P**	**MS+P**	**M+S+MS**	**M+S+P**	**M+MS+P**	**S+MS+P**	**M+S+MS+P**
H	59 ± 22	66 ± 17**	68 ± 17**	44 ± 12	62 ± 20*	**69 ± 15****	47 ± 20	58 ± 16	57 ± 14	62 ± 14*	58 ± 18	61 ± 22	63 ± 13**	56 ± 17	62 ± 18*
E	47 ± 7	43 ± 13	52 ± 3	52 ± 3	48 ± 12	48 ± 7	52 ± 3	45 ± 13	46 ± 13	52 ± 3	48 ± 10	46 ± 13	52 ± 3	49 ± 11	52 ± 3
X	52 ± 12	46 ± 9	48 ± 12	53 ± 15	51 ± 4	48 ± 10	59 ± 17	46 ± 13	50 ± 12	**60 ± 12***	49 ± 7	51 ± 7	**61 ± 14***	50 ± 12	54 ± 9
A	54 ± 9	52 ± 10	54 ± 8	55 ± 14	53 ± 15	55 ± 6	56 ± 15	53 ± 17	58 ± 18	59 ± 15	62 ± 10*	53 ± 12	54 ± 20	60 ± 15	**65 ± 14***
C	46 ± 19	49 ± 19	57 ± 13	46 ± 8	52 ± 16	55 ± 13	42 ± 19	56 ± 12	53 ± 13	50 ± 13	66 ± 15**	55 ± 14	49 ± 16	55 ± 20	**69 ± 15****
O	58 ± 1	56 ± 5	58 ± 1	**69 ± 17***	55 ± 9	53 ± 9	63 ± 17	58 ± 1	66 ± 14	60 ± 19	53 ± 13	48 ± 17	65 ± 18	51 ± 12	56 ± 19

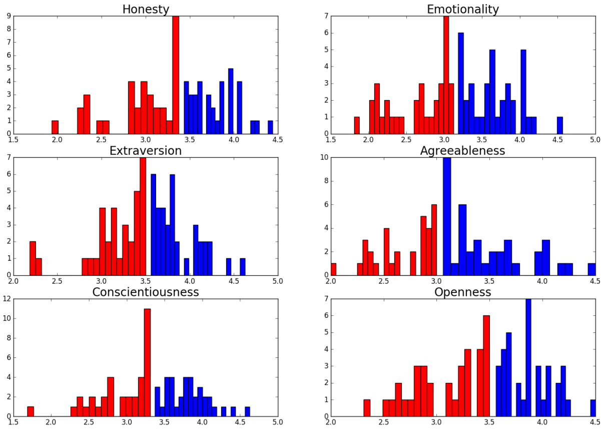

Figure 2: Distributions of personality scores per trait (Red: Negative class Blue: Positive class)

This choice, treating the problem as classification instead of regression, is based on our final aim which benefits from separating people with high/low levels in each trait. This labeling is obtained by using the median value for each item and using it as a threshold, with higher values in the positive class. This procedure resulted in fairly balanced class distributions, which are shown in Figure 2. The red and blue parts are the negative and positive classes, respectively.

By extracting the features in Table 1, we obtained 71 samples with 13 dimensions each (when all the features were used). Since we have low number of samples and feature dimensions, we selected the logistic regressor as our classifier. For performance evaluation, we used 10-fold cross validation. The optimal regularization parameter C for the logistic regressor was set using nested cross validation. The accuracies obtained for each item and with different feature combinations are provided in Table 3.

Apart from Emotionality, we were able to classify items in the HEXACO inventory significantly better than a random baseline classifier. This random classifier assigns all samples the most frequent label in the training set. To test significance for a given trait detection, we applied a paired one-tailed t-test to the performance values of our method and the random baseline classifier which are computed from each stratified test-fold. For each set of features, we compared our method against a random baseline classifier with the same set of features for statistical significance. When 2 different sets of features are compared, we compare their respective accuracies against their own baseline.

