MAPTRAITS 2014: The First Audio/Visual Mapping Personality Traits Challenge

Perceived Personality and Social Dimensions

Oya Celiktutan
School of Electronic Eng. and Computer Science
Queen Mary University of London
E1 4NS London, UK
o.celiktutandikici@qmul.ac.uk

Florian Eyben
Machine Intelligence & Signal Processing
Technische Universität München
80290 München, DE
eyben@tum.de

Evangelos Sariyanidi
School of Electronic Eng. and Computer Science
Queen Mary University of London
E1 4NS London, UK
e.sariyanidi@qmul.ac.uk

Hatice Gunes
School of Electronic Eng. and Computer Science
Queen Mary University of London
E1 4NS London, UK
h.gunes@qmul.ac.uk

Björn Schuller*
Machine Intelligence & Signal Processing
Technische Universität München
80290 München, DE
schuller@tum.de

ABSTRACT

The Audio/Visual Mapping Personality Challenge and Workshop (MAPTRAITS) is a competition event aimed at the comparison of signal processing and machine learning methods for automatic visual, vocal and/or audio-visual analysis of personality traits and social dimensions, namely, extroversion, agreeableness, conscientiousness, neuroticism, openness, engagement, facial attractiveness, vocal attractiveness, and likability. The MAPTRAITS Challenge aims to bring forth existing efforts and major accomplishments in modelling and analysis of personality and social dimensions in both quantised and continuous time and space. This paper provides the details of the two Sub-Challenges, their conditions, datasets, labels and baseline features that made available to the researchers who are interested in taking part in this challenge and workshop.

Categories and Subject Descriptors

J.4 [**Social and Behavioural Sciences**]: Psychology, Sociology; H.1.2 [**Human/Machine Systems**]: Human Information Processing; I.5 [**Pattern Recognition**]: Applications—*Computer Vision, Signal Processing*

Keywords

Personality; Big Five Model; Attractiveness; Challenge

1. INTRODUCTION

The problem of assessing people's personality is very important for multiple research domains including human-computer and human-robot interaction. Despite a growing interest and emphasis on personality traits and their effects on human life in general, and recent advances in machine analysis of human behavioural signals (e.g., vocal expressions, and physiological reactions), pioneering efforts focusing on machine analysis of personality traits have started to emerge only recently. These efforts have mostly focused on unimodal cues such as written texts, audio, speech, face and body gestures, with some tentative efforts on multimodal analysis. Although personality analysis research suggests that a trait exists in all of us to a greater or lesser degree, to date none of the proposed efforts have attempted to assess personality traits continuously in time and space along the multiple trait dimensions at a given interaction time and context. The MAPTRAITS Challenge aims to focus on the aforementioned open issues and to encourage the participation of diverse research groups from different disciplines, in particular the audio and video analysis communities and those in the social sciences who study personality, traits, expressive and nonverbal behaviour, by providing a forum for interdisciplinary solutions to these challenges.

The major focus of MAPTRAITS Challenge can be summarised as follows:

- Automatic analysis and prediction of perceived trait and social dimensions of extraversion, agreeableness, conscientiousness, neuroticism, openness, engagement, facial attractiveness, vocal attractiveness, and likability.

*Dr. Schuller is also with Department of Computing, Imperial College London, SW7 2AZ London, UK.

MAPTRAITS'14, November 12-16 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2668024.2668026.

- New and/or optimal ways of extracting static vs. dynamic features from face, head, body, and voice.
- New and/or efficient techniques for automatic analysis and prediction: quantised prediction vs. continuous prediction.
- Investigating similarities and differences of video based trait and social dimension prediction versus audio-visual prediction.
- Proposing methods to synergistically combine the information in the audio stream – including acoustic and linguistic as well as non-linguistic information – and the video streams to improve trait and social dimension prediction performance.

To achieve these ambitious goals, MAPTRAITS 2014 consists of two sub-challenges for unimodal/multimodal prediction of perceived personality traits and social dimensions: (i) Mapping Personality in Quantised Space-Time, and (ii) Mapping Personality in Continuous Space-Time.

In Mapping Personality in Quantised Space-Time Sub-Challenge, ground truth ratings were built upon person's perceived personality where external observers were asked to view a video clip of a person and rate him/her along nine social dimensions based on their impressions. A rating that can take values between 1 and 10 was obtained for the whole video clip/audio clip. The goal of this sub-challenge is to find the relationship between unimodal/multimodal personality cues and these ratings.

