Journal on Multimodal User Interfaces (2021) 15:337–358
https://doi.org/10.1007/s12193-020-00347-7

ORIGINAL PAPER

Multimodal analysis of personality traits on videos of self-presentation and induced behavior

Dersu Giritlioğlu[1] · Burak Mandira[1] · Selim Firat Yilmaz[2] · Can Ufuk Ertenli[3] · Berhan Faruk Akgür[4] · Merve Kınıklıoğlu[4] · Aslı Gül Kurt[4] · Emre Mutlu[5] · Şeref Can Gürel[6,7] · Hamdi Dibeklioğlu[1]

Received: 26 May 2020 / Accepted: 29 September 2020 / Published online: 2 November 2020

Abstract

Personality analysis is an important area of research in several fields, including psychology, psychiatry, and neuroscience. With the recent dramatic improvements in machine learning, it has also become a popular research area in computer science. While the current computational methods are able to interpret behavioral cues (e.g., facial expressions, gesture, and voice) to estimate the level of (apparent) personality traits, accessible assessment tools are still substandard for practical use, not to mention the need for fast and accurate methods for such analyses. In this study, we present multimodal deep architectures to estimate the Big Five personality traits from (temporal) audio-visual cues and transcribed speech. Furthermore, for a detailed analysis of personality traits, we have collected a new audio-visual dataset, namely: Self-presentation and Induced Behavior Archive for Personality Analysis (SIAP). In contrast to the available datasets, SIAP introduces recordings of induced behavior in addition to self-presentation (speech) videos. With thorough experiments on SIAP and ChaLearn LAP First Impressions datasets, we systematically assess the reliability of different behavioral modalities and their combined use. Furthermore, we investigate the characteristics and discriminative power of induced behavior for personality analysis, showing that the induced behavior indeed includes signs of personality traits.

Keywords Big five · Estimation of personality traits · Deep learning · Multimodal fusion · Self-presentation · Induced behavior

Dersu Giritlioğlu, Burak Mandira, Selim Firat Yilmaz, Can Ufuk Ertenli have equally contributed.

✉ Dersu Giritlioğlu
dersu@bilkent.edu.tr

Burak Mandira
burak.mandira@bilkent.edu.tr

Selim Firat Yilmaz
syilmaz@ee.bilkent.edu.tr

Can Ufuk Ertenli
ufuk.ertenli@metu.edu.tr

Berhan Faruk Akgür
faruk.akgur@bilkent.edu.tr

Merve Kınıklıoğlu
m.kiniklioglu@bilkent.edu.tr

Aslı Gül Kurt
gul.kurt@bilkent.edu.tr

Emre Mutlu
mutluemre12@gmail.com

Şeref Can Gürel
scgurel@hacettepe.edu.tr

Hamdi Dibeklioğlu
dibeklioglu@cs.bilkent.edu.tr

[1] Department of Computer Engineering, Bilkent University, Ankara, Turkey

[2] Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey

[3] Department of Computer Engineering, Middle East Technical University, Ankara, Turkey

[4] Department of Neuroscience, Bilkent University, Ankara, Turkey

[5] Psychiatry Clinic, Etimesgut Şehit Sait Ertürk State Hospital, Ankara, Turkey

[6] Department of Psychiatry, Hacettepe University, Ankara, Turkey

[7] Department of Cognitive Neuroscience, Maastricht University, Maastricht, The Netherlands

1 Introduction

Among many other theories of personality analysis, trait-based ones are widely accepted in the literature [59]. Trait theory bases on the measurement of general patterns of behaviors, thoughts, and emotions. These patterns (traits) are relatively stable over time and in different contexts [21]. Although there have been different trait models broadly examined such as the Big Five [59], the Big Two [2], and the HEXACO model [7], the Big Five Model is currently the dominant and well-established paradigm in personality research. According to Goldberg [32], the model consists of five independent traits, namely, Openness to experience (being curious to experience new things and imaginative), Conscientiousness (being dutiful and self-disciplined), Extraversion (being gregarious and active), Agreeableness (being tolerant and trusting), Neuroticism (being inclined to notice threatening factors in non-threatening situations and tendency to experience negative emotions). McCrae and John provide empirical and theoretical foundations of the Five Factor model: It integrates various personality constructs; it is comprehensive (provides a way of systematic exploration of the relations between personality and other phenomena) and it is efficient (providing a global description of personality with as few as five scores) [59]. Table 1 demonstrates some example adjectives for the high scorers of each of these traits.

