CONTINUOUS PREDICTION OF PERCEIVED TRAITS AND SOCIAL DIMENSIONS IN SPACE AND TIME

Oya Çeliktutan and Hatice Gunes

School of Electronic Engineering and Computer Science
Queen Mary University of London, United Kingdom
{o.celiktutandikici,h.gunes}@qmul.ac.uk

ABSTRACT

Developing automatic personality predictors requires generating reliable annotations, i.e., ground truth. To date, researchers have relied on the overall ratings provided for a whole video sequence, either obtained by self-assessment or provided by external observers. In this paper, we propose a novel personality assessment approach, where we ask external observers to continuously provide ratings along multiple dimensions ranging from 0 to 100 along time, and we generate continuous annotations in space and time. In addition to the widely used Big Five personality dimensions, we introduce three more dimensions that have the potential to gauge the reliability of the perceived social and trait judgements in the context of varying situational interactions between a human subject and virtual characters. Our results demonstrate the viability of the proposed approach and the plausible relationship between the extracted features and perceived trait and social dimensions. Annotations obtained continuously in time and in trait-social dimensional space showed that a number of dimensions appear to be more static and stable over time while other dimensions appear to be more dynamic.

Index Terms— Personality, Big Five model, data annotation, continuous prediction.

1. INTRODUCTION

Personality traits are essential cues to predict human behaviours, abilities and preferences in daily life such as success in academic career, personal life and relationships. In the context of human-computer interaction, personality prediction is crucial to enhance intelligent user interfaces that adapt and better respond to users' need with applications in ambient intelligence, virtual reality systems, entertainment and game technology.

The commonly used Big Five model of personality suggests that personality traits manifest themselves along five major dimensions, namely, agreeableness, conscientiousness, extroversion, neuroticism and openness to experience. Although most of the existing literature focused on a subset of personality of the Big Five model [1, 2, 3, 4], few studies also took into account other social dimensions such as likeability [5], group involvement and individual engagement [6], persuasiveness as well as its potential correlation with the five main dimensions [7].

To predict five major personality dimensions, Batrinca *et al.* [1] used short self-presentations where the subjects were asked to introduce themselves in front of a camera. Each subject, at the same time, completed a questionnaire for self-personality assessment. They extracted both vocal features (pitch, acoustic intensity, etc.) and visual features (motion vector magnitude, eye-gaze, head-face gestures, hand use and posture etc.). In another work [2], they used the same scheme for personality prediction in a context where each subject instructed by an agent had to perform a task on the computer screen and the agent could display different levels of collaboration, from agreeable and stable to less likely to compromise and neurotic.

A number of works [8, 3] focused on "video blogs" and generated annotations through the crowdsourcing scheme. Biel *et al.* [8] detected facial expressions of emotions (anger, happiness, sadness, etc.) on a frame-by-frame basis and fed emotion activity cues extracted from these sequences to Support Vector Regression (SVR) for predicting five dimensions. Their latter study [3] exploited verbal content together with the features extracted from facial expressions [8], Motion Energy Images [4] and other nonverbal cues based on speaking activity, distance to camera, and looking at the camera while speaking. The work in [9] used similar features as well as head and body activity, and social attention features (attention given by the target subject to the others, attention received from the others) in a small group meeting scenario. Although they obtained the annotations only for one minute segments, namely for thin slices, they also considered possible generalization both of regression and classification problems from slices to whole video. Following this, they proposed a cross-domain approach in [4] where Motion Energy Images were employed to train Rigde Regression and SVR classifiers from Youtube video blogs, and then the trained classifiers were tested on small group meeting data for recognizing extroversion. In the same vein, Subramanian *et al.* [10] also explored features extracted from thin slices, in particular, proximity features such as distance between the target subject and the others, the velocity of the target subject in a given time window and social attention features based on head pose in a cocktail party scenario.