Mapping Personality in Continuous Space-Time Sub- Challenge focuses on continuous prediction of personality/social dimensions in time and in space. The external observers were asked to generate ratings continuous in time and in space ranging from 1 to 100 as the clip of the target subject was playing. The goal of this sub-challenge is to deliver a predicted rating at each time instant, and to exploit appropriate methods to model the temporal relationships between the cues and the continuously generated ratings.

A dataset pertaining to each Sub-Challenge is provided to the participants (Section 2). Both Sub-Challenges focus on the Big Five personality traits (extraversion, agreeableness, conscientiousness, neuroticism, and openness) and four additional social dimensions. These social dimensions are engagement (how engaged the person appears in the interaction), facial attractiveness (how attractive the person appears based on the face), vocal attractiveness (how attractive the person appears based on the voice) and likability (how one likes the person in the given context). Both of the sub-challenges are open to participants using their own features and their own machine learning algorithms. However, a standard feature set is made available to the registered participants either for their own use or for comparative purposes (Section 3). The evaluation measure are mean square error, correlation, coefficient of determination (R^2) and unweighted average recall (Section 4). The datasets are released to the participants in two parts: training/validation set and test set.

2. CHALLENGE DATASETS

MAPTRAITS 2014 Challenge dataset is a subset of the well-known SEMAINE corpus [9]. This corpus has been recorded to study the behavioural changes and different affect manifestations of a user interacting with four virtual characters. Each character has a distinct emotional style and a conversational goal of shifting the user towards that emotional state, and creates a different situational context. All sequences are stored using AVI file format and are available upon request (frame rate = 50 fps, compressed with x264, CRF=12). The challenge dataset consists of two datasets: Quantised Dataset and Continuous Dataset.

2.1 Quantised Dataset

The Quantised Dataset contains 44 clips of audio-visual recordings of 11 different subjects interacting in four different situational contexts (i.e., interaction with four virtual characters). In order to analyse the effect of visual-only behavioural cues on the perception of traits, these 44 clips were first assessed by 6 raters along the five dimensions of the BF model and the four additional dimensions (engagement, facial attractiveness, vocal attractiveness, and likability). Furthermore, to analyse the effect of audio-visual behavioural cues on the perception of traits, the same 44 clips were rated by the same 6 raters together with the audio channel. The dimensions were scored on a Likert scale with ten possible values, from strongly disagree to strongly agree, mapped into the range from $[1, 10]$. All 6 raters were shown audio-visual clips and visual-only clips in random order and the responses were recorded. In total 6 raters assessed a total of 88 clips. For each setting (audio-visual and visual-only), the ground truth labels were generated by taking the average of the ratings per clip and per dimension.

As the overall objective was to analyse thin slices of behavioural responses, the extracted clips are curtailed on average to 15 s. The dataset is divided with respect to the subjects into a training/validation set (6 subjects, 24 videos) and a test set (5 subjects, 20 videos).

2.2 Continuous Dataset

The Continuous Dataset has been created for continuous prediction of traits in time and in space. The raters used AnnotationMaster tool developed by Motichande [10] to view each clip and to continuously provide scores over time by scrolling a bar between 0 and 100 as shown in Figure 1. In order to reduce the burden on the raters, recordings with Prudence were excluded and the duration of each clip was set to 60 s. Each annotation was determined in terms of the rating trace sampled at certain time intervals, e.g., at every 50 ms time interval in the 60 s period. In total, clips from 10 subjects were taken into account in three situational contexts, resulting in 30 clips.

The clips were annotated by (paid) 21 raters aged between 23 and 53 years ($mean$ = 29), from different ethnic backgrounds. The annotations were collected in three scenarios, namely, visual-only, audio-only and audio-visual.