These traits are mainly associated with cognition, affect, and non-verbal behavior, such as gaze, head movement, body pose and facial appearance. Conscientiousness is dominated more by behaviour, Neuroticism by negative affect alongside these kind of behaviours, Extraversion by both affective and behavioral, and lastly Openness and Agreeableness by cognitions [91]. [91] suggests that certain personality traits are more visible to eye than some. In this sense, traits such as Extraversion, Conscientiousness or Neuroticism would be more apparent as we observe an individual at first sight. Besides the studies based on facial appearance and head pose, previous evidence suggests using audio-visual recordings to improve outcomes of automatic personality analysis [18,39,72]. Effect of the Big Five personality traits on emotions has also been studied [90], however, there is limited to none research that investigates possible models for finding implications of these traits. Therefore, in this study, the optimal solution is to design a computational model that can accurately identify implications of certain traits and use multimodal cues of these traits in further annotations to validate each other.

Earlier studies have repeatedly demonstrated that the personality traits affect clinical features, prognosis and treatment response of certain mental disorders such as depression [49] and personality disorders [69]. Even though evaluation of personality traits holds high potential to be effectively used in clinical settings for the management of certain disorders, this is hampered by certain aspects of current evaluation methods like requirement of specific training for application and interpretation, employment of extra personnel, and high time expenditure [15]. Therefore, automated reliable computerized methods for personality trait assessment could potentially overcome such limitations and increase their utilization, enabling better management of mental disorders. Despite its potential benefits in clinical practice, personality assessment is a phenomenon of daily routine for everybody. Associated topics getting growing attention are first impression analysis and hiring recommendation systems [26,72]. Willis and Todorov found that 100 milliseconds would be enough to get a trait impression about someone, and this immediate impression is correlated with important decisions [86]. But one should keep in mind that the evidence for the validity of these first impressions is still unsettled, leading us to develop complex methods for personality trait analysis.

Table 1 Associated Adjectives for the Big Five Personality Traits traits

Factor	Adjectives
Agreeableness (AGR)	Appreciative, forgiving, generous, kind, sympathetic
Conscientiousness (CON)	Efficient, organized, planful Reliable, responsible, thorough
Extraversion (EXT)	Active, assertive, energetic, Enthusiastic, outgoing, talkative
Neuroticism (NEU)	Anxious, self-pitying, tense, touchy, unstable, worrying
Openness to Experience (OPE)	Artistic, curious, imaginative, Insightful, original, wide interests

In this study, as well as using an existing dataset, we have collected a new one. We develop methods employing deep architectures to analyze audio-visual cues in the videos also with the help of the transcribed speech. The contributions of this study can be listed as follows:

- A new audio-visual dataset for personality analysis is collected, including 60 subjects. The collected dataset consists self-presentation (speech) videos in an interview-like setting as well as videos of trait-based induced behavior (obtained using video stimuli). While a few studies exist which aim to induce different levels of affect (valence and arousal) and emotions to correlate them with personality traits [1,61], we aim to explicitly induce behavioral patterns related to personality traits which is the first of its kind to the best of our knowledge.
- We present deep spatiotemporal models for the estimation of personality traits from multimodal cues.
- We systematically assess the reliability of several behavioral modalities for personality analysis.

– Different fusion techniques are implemented and evaluated in a detailed manner.
– Our results suggest that the (trait-based) induced behavior includes signs/cues of personality.
– We analyze the relation between self-reported and observed (by experts) personality traits.
– We investigate and visually compare the (facial) patterns displayed in different datasets in two dimensional feature spaces.

2 Related work

Analysis of personality traits is an important task for various applications from evaluating job candidates to providing personality-aware recommendations. In addition, objective assessment of personality is crucial for the assessment of several mental disorders. In this section, we overview the literature on automated analysis of apparent personality, and discuss the importance of personality analysis from a medical point of view.

2.1 Analysis of apparent personality traits

Personality trait analysis has been a common area of interest for psychologists and psychiatrists for many years. With the advances in deep learning, use of computational methods to assess apparent personality traits has started to attract more interest from computer scientists. Recent studies mainly utilize three different modalities: vision-based, focusing on images [37,66,89], and focusing on videos [6,10,17], audio-based [77], and language-based [4]. Vision-based methods are generally utilized more commonly than any other modalities [72]. To provide more robust predictions, while some studies only combine audio and visual information [11,36,76], many others propose using additional modalities, namely, combining language information with audio-visual features [3,28,35] or using facial landmark locations and action units (AU) [78,85].