In this paper, we propose a novel personality assessment approach based on continuous prediction of perceived trait and social dimensions in space and time. Although continuous prediction of affect has been well studied in the literature [11, 12], to the best of our knowledge, such changes in the personality impressions have not been explored yet. Annotations obtained continuously in time and in trait-social dimension have the potential to provide insight into which trait-social dimensions appear more dynamic, and which ones appear more static and stable over time. To test this hypothesis, we first conduct a study by asking a number of external observers to rate clips of subjects interacting with virtual agents, continuously in time and in space (i.e, ranging from 0 to 100). In addition to the Big Five model, in this study we consider three other social dimensions, namely, engagement (how engaged the person appears in the interaction), facial attractiveness (how attractive the person appears based on the face), and likeability (how one likes the person in the given context). While the engagement dimension is for evaluating the interaction between a human subject and virtual characters, other social dimensions are related to the "Halo Effect" [13] which states that observers tend to assign good attributes to the person that they find attractive or they like. Measuring the correlation between these dimensions and the Big Five model can offer insight into obtaining better interpretations of the individual judgements.

We address the prediction problem of perceived trait-social dimensions by employing histogram of gradient and histogram of optical flow [14] in conjunction with a linear regression method. Continuous annotations are used in two separate frameworks: (i) continuous prediction in space (CPS) and (ii) continuous prediction in space and in time (CPST). Our results demonstrate the viability of the proposed approach in the context of human-virtual agent interaction. Annotations obtained continuously in time and in trait-social dimensional space show that a number of traits appear to be more static and stable over time while other dimensions appear to be more dynamic.

2. DATA AND ANNOTATION

Data. In this work, we used the audio-visual recordings from the SEMAINE database [15]. This database contains recordings of human subjects interacting with virtual agents in a naturalistic scenario. We took into account 10 different subjects, each communicating with 3 semi-automatic Sensitive Artificial Listener (SAL) agents, namely, *Poppy*, *Obadiah* and *Spike*. In total, this resulted in 30 video recordings. To reduce the burden on the annotators, we shortened and segmented each video into a 60-sec clip containing several instances of turn taking. Each clip starts with the human subject engaged in the interaction as a listener, and as time progresses, the interaction becomes more established, and the subject is perceived to be more engaged with the virtual character. The 60-sec length was appeared to be sufficiently long to capture these behavioural changes and was reasonable for obtaining effective annotations.

Table 1. The annotated trait-social dimensions.

Acronym	Trait/Dimension	Acronym	Trait/Dimension
AG	Agreeableness	FA	Facial Attractive.
CO	Conscientious.	LI	Likeability
EN	Engagement	NE	Neuroticism
EX	Extroversion	OP	Openness

Annotation. For the annotation task, we used an in-house tool [16] that requires the annotator to scroll a bar between a range of values (1 and 100) along time. The clips were annotated by 21 paid participants aged between 23 and 53 years (mean = 29). Each participant rated the clips without hearing any audio and only the human subject was visible to them. Since rating along one dimension (30 videos at once) lasts approximately 45 min per participant, we divided the participants into two separate groups, each one rating the clips along a set of 5 dimensions. In total, 16 annotators (female/male: 9/7) rated all clips along the five major dimensions as well as engagement, likeability and facial attractiveness, which resulted in 8-10 annotations per clip and per dimension. The annotated trait-social dimensions are summarized in Table 1.

Analysis of Annotations. A key challenge in designing socially and emotionally intelligent user interfaces is establishing a reliable ground truth from multiple annotators. Especially, in the continuous case, solution to this problem has proven to be extremely difficult due to missing data, and variations in the speed and the style of the annotators. More specifically, time lags are likely to occur when responding to the conveyed cues or the perception and usage of rating scales may drastically differ among the annotators. Figure 1 illustrates representative annotations for one clip along *engagement* and *agreement* dimensions. One can observe that annotators hardly agree, however their judgements show similar trends. Therefore, a widely adopted approach in the literature is to compare two ratings in relative terms rather than in absolute terms, e.g., whether there has been a rise, fall or level stretch [12]. We follow this approach by applying z-score normalization and mitigate for the effects of rating scales.