Visual-Only (VO) Annotations. In visual only annotation, raters annotated the clips only taking into account the visual cues (audio tracks were removed) such as face/head gestures, use of hand etc., and only the human subjects were visible to them as shown in Figure 1. Since annotating one dimension (30 videos at once) lasts approximately 45 minutes per person, we also divided the social dimensions into two separate groups, each containing 5 dimensions. We nevertheless asked both of the groups to annotate *conscientiousness* and *neuroticism* as these dimensions have been found to be most challenging to perceive and annotate by the raters. In total, 16 raters (9 females, 7 males)

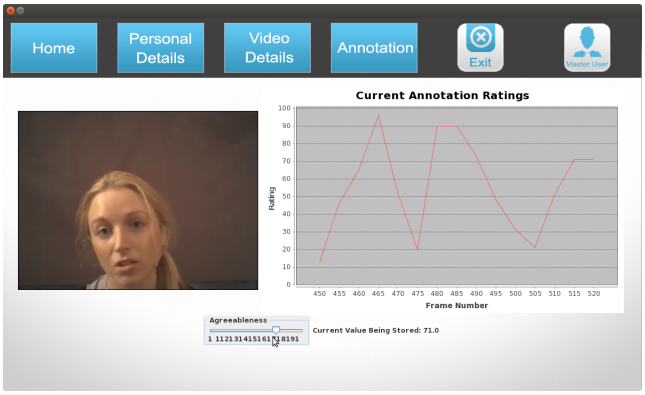

Figure 1: Illustration of the annotation tool used for generating the continuous ratings.

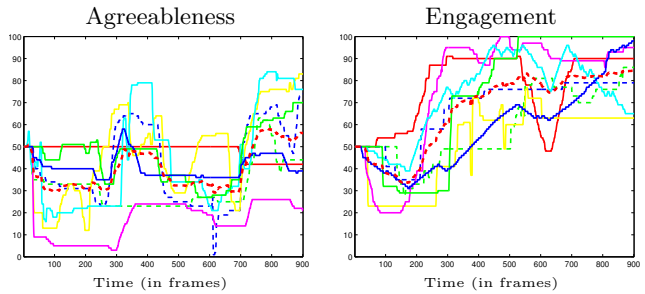

Figure 2: Representative visual-only annotations along agreeableness and engagement dimensions. The red dashed line represents the mean trajectory (best seen in colour).

annotated all clips along the BF dimensions as well as *engagement*, *likability* and *facial attractiveness*, which resulted in 32-40 annotations per video.

Audio-Only (AO) Annotations. Contrary to visual-only annotation, the focus of this annotation task was only the audio channel (human subjects were not visible to the raters). 6 raters (2 females, 4 males) were selected out of the visual-only and audio-visual raters, and employed to annotate the same clips along the BF dimensions, and *engagement*, *vocal attractiveness* and *likability*. SEMAINE recordings have stereo audio such that the subject's audio is on the left channel and the operator's audio is on the right channel. During the annotation process, we suppressed the right channel (the operator's audio), and obtained 3 annotations for each audio track.

Audio-Visual (AV) Annotations. Audio-visual annotation complements the above annotation tasks in that raters annotated the video clips together with audio channel by taking into account both visual and audio cues. All clips were assessed by 5 raters (2 females, 3 males), resulting in 25 annotations per video for all 9 social dimensions.

In the case of time-varying annotations, it is not straightforward to generate ground-truth by simply evaluating the mean value. Therefore, we first used Dynamic Time Warping (DTW) which is an effective technique for dealing with temporal operations to align the pairs of annotation trajectories, and then measured the agreement between the warped annotations in terms of Pearson's correlation. We used this approach to eliminate the inconsistent raters based on the correlation values, and to select at least 3 reliable raters per video clip/audio clip. Once the reliable raters were determined, we evaluated the mean of their corresponding original annotation trajectories per video clip/audio clip as illustrated in Figure 2 where the red dashed line represents the mean trajectory, namely the ground-truth. The details of this approach can be found in [1]. The data is divided with respect to the subjects into a training/validation set (6 subjects, 18 videos) and a test set (4 subjects, 12 videos).

3. BASELINE FEATURES

For the baseline features, MAPTRAITS 2014 Challenge provides (i) a set of visual features extracted using Quantised Local Zernike Moments (QLZM) [12], and (ii) a set of audio features extracted using the openSMILE toolkit [4] that was ranked 2nd in the open-source software competition of ACM MM 2010 and ACM MM 2013 and has become a standard toolkit in the field.

3.1 Visual Features

Visual features were extracted on a frame by frame basis. Prior to any feature extraction, we used the facial landmark point detector[1] developed by Xiong and De la Torre [16] and detected 49 landmark points per frame. These landmark points were later used to determine a rectangle enclosing the face, to crop faces based on these rectangles, and to align the faces based on the coordinates of the eye centers by affine transformation. The cropped and aligned faces were finally resized such that each face has the size of 128×128.