[37,66] employ image-based analysis to estimate personality trait scores where authors explore the use of selfies (self-portrait images) and show that selfies contain behavioral cues that help assessing apparent personality traits. [23,70] use data collected from Instagram and utilize deep Convolutional Neural Networks (CNNs). While [70] proposes to use Instagram images that users liked, to predict their personality traits, [23] builds a combined image- and language-based method that estimates personality from users' Instagram posts utilizing images and the corresponding captions.

A variety of methods are explored to model personality traits from videos. While earlier studies use handcrafted features, recent ones focus on deep learned representations. For instance, [10,11,85] use Weighted Motion Energy Images to capture the overall motion of the person in the video to predict personality trait scores. [25] employs texture descriptors, namely Local Phase Quantization (LPQ) and Binarized Statistical Image Features (BSIF). [35,36,47,76,83], on the other hand, utilize deep learning models to obtain latent representation of videos. [35,36] use a variation of the ResNet-18 architecture [43] to extract features from each frame of the videos, whereas [76] uses 3D convolutions. [83] proposes an extension to the Descriptor Aggregation Networks that utilize max and average pooling at two different layers of the CNN and normalize these values. Outputs of these pooling layers are concatenated before they are fed to fully connected layers that perform multi-target regression. [47] proposes combined use of deep facial and deep scene features. They extract facial features using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [5] and the pretrained VGG-Face [63] model. For modeling the scene, a pretrained object detection model is used.

In terms of modeling audio, [35,36,83] use deep architectures to learn effective representations of audio signals for personality analysis. [83] employs the Mel Frequency Cepstral Coefficients (MFCC) and learned representations obtained from raw waveform signals using a combined CNN and Long Short-Term Memory (LSTM) architecture. [35,36] modify the ResNet-18 architecture [43] to model mel power spectrograms of audio. [47], on the other hand, employs various handcrafted features such as MFCC, pitch, energy and their temporal derivatives, in the analysis.

When it comes to language processing, [85] makes use of the transcriptions of speech. They compute several readability indexes and combine these with the total word count and the number of unique words used in the transcripts to obtain two additional statistical features. Differently, [35] uses two different word embedding models, namely, the bag-of-words embedding and the skip-thought vector embedding followed by a fully connected layer to describe language patterns.

In multimodal analysis, it is important to choose a reliable method to combine different modalities for more accurate estimation of trait scores. To this end, various models are presented in the literature, e.g., [36,76] explore feature level (*early*) fusion after extracting features from deep models and [36] combine audio- and vision-based features by concatenation after a temporal average pooling. In [35], a ridge regressor is used at the output level to combine audio-visual and language modalities. [76] presents two different architectures for fusion, namely a fully connected network and an LSTM network, both follow concatenation of audio- and vision-based representations. Whereas [38,39,83,85] investigate score level (*late*) fusion by computing a weighted average of the trait scores predicted across different modalities to make a final prediction. [85] combines facial appearance and language models' predictions, [38,83] combine audio and visual models' predictions, and [39] combines facial appearance, scene and audio models' predictions. Differently, Kaya *et al.* [47] propose a *hybrid* fusion method. They first perform an early fusion by aggregating the similar features into two groups (by concatenation). Using a separate model on each feature group, two prediction scores are obtained. In late fusion, these predictions are fed to a random forest regressor for the final prediction.

In the current study, we analyze several modalities such as facial appearance, action units, head pose & gaze, body pose, voice, and transcribed speech, and investigate the effectiveness of different fusion strategies in a systematic manner. For further studies and details on personality analysis, we refer the reader to [80].

2.2 Medical perspective on personality traits

Psychiatrists and clinical psychologists have extensively studied personality traits for their role in the general diagnosis and prognosis of psychiatric disorders. However, the hardship of evaluation of these facets fundamental to the human personality, and behavior have hampered their sustained utility in everyday clinical practice.

The most common psychiatric disorder in the clinical sense, depression has a high lifetime prevalence with substantial disability and burden for society [14]. While evaluating the etiopathogenesis of depression, functional imaging has provided evidence for the interaction of depression diagnosis, personality traits, and brain state [87]. Moreover, personality traits affect treatment outcome; especially high levels of Neuroticism, producing a negative impact on remission [60]. Whereas, traits of Extraversion and Openness to experience were observed to be higher in responders [67].