For consensus analysis, correlation-based approaches, e.g., Cronbach's α coefficient, have been widely used in the literature. However, in the case of time-varying data, it is not straightforward to use these approaches. Although correlation permits comparison by shifting operations, its main limitation is not being able to incorporate warping operations such as insertion and deletion. On the other hand, Dynamic Time Warping (DTW) is an effective technique for dealing with such temporal operations. DTW algorithm searches for the optimal alignment between two sequences that minimizes

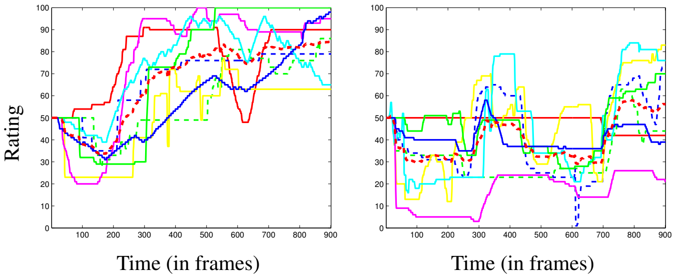

Fig. 1. Continuous annotations in time provided for two clips and two different dimensions: engagement dimension (left) and agreement dimension (right). Red dashed line illustrates the mean trajectory of the continuous annotations (best seen in colour).

Table 2. Measure of agreement among *all* and *selected* annotators in terms of mean Pearson's correlation and mean Cronbach's α in both settings.

	All Annotators		Selected Annotators	
	Pearson	Cronbach	Pearson	Cronbach
AG	0.40	0.81	0.48	0.85
CO	0.17	0.63	0.40	0.80
EN	0.41	0.82	0.51	0.87
EX	0.39	0.81	0.48	0.86
FA	0.28	0.71	0.44	0.80
LI	0.36	0.78	0.47	0.84
NE	0.35	0.82	0.45	0.88
OP	0.28	0.70	0.44	0.82

the sum of cumulative distances with respect to a locality constraint. In our experiments, we set the locality constraint to 2 sec.

After all pairs of annotations were aligned by using DTW, we measured the agreement between the annotators in terms of Pearson's correlation and Cronbach's α coefficient. We also eliminated the outliers based on the correlation values. Let N be the number of annotations for a clip, i.e, $\{x^1, .., x^N\}$. We first computed the pairwise correlations between each annotation x^i and the remaining $N-1$ annotations, $\{x^j\}_{j \neq i}$. If only the mean of its pairwise correlations was greater than a threshold, we took into account x^i to compute the ground-truth for the corresponding clip. As shown in Table 2, this approach resulted in high level of consensus among the selected annotators.

3. AUTOMATIC PREDICTION

In this section, we address the prediction problem using two separate frameworks: (i) continuous prediction in space (CPS) and (ii) continuous prediction in space and time (CPST). We only consider visual cues that are represented by low-level features derived from shape and motion. While, in CPS, our goal is to produce an overall score for the whole clip, CPST predicts a score at each time instant.

3.1. Continuous Prediction in Space (CPS)

In this case, we calculated an overall score per annotator by averaging continuous annotations over 60 sec. The ground truth was set as the mean of the aggregated scores of the annotators selected by the approach described in Section 2. This procedure resulted in an overall score in the range of 0 and 100 per clip.

As features, we used Histogram of Gradient (HoG) and Histogram of Optical Flow (HoF) extracted from the local neighbourhoods in the spatio-temporal domain. We first detected spatio-temporal interest points (STIPs) by using 3D Harris detector [14]. The local neighbourhood of a STIP was then described by concatenation of HoG and HoF values as follows. The volume was divided into a grid with $M \times M \times N$ (i.e., $3 \times 3 \times 2$) spatio-temporal blocks. For each block, 4-bin gradient and 5-bin optical flow histograms were computed and concatenated into a 162-length feature vector. Each clip was then represented by using Bag-of-Words (BoW) formalism [17]. BoW formalism provides a compact and rich description in terms of the number of local feature occurrences. It performs K-means clustering to find the representative cluster centers, and then converts the congregated local features around these centers into a histogram. In our experiments, we divided the video volume into 4 sec-long slices along the time dimension and, for each slice, we constructed histograms separately where the number of clusters was set to $K = 32$. The final representation was obtained by calculating the average of time-interval-dependent histograms.