As baseline visual features, Quantised Local Zernike Moments (QLZM) [12] were extracted on the cropped, aligned, resized faces per frame. QLZM can be interpreted as a low-level representation that calculates local Zernike Moments (ZMs) in the neighbourhood of each pixel of the input face image and converts the accumulated local features into position dependent histograms. Each ZM coefficient describes the texture variation at a unique scale and orientation, and the information conveyed by different ZMs does not overlap [14]. Once the ZMs are computed for all pixels, the QLZM descriptors are obtained by quantising all ZM coefficients around a pixel into a single integer. The quantization process yields the QLZM image, which is then divided into subregions based on two grids, an inner partitioning and an outer partitioning (cf. Figure 3). The two-fold partitioning intends to reduce the sensitivity against registration errors. A position dependent histogram is computed for each subregion, and the final representation is obtained by concatenating these position-dependent histograms.

The dimensionality of the QLZM representation depends on the number of ZMs used, and the number of subregions. In this Challenge, we considered two ZMs that result in 16-bin local histograms as in [12]. We divided the face into subregions by applying a 5×5 outer grid and a 4×4 inner grid which yielded a 656-dimensional feature vector.

We also considered a part-based representation that enables the extraction of a smaller set of features. We first determined the facial parts, namely, the two eye areas and the mouth area, and extracted QLZM features on these parts. For part-based representation, we did not divide the eye and

[1]This code is publicly available at: http://www.humansensing.cs.cmu.edu/intraface/

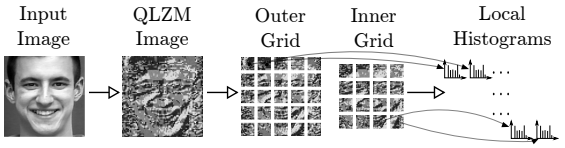

Figure 3: Illustration of the QLZM representation.

mouth areas into subregions. This resulted in feature vectors having a length of 48.

The 656-dimensional and 48-dimensional QLZM features as well as 49 landmark points per frame are readily provided to the challenge participants. We also provide the code to extract QLZM features [12].

3.2 Audio Features

We considered the same set of audio features used for the the INTERSPEECH 2013 Computational Paralinguistics ChallengE (COMPARE) [13]. Features were extracted by employing TUM's open-source openSMILE feature extractor [4] in its recent 2.0 release [2]. The feature set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs) as well as a few LLDs including logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. First order delta regression coefficients are applied to all LLDs and a sliding window mean smoothing over three frames is applied to the LLDs and deltas. A comprehensive set of functionals is applied to the smoothed LLDs and deltas in order to summarise them over segments of interest (sliding windows or the full recording, more details given in the next paragraph), resulting in 6,373 features per segment. Details on these features are given in [15]. Since the recordings provided for the MAPTRAITS 2014 Challenge are up to 60 s long and contain long speech pauses, a voice activity detection (VAD) score (cf. [3]) was included as an additional LLD with the functionals mean, flatness, and standard deviation applied, resulting in 6,376 features in total.

The LLDs for all files in the quantised and the continuous set are readily provided to the challenge participants. The LLDs are extracted at a rate of 10 ms from overlapping windows, where pitch based LLDs are extracted from 60 ms windows and all other LLDs (spectral, cepstral, energy based) are extracted from 20 ms windows. Further, for both sets, a sliding window scheme was used to apply functionals to the LLDs and obtain feature vectors at a constant rate. Overlapping windows of 4 seconds and 2 seconds length are used, shifted forward at a rate of 0.5 s and 1 s. This results in four different variations of the sliding window features, which participants can use to find the best set. Baseline results for these features will be published soon. For the quantised data set, an additional set of features is provided: functionals are there applied to the full length of the provided clips (14 seconds), yielding one feature vector for each clip. For the LLDs and the sliding window functional features, the mean VAD score can be used to remove segments which have no or only little speech content. The VAD scores are in the range from -1 to +1, where a logical threshold for the voiced/un-voiced decision is given at 0 by the employed VAD method. However, in practice, for the mean VAD score, we recommend a threshold of 0.25, i. e., segments (= feature vectors) where the mean VAD score is < 0.25 should be removed.

4. PERFORMANCE EVALUATION

In this section, we introduce the evaluation metrics for experimental assessment and performance comparison of different algorithms for both tasks, i.e., quantised space-time and continuous space-time.