Another affective disorder, while not as prevalent as depression, is bipolar affective disorder, namely manic depression. Previous studies have identified a state (i.e., manic or depressive state of the disorder) independent association between bipolar disorder and personality traits [51]. Similar to depression, personality traits could also provide guidance on treatment outcomes as well as prognosis of the disease, even displaying an association with switches in the disorder states [50].

Personality traits also have a close relationship with psychotic and anxiety disorders. Compared to healthy controls, patients with schizophrenia have a higher Neuroticism level, and lower Extraversion, Openness, Agreeableness and Conscientiousness levels [62]. Previous studies show that, Big Five traits are associated with social functioning [57], life satisfaction [12], and non-adherence to treatment and treatment delay [20,56] in patients with schizophrenia. From a genetic perspective, a higher level of Neuroticism increases the familial risk of psychosis [13] and the Big Five traits and schizophrenia share some common genetic loci [73].

Personality traits seem to have a role in the development of anxiety disorders. Recent studies have found that a high level of Neuroticism predicts social phobia, panic disorder and generalized anxiety disorder [82]. Higher Neuroticism and lower Conscientiousness levels are related to an increase in anxiety and lower Extraversion is associated with social phobia [52].

Personality traits have a key role for both diagnosis and prognosis of four major mental disorders. Although the significance of the influence of personality traits over treatment outcome on the individual patient level is debatable, the value of personality trait assessment in clinical trials is unobjectionable, if could be done effortlessly [45]. As they rely on the individuals subjective views on their own personality, the non-objective nature of the current assessment methodologies hinder their use on the individual level to guide clinical practice. Thus, future studies need to merge self-reported results with an observer's reports, and use momentary behavioral signs for the assessment of personality traits. Moreover, using mobile applications have received a growing attention in management of mental disorders [19]. Thereupon, any such accurate automated systems of assessment would be indispensable.

3 Methodology

To model and estimate the level of (observed) personality traits, we employ several methods on various modalities including facial appearance, action units, head pose & gaze, body pose, voice, and transcribed speech. Information obtained from individual modalities are then fused employing different strategies. Observed scores for each trait (normalized to [0, 1] range) are used as labels. Details of modeling each of these modalities and fusion strategies will be described in the following sections.

3.1 Facial appearance

3.1.1 Face normalization

As the first step of analyzing facial appearance, we track 68 landmarks on the facial boundary (17 points), eyes & eyebrows (22 points), nose (9 points), and mouth (20 points) regions in the videos using a state-of-the-art tracker , namely OpenFace [8] (see Fig. 1a). Once the landmarks are obtained, the facial image in each frame of the videos are normalized in terms of translation, rotation and scale to obtain frontal view of the faces.

The tracked 2D coordinates of the landmarks are first normalized by removing the global rigid transformations such as translation, rotation and scale. To shape-normalize facial texture, each face image is warped using piecewise linear warping so as to transform the X and Y coordinates of the detected landmarks onto those of normalized landmarks. Obtained images are then scaled and cropped around the facial boundary and eyebrows as shown in Fig. 1b. As a result, each normalized face has a resolution of 224 × 224 pixels. Note that the deformations in the facial surface can better be interpreted since the normalized faces are directly comparable in a pixel-to-pixel manner.

3.1.2 Modeling

Once the normalized facial videos are obtained, we model the spatio-temporal patterns using two different deep architectures, namely by the 3D ResNext-101 [88] and by a Convolutional Neural Network, Gated Recurrent Unit combination (CNN-GRU).

Since our input is a facial video, our aim is to capture both facial appearance and facial dynamics. To this end, we opt for employing a CNN-based architecture that also takes into account the dynamics between the frames through spatio-temporal kernels. Therefore, we first use **3D-ResNext** model to utilize temporality thoroughly. The novelty of the ResNeXt architecture [88] is the introduction of the *cardinality* concept, which is a different dimension from deeper and wider. ResNeXt block introduces group convolutions (whose numbers are called cardinality), which divide the feature maps into small groups different from the original ResNet [43] bottleneck block. Xie *et al.*[88] shows that increasing the cardinality of 2D architectures is more effective than using wider or deeper architectures.