Finally, we modelled the relationship between the BoW histograms and the aggregated scores by using Lasso and Ridge Regression [18]. A separate regression model per trait-social dimension was trained by applying *leave-one-subject-out* cross-validation strategy. Each time the parameters of the regression model were optimized over the training-validation samples with respect to the subjects. In each fold, we used 9 subjects (27 clips) for training-validation and the remaining one subject (3 clips) for testing.

3.2. Continuous Prediction in Space and Time (CPST)

For CPST, we generate the ground truth using two different strategies. The first strategy is based on taking the mean of the selected annotation trajectories per clip. Instead of creating a non-existing annotation, the second strategy selects the annotation trajectory that has the maximum correlation with the remaining annotations per clip.

CPST can be interpreted as a frame-based regressor where we treated each frame independently during prediction. In other words, we first extracted separate features per frame and then mapped these features onto the rating values at their corresponding time instant. We used the facial landmark point detector of [19] which results in 49 landmark points per frame. From each landmark, we calculated HoG+HoF values and obtained $49 \times 162 = 7938$ length feature vectors. Prior to the regression analysis, we reduced the dimension of each feature vector to 100 by applying Principal Component Analysis (PCA). We learned the relationship between the features and instantaneous rating values by using Lasso and

Table 3. Prediction results in terms of MSE and COR. The best result (high COR and low MSE) is highlighted in bold per row (social dimension) for both frameworks. MSE values are provided in the parentheses. Note that the maximum MSE value can be 4 and * indicates negligible values, i.e., $p > 0.05$.

	(a) CPS		(b) CPST	
	Lasso	Ridge	Lasso	Ridge
AG	*	**0.56** (0.03)	0.16 (0.55)	**0.24** (0.44)
CO	*	*	**0.11 (0.42)**	0.10 (0.37)
EN	*	**0.42** (0.04)	0.15 (0.47)	**0.19** (0.41)
EX	*	**0.56** (0.04)	0.17 (0.47)	**0.19** (0.39)
FA	0.77 (0.07)	**0.85** (0.04)	0.11 (0.55)	**0.13** (0.40)
LI	0.53 (0.18)	**0.75** (0.03)	0.17 (0.54)	**0.21** (0.41)
NE	**0.66 (0.07)**	0.52 (0.05)	0.13 (0.49)	**0.18** (0.38)
OP	*	**0.51** (0.03)	0.10 (0.56)	**0.14** (0.41)

Ridge Regression as in Section 3.1.

4. RESULTS AND DISCUSSION

We demonstrate the utility of the two proposed frameworks for automatically predicting eight social dimensions using visual-only annotations. Table 3 summarizes the regression results for CPS and CPST. Performance is evaluated in terms of Mean Square Error (MSE) and Pearson's correlation coefficient (COR) between the predicted values and the ground truth. As we apply z-score normalization to the annotation values for CPST, for a fair comparison, all MSE measures are normalized such that the maximum value is 4.

Table 3-a shows that the best results for CPS are obtained for facial attractiveness and likeability dimensions (COR> 0.7). This result is in accordance with our expectations as annotators felt more confident when rating the dimensions of facial attractiveness and likeability. The proposed scheme is also successful (COR> 0.55) in predicting agreeableness, extroversion and neuroticism, however, COR values for engagement and openness are found to be slightly lower. This might be due to the fact that these dimensions, especially engagement and extroversion, are perceived more dynamic and deducing an aggregated score from a time-varying annotation does not result in a representative scale.

Note that most of the results published in the literature are not directly comparable, as the annotation procedure, the data and the performance evaluation metrics employed are all different. Nevertheless, we attempt to compare our results with the methods having the most similar setup. For example, Aran and Gaticia-Perez [9] also used visual-only annotations in a meeting scenario. While they obtained the best result with weighted Motion Energy Images for extroversion ($R^2 = 0.31$), R^2 measures were found to be less than 0.1 for the other dimensions. Similar phenomenon was also reported with facial cues and audio-visual cues in [3]. Batrinca *et al.* [2] handled the problem in a classification framework and achieved high performance both for extroversion and neuroticism using audio-visual cues. We obtained comparative results with $R^2 = 0.31$[1] for extroversion and $R^2 = 0.43$ for neuroticism. We also obtained high values for agreeableness ($R^2 = 0.31$) and openness ($R^2 = 0.26$).