4.1 Evaluation Metrics for Quantised Space-Time

For mapping personality in quantised space-time, we propose the following metrics: i) mean square error (MSE), ii) sample Pearson's correlation coefficient (COR), iii) coefficient of determination (R^2), and iv) unweighted average recall (UAR).

MSE is frequently used to measure the differences between the values predicted by a model and ground-truth values, namely, to measure the model prediction error. Let y_k and $\hat{y}_k$ be the ground-truth and predicted values, respectively, MSE can be defined as:

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2 \tag{1}$$

where N is the number of predictions. One shortcoming of MSE is being scale-dependent. On the other hand, COR and R^2 are invariant to scale changes. First, COR measures the linear correlation between the ground-truth and the predicted values, and can be formalised as:

$$COR = \frac{\sum_{k=1}^{N}(y_k - \mu_{y_k})(\hat{y}_k - \mu_{\hat{y}_k})}{\sqrt{\sum_{k=1}^{N}(y_k - \mu_{y_k})^2}\sqrt{\sum_{k=1}^{N}(\hat{y}_k - \mu_{\hat{y}_k})^2}} \tag{2}$$

where μ_{y_k} and $\mu_{\hat{y}_k}$ are the sample mean of ground-truth and predicted values, respectively. Secondly, R^2 grades the goodness of the fit of the model, in our case, it measures how well the learned model fits the unseen samples, and can be defined as:

$$R^2 = 1 - \frac{\sum_{k=1}^{N}(y_k - \hat{y})}{\sum_{k=1}^{N}(y_k - \mu_{y_k})} \tag{3}$$

While COR yields a value between -1 and 1, i.e., $COR > 1$ is positive correlation, and $COR < 1$ is negative correlation, R^2 gives a measure between 0 and 1, i.e., large values indicate high goodness of the fit, and negative values are not meaningful.

As an alternative measure for quantised space-time, we can treat the regression problem as a classification task and use unweighted average recall (UAR) to measure the performance. UAR is defined as follows:

$$UAR = \frac{1}{N_c}\sum_{i=1}^{N_c}\frac{\sum_{k=1}^{N_i}\delta({y_k}^i, \hat{y}_k^i)}{N_i} \tag{4}$$

where N_c is the number classes, in our case, the number of values that the ratings can take, i.e., between 1 and 10. ${y_k}^i$ and $\hat{y}_k^i$ are the ground-truth and the predicted values pertaining to the ith class, respectively. $\delta(x, y)$ gives 1 if $x = y$, otherwise it is equal to 0.

4.2 Evaluation Metrics for Continuous Space-Time

For mapping personality in continuous space-time, we propose the following metrics: i) mean square error (MSE), ii) sample Pearson's correlation coefficient (COR), iii) cross

correlation coefficient ($XCOR$), and iv) concordance correlation coefficient ($CCOR$). Differently from the quantised space-time, we evaluate these measures between the pairs of rating trajectories – a ground-truth trajectory and a predicted trajectory – and calculate the mean of the measures over all predictions. More explicitly, average MSE can be re-formalised as follows:

$$MSE_k = \sum_{t=1}^{T} (y_k(t) - \hat{y}_k(t))^2 \quad (5)$$

where $y_k(\cdot)$ and $\hat{y}_k(\cdot)$ are the respective ground-truth and predicted trajectories, t indexes the time instances and T is the number of time instances, namely, the length of the trajectory. We can now calculate average MSE by $MSE = \frac{1}{N}\sum_{k=1}^{N} MSE_k$. This also applies to correlation measures, for example, we can define COR_k as:

$$COR_k = \frac{\sum_{t=1}^{T}(y_k(t) - \mu_{y_k})(\hat{y}_k(t) - \mu_{\hat{y}_k})}{\sqrt{\sum_{t=1}^{T}(y_k(t) - \mu_{y_k})^2}\sqrt{\sum_{t=1}^{T}(\hat{y}_k(t) - \mu_{\hat{y}_k})^2}} \quad (6)$$