Table 3 also shows that using the multimodal set that includes all features (M+S+MS+P) provides the best general result where significant performances are obtained for three items: Honesty (H), Agreeableness (A) and Conscientiousness (C). For Honesty, significant results are obtained when speaking turn based features are in the feature set. This is quite interesting, when compared to the non-significant result obtained with just the movement energy features as it shows that we were able to extract distinguishing information (that imitates another modality) from acceleration. Also, we have seen that Feature 10, the mean of the acceleration magnitude variance in speaking turns, has the largest weight of all the features in the feature set M+MS.

Compatible with the correlation analysis of Section 5.2, we see that significant results for Openness (O) and Extraversion (X) are obtained with feature sets that include proximity based features. Significant results for Extraversion (X) are obtained when movement and proximity features are used together. This is most probably caused by the fact that extroverts tend to (i) interact with more people (captured by the proximity data), and (ii) display more body movement energy. For Openness (O), using only proximity based features was enough to obtain significant results. The contribution of multimodality is more apparent for Agreeableness (A) and Conscientiousness (C), where satisfying results are only obtained by using all features (different behavioral modalities but extracted from the same digital modality; acceleration) in combination. Adding features from other digital modality (proximity) to this combination resulted in noticeable increased performance.

6. CONCLUSION

We presented a novel approach to recognize self-assessed personality during crowded mingling events using accelerometers and proximity sensors embedded in wearable devices. To the best of our knowledge, we are the first to address this complex problem using wearable devices alone and with such a high number of subjects. We also applied a novel transfer learning method, TPT [14], to our problem to extract reliable speech information from acceleration. This allowed us to have a proxy for speech in a noisy environment like a crowded mingle event and improve our performance by fusing cues from two behavioral modalities originating from the same digital modality. Our best performing traits were Honesty (H) with a 69% accuracy when using movement (M) in combination with speech-based movement (MS), and Conscientiousness (C) with 69% accuracy when using all features. When estimating all other traits, except for Emotionality, our method performed significantly above a random baseline. Finally, we show that adding the proximity information (therefore exploiting multiple digital modalities) increases the accuracy of almost all traits. A more detailed analysis of the contribution of the behavioral cues to the different personality traits is left for future work.

7. ACKNOWLEDGMENTS

We would like to thank Leander van der Meij and Andrew Demetriou for their collaboration while collecting the data used in this work. This publication was partially supported by the Dutch national program COMMIT.

8. REFERENCES

[1] M. Ashton, K. Lee, M. Perugini, P. Szarota, R. De Vries, L. Di Blas, K. Boies, and B. De Raad. A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of personality and social psychology*, 2004.

[2] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: automatic personality assessment using short self-presentations.

[3] O. Celiktutan, F. Eyben, E. Sariyanidi, H. Gunes, and B. Schuller. MAPTRAITS 2014: The First Audio/Visual Mapping Personality Traits Challenge . *ICMI*, 2014.

[4] E. Gedik and H. Hung. Speaking status detection from body movements using transductive parameter transfer. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 69–72. ACM, 2016.

[5] H. Hung, G. Englebienne, and J. Kools. Classifying social actions with a single accelerometer. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 207–210. ACM, 2013.

[6] K. Lee and M. Ashton. Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Researc*, 2004.

[7] C. Martella, M. Dobson, A. van Halteren, and M. van Steen. From Proximity Sensing to Spatio-Temporal Social Graphs. *PerCom*, 2014.

[8] D. McNeill. *Language and gesture*, volume 2. Cambridge University Press, 2000.

[9] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[10] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal Recognition of Personality Traits in Social Interactions. *ICMI*, 2008.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000.

[12] A. Vinciarelli and G. Mahammadi. A survey of personality computing. *IEEE Trans. on Affective Computing*, 2014.

[13] G. Zen, B. Lepri, E. Ricci, and O. Lanz. Space Speaks-Towards Socially and Personality Aware Visual Surveillance . *MPVA*, 2010.

[14] G. Zen, E. Sangineto, E. Ricci, and N. Sebe. Unsupervised domain adaptation for personalized facial emotion recognition. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 128–135. ACM, 2014.