In Table 3-b, for CPST framework, we only reported the prediction results with the ground truth that was generated by averaging the annotation trajectories since this strategy gave the best results. At first sight it might come as a surprise that the COR values are too low compared to the results provided by the CPS framework. However, such values are commonly found in automatic continuous prediction problems. For example, in the context of affect recognition, a similar framework obtained the best correlation value of COR$= 0.22$ for predicting the valence dimension [20]. Overall, the proposed method finds a plausible relationship (COR> 0.1) for all dimensions. Table 3 also indicates that Ridge regression provides better prediction results compared to Lasso regression. This could be due to the fact that the extracted features were not sufficiently sparse for Lasso regression.

5. CONCLUSIONS AND FUTURE DIRECTIONS

This paper introduced a novel approach for personality assessment, namely, continuous prediction of perceived trait and social dimensions in space and in time. On the one hand, the CPS framework showed competitive performance for predicting extroversion in comparison with the state-of-the-art techniques in the literature, and improved prediction results for neuroticism, agreeableness and openness. On the other hand, our results demonstrated the viability of the CPST framework for automatic continuous prediction of trait and social dimensions from visual cues.

As an extension of the work proposed here, we conducted another annotation study using audio-visual clips, where the annotators watched and heard the interaction between the human subjects and the virtual characters. All clips were rated by another set of 5 annotators (female/male: $2/3$) along 8 trait-social dimensions as well as vocal attractiveness. However, at this stage we could not find any significant relationship between the visual features and the audio-visual ratings. We conjecture that audio-visual annotators might have concentrated more on the verbal content rather than the visual cues. This would require extending the work presented here to combine visual and audio features.

Recent works in psychology [21] introduced the concept of personality *state* as a behavioral episode wherein a person behaves more/less extravertedly depending on the situation. A recent study on automatic prediction of perceived traits also showed that interactive context affects the trait perception of the external observers [22]. Therefore, investigating the effects of situational context appears to be an interesting research question to focus on as future work.

Acknowledgment
This work is supported by the EPSRC MAPTRAITS Project (Grant Ref: EP/K017500/1).

[1] R^2 is calculated by taking square of the Pearson correlation coefficient.

6. REFERENCES

[1] L. M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: Automatic personality assessment using short self-presentations," in *Proc. of ACM Int. Conf. on Multimodal Interaction*, 2011.

[2] L. Batrinca, B. Lepri, N. Mana, and F. Pianesi, "Multimodal recognition of personality traits in human-computer collaborative tasks," in *Proc. of ACM Int. Conf. on Multimodal Interaction*, 2012.

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

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

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

[6] C. Oertel and G. Salvi, "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue," in *Proc. of ACM Int. Conf. on Multimodal Interaction*, 2013.

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

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

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

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

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

[12] A. Metallinou and S. S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. of Int. Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space*, 2013.

[13] Wikipedia, "Halo effect," `http://en.wikipedia.org/wiki/Halo_effect`, accessed at January 2014.

[14] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

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

[16] B. P. Motichande, "A graphical user interface for continuous annotation of non-verbal signals," Final Project, BSc FT Computer Science, Queen Mary University of London, UK, 2013.

[17] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *Proc. of IEEE Int. Conf. on Computer Vision*, 2003.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2008.

[19] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2013.

[20] E. Sariyanidi, H. Gunes, M. Gökmen, and A. Cavallaro, "Local Zernike moment representations for facial affect recognition," in *Proc. of British Machine Vision Conf.*, 2013.

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

[22] J. Joshi, H. Gunes, and R. Göcke, "Automatic prediction of perceived traits using visual cues under varied situational context," in *Proc. of Int. Conf. on Pattern Recognition (ICPR)*, 2014.