Similarly, $XCOR_k$ can be defined as:

$$XCOR_k = \max_{\Delta} \frac{\sum_{t=1}^{T-\Delta-1}(y_k(t+\Delta) - \mu_{y_k})(\hat{y}_k(t) - \mu_{\hat{y}_k})}{\sqrt{\sum_{t=1}^{T}(y_k(t) - \mu_{y_k})^2}\sqrt{\sum_{t=1}^{T}(\hat{y}_k(t) - \mu_{\hat{y}_k})^2}} \quad (7)$$

where Δ is the shift in the time domain. For example, in our experiments (Section 5.2), we set $\Delta = 15$, which corresponds to 1 s when the sampling rate is 66 ms (as in the visual-only annotations). Finally, $CCOR$ measures the agreement between two trajectories as follows

$$CCOR_k = \frac{\sum_{t=1}^{T}(y_k(t) - \mu_{y_k})(\hat{y}_k(t) - \mu_{\hat{y}_k})}{s_{y_k}^2 + s_{\hat{y}_k}^2 + (\mu_{y_k} - \mu_{\hat{y}_k})^2} \quad (8)$$

$$s_{y_k}^2 = \sqrt{\sum_{t=1}^{T}(y_k(t) - \mu_{y_k})^2}; \quad s_{\hat{y}_k}^2 = \sqrt{\sum_{t=1}^{T}(\hat{y}_k(t) - \mu_{\hat{y}_k})^2}.$$

$CCOR$ is frequently used to measure the inter-rater reliability. While COR and $XCOR$ gives an idea whether the ground-truth and the predicted values lie in the same direction or not, $CCOR$ also measures how far the ground-truth values lie from the predicted values. Similar to the COR, $CCOR$ yields values between -1 and 1, i.e., $CCOR > 0$ indicates a positive agreement.

5. BASELINE RESULTS

For both of the Sub-Challenges, we presented the baseline results in three settings: visual-only, audio-only and audio-visual.

5.1 Mapping Personality in Quantised Space-Time

Visual-Only (VO) Results. We represented each video by the mean and the standard deviation of 48-length QLZM feature vectors. We first divided the video volume into 2 s-long overlapping slices at a rate of 2 s along the time dimension. For each slice, we computed the mean of QLZM feature vectors. The final representation was obtained by calculating the standard deviation of time-interval-dependent mean QLZM feature vectors. We modelled the relationship between these features and the annotations by using ridge regression [7]. Prior to any analysis, we scaled each feature so that it falls to the range of $[-1, 1]$. In order to learn the optimum parameters, i.e. λ in ridge regression, we applied 6-fold cross validation with respect to the subjects by scanning over the values $\lambda \in \{e^{0.5}, e^{0.505}, \ldots, e^{5}\}$. Baseline results are provided in Table 1.

Audio-Only (AO) Results. The audio baseline results were obtained with SVM (multiple one-vs-one SVMs) using the WE-KA data-mining toolkit [6][2]. Models were trained with the Sequential Minimal Optimisation (SMO) algorithm [11]. Based on the size of the quantised data-set and our past experience, the complexity C was set to $C = 0.5$. All baseline features were used, i.e., no feature selection was performed. z-normalisation (mean zero and unit variance) was applied to the training and test set independently. The predictions obtained on the test set are at a rate of 2 Hz or 1 Hz, depending on the window and step size of the audio baseline features, or a single prediction in case of the full length feature vectors.

We obtained the best results when functionals of LLD are computed over sliding windows 4 s long shifted forward at a rate of 1 s. We therefore reported only the results with the corresponding features in Table 1. In terms of the MSE metric, audio-only results were better than visual-only results for the *agreeableness, conscientiousness* and *neuroticism*. In terms of the COR metric, the visual-only results constantly outperform audio-only results. This may be due to the usage of a regressor for visual-only results while using a classifier for audio-only results.

Audio-Visual (AV) Results. We applied a rather simple approach to obtain multimodal prediction results. We combined two single-modality prediction results at the decision level by taking average of the predictions of the two modalities. As shown in Table 1, this simple approach improved the MSE performance compared to both audio- and visual-only results for all dimensions, except for *conscientiousness*. However, COR performance is lower than visual-only for all dimensions, except for *neuroticism* and *agreeableness*.

5.2 Mapping Personality in Continuous Space-Time

Visual-Only (VO) Results. In this setting, we treated each frame independently during prediction. In other words, we used the 656-length QLZM features per frame and then mapped these features to the rating values at the corresponding time instant. The feature vectors were also sampled at every 66 ms similarly to obtaining the visual-only annotations. We trained a model to learn the relationship between the features and instantaneous rating values by using ridge regression [7]. Prior to any analysis, we scaled each feature so that it falls to the range of $[-1, 1]$. In order to learn the optimum parameters, i.e. λ in ridge regression, we applied 6-fold cross validation with respect to the subjects by scanning over the values $\lambda \in \{e^{0}, e^{0.25}, \ldots, e^{30}\}$. We observe that, in general, raters were not active in the first 1 s of the annotation task. For this reason, we do not take into account the first 10 time instants while learning/fitting the models.

Baseline results are provided in Table 2. $CCOR$ has been found to be lower as compared to COR and $XCOR$ measures. We conjecture that the ground-truth ratings and the

[2]The SMOreg class is used in our experiments.

Table 1: Prediction results for quantised space-time under three settings.

	Visual-Only				Audio-Only				Audio-Visual			
	MSE	COR	R^2	UAR	MSE	COR	R^2	UAR	MSE	COR	R^2	UAR
AG	1.58	0.06	-56.75	0.20	1.45	-0.14	-30.89	0.27	1.24	0.07	-12.00	0.25
CO	1.11	-0.16	-71.22	0.19	0.57	-0.47	-148.66	0.29	0.61	-0.37	-167.17	0.24
EN	1.05	-0.14	-3.81	0.20	1.16	-0.22	-68.10	0.15	0.97	-0.26	-40.83	0.15
EX	0.71	0.38	-8.71	0.22	1.07	-0.11	-55.55	0.23	0.66	0.25	4.34	0.29
FA	2.29	0.12	-72.13	0.20	-	-	-	-	-	-	-	-
LI	1.66	0.02	-47.34	0.20	2.29	-0.52	-98.92	0.25	1.83	-0.39	-59.01	0.25
NE	3.98	0.02	-53.84	0.14	3.85	0.02	-86.69	0.17	3.20	0.06	-55.47	0.22
OP	0.70	0.22	-64.26	0.25	0.96	-0.21	-123.48	0.19	0.67	-0.24	-56.27	0.30
VO	-	-	-	-	2.10	-0.13	-104.46	0.21	-	-	-	-

Table 2: Prediction results for continuous space-time under three settings.

	Visual-Only				Audio-Only				Audio-Visual			
	MSE	COR	$XCOR$	$CCOR$	MSE	COR	$XCOR$	$CCOR$	MSE	COR	$XCOR$	$CCOR$
AG	0.41	0.21	0.27	0.01	0.50	0.13	0.15	0.10	0.80	0.23	0.26	0.23
CO	0.28	0.11	0.14	0.00	0.51	0.07	0.09	0.03	1.04	0.07	0.10	0.07
EN	0.45	0.12	0.15	0.01	0.65	0.16	0.17	0.13	1.03	0.05	0.08	0.05
EX	0.38	0.24	0.27	0.02	0.66	0.21	0.24	0.17	1.00	0.09	0.11	0.09
FA	0.34	0.09	0.15	0.00	-	-	-	-	-	-	-	-
LI	0.37	0.16	0.22	0.01	0.56	0.19	0.21	0.15	0.78	0.27	0.30	0.27
NE	0.35	0.11	0.17	0.00	0.77	0.01	0.02	0.01	0.87	0.13	0.16	0.12
OP	0.39	0.09	0.12	0.00	0.53	0.26	0.28	0.21	0.91	0.24	0.26	0.24
VO	-	-	-	-	0.41	0.12	0.15	0.11	-	-	-	-

predicted ratings have similar trends, however their difference in absolute terms is rather large. Based on the current baseline results, *extroversion*, *agreeableness* and *likability* seem to be better modelled in continuous space and time. On the other hand, *facial attractiveness*, *openness*, *consciousness* are better predicted in quantised space and time.

Audio-Only (AO) Results. The audio baseline results were obtained using SVR using the WEKA data-mining toolkit [6][2]. Models were trained with the Sequential Minimal Optimisation (SMO) algorithm [11]. Based on the size of the continuous data-set and our past experience, the complexity C was set to $C = 0.005$. All baseline features were used, i.e., no feature selection was performed. z-normalisation (mean zero and unit variance) was applied on the data of each speaker, both in the training and the test set. The predictions obtained on the test set are at a rate of 2 Hz or 1 Hz, depending on the window and step size of the audio baseline features. In order to have a common rate for scoring, the predictions were resampled to a rate of 20 Hz (50 ms period) to match the original ratings. The re-sampling was done by repeating each prediction 10 or 20 times, respectively. The first 20 (or 40, respectively) values of the 20 Hz predictions were filled with zeros, to align the first prediction result with the center of the first audio feature analysis window (2 or 4 seconds, respectively). Similarly, the last prediction was repeated for the last 20 (or 40, respectively) values to match the number of the original predictions.

We obtained the best results when functionals of LLD are computed over sliding windows 4 s long shifted forward at a rate of 1 s. We therefore reported only the results with the corresponding features in Table 2. Taking into account MSE and COR, results, visual-only-based approach yielded better results in overall, while audio-only-based approach significantly improved the prediction of *openness*. On the other hand, considering $CCOR$ results, we obtained better results with audio-only-based approach in all dimensions. This can be due that support vector regression works better in absolute terms. In addition to *openness*, the prominent dimensions are *extroversion* and *likability*.

Audio-Visual (AV) Results. To obtain multimodal prediction results, we applied the same approach introduced in Section 5.1. Briefly, we combined two single-cue prediction results at the decision level by taking average of the predicted trajectories along time dimension. As given in Table 2, decision-level fusion developed into better results for *likability* and slightly better results for *agreeableness*. However, to further improve the results, we reached the conclusion that one should combine two modalities at the feature level as well.

6. CONCLUSIONS

MAPTRAITS 2014 Challenge & Workshop aimed at speeding up the progress in automatic analysis of personality traits and social dimensions by providing a benchmarking protocol for this problem and encouraging various research groups to participate in developing novel solutions. We designed two sub-challenges, the Quantised Space-Time Sub-Challenge and the Continuous Space-time Sub-Challenge. We provided a dataset for each sub-challenge and used a variety of metrics to evaluate the performance of participating systems from multiple perspectives. These datasets and

metrics will serve as a benchmarking protocol not only for the systems participating in this competition, but also for future studies that will consider the automatic analysis of personality traits and social dimensions.

The baseline systems have been designed taking into account the trends and progress in other closely related fields such as automatic facial/vocal affect analysis. Yet, our baseline is merely a starting point as the automatic analysis of personality traits and social dimensions is a problem that has its own patterns and dynamics and is likely to benefit from tailored solutions. In fact, our discussion in Section 5 suggests that each sub-problem (i.e. analysis in quantised vs. continuous space-time) may differ in their optimal solutions as the features and modalities that proved useful were not necessarily similar for both problems.

Several works in psychology [5] have emphasized the dynamic components of the personality and the concept of the *personality states*. While personality traits are defined to be fairly stable across time, *personality states* describe personality changes on shorter time scales and across different situations. A recent study on automatic prediction of perceived traits also showed that interactive context affects the trait perception of the external observers [8]. The future MAPTRAITS Challenge and Workshops will be designed by taking into account these new trends.

Despite the increasing research efforts, there still exist many open problems and a large room for improvement in automatic analysis of personality traits and social dimensions. We hope that the first version of MAPTRAITS Challenge and Workshop will push the state of the art in the field towards novel approaches for unimodal and multimodal feature extraction, fusion and prediction, and towards new research directions.

Acknowledgments

The work of Oya Celiktutan, Evangelos Sariyanidi and Hatice Gunes has been supported by the MAPTRAITS Project funded by the Engineering and Physical Sciences Research Council UK (EPSRC) (Grant Ref: EP/K017500/1).

7. REFERENCES

[1] O. Celiktutan and H. Gunes. Continuous prediction of perceived traits and social dimension in space and time. In *Proc. of IEEE Int. Conf. on Image Processing*, Paris, France, 2014.

[2] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 835–838. ACM, 2013.

[3] F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life Voice Activity Detection with LSTM Recurrent Neural Networks and an Application to Hollywood Movies. In *Proceedings 38th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2013*, pages 483–487, Vancouver, Canada, May 2013. IEEE, IEEE.

[4] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia*, pages 1459–1462. ACM, 2010.

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

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations Newsletter*, 11(1):10–18, June 2009.

[7] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2008.

[8] 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.

[9] 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*, 3(1):5–17, 2012.

[10] 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.

[11] J. Platt. Sequential Minimal Optimization: A fast algorithm for training Support Vector Machines. Technical Report MSR-98-14, Microsoft Research, 1998.

[12] 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.

[13] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. *Proceedings of Interspeech*, 2013.

[14] M. R. Teague. Image analysis via the general theory of moments. *Journal of the Optical Society of America (1917-1983)*, 70:920–930, Aug. 1980.

[15] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer. On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Emotion Science*, 4(Article ID 292):1–12, May 2013.

[16] 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.