To model normalized facial videos, we fine-tune 3D ResNext-101 [40] that is pretrained on the Kinetics dataset [46], starting from the third block (based on our preliminary experiments). We use random temporal sampling of 45 frames (RTS-45), which corresponds to 1.5 seconds, during training, and non-overlapping sliding window of the same size during test and validation. Window size is chosen among the values [30, 45, 60] through validation error. Finally, the last fully connected layer of the network is replaced with a linear regression layer and L1 loss is utilized. Notice that the latent representation fed to regression layer is 2048D.

CNN-GRU is employed as a second spatio-temporal deep architecture for modeling facial videos. It is widely used [27,81] in the literature, as it can model the spatial relations via CNN and temporal relations via the recurrent network at the same time. In our implementation, as shown in Fig. 2, AlexNet is used as the CNN module by connecting its FC7 layer to a two-layered GRU structure, where the dimensionality of both GRU layers are set to 512. In this way, 4096D spatial representation of faces is fed to the temporal model. As the final layer linear regression with L1 loss is employed. The obtained model is trained in an end-to-end manner. We initiate the training from the pretrained weights of the original AlexNet, in order to accelerate the process and start from an effective set of parameters.

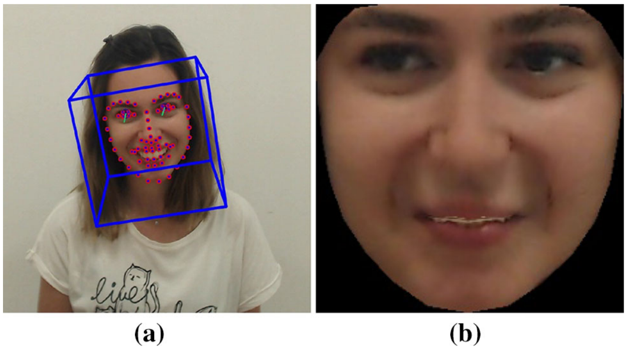

Fig. 1 **a** Visualization of the facial landmarks, gaze direction, and head pose obtained from OpenFace, and **b** the corresponding normalized face

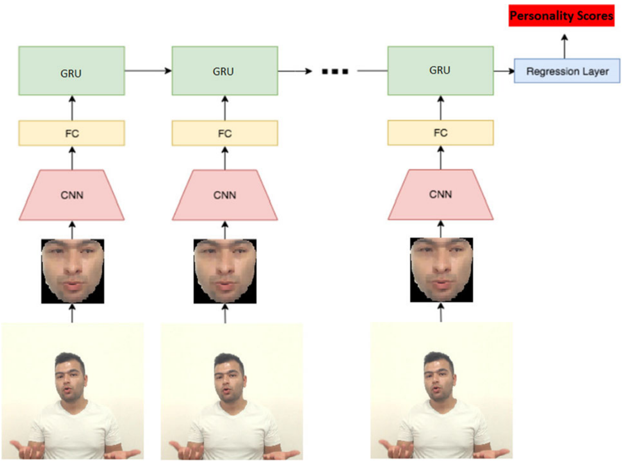

Fig. 2 CNN-GRU architecture, followed by a regression layer

During training, average mean absolute error of the five traits is minimized for both 3D-ResNext and CNN-GRU models.

3.2 Facial action units and head pose & gaze

3.2.1 Feature extraction

To obtain measures for facial shape, displayed facial action units (AU), head pose, and gaze, we process the videos using OpenFace [8] as visualized in Fig. 1a. In order to describe facial action units, we use the 18 AU occurrence and 17 AU intensity features provided by OpenFace. While the binary occurrence features indicate the presence of AU1, AU2, AU4, AU5, AU6, AU7, AU9, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU28, and AU45, the intensity features (in the range of [0,5]) represent the intensity of the aforementioned AUs except AU28 (lip sucking).

To represent head pose, 3 degrees of out-of-plane rigid head rotations (i.e., pitch, yaw, and roll) in radians, and the

3D location of the head with respect to camera in millimeters are used. Finally, for describing the gaze, we employ the 3D gaze directions for both left and right eyes (yielding six feature values) together with the 2D coordinates of 28 eye landmarks for each eye (yielding 112 feature values). Then head pose and gaze features are concatenated, to be used as the representation of the head pose & gaze.

Obtained frame-level feature vector of each of these modalities can be used as a time step in temporal models. In other words, each video can be represented by the multivariate time series of the aforementioned modality-specific feature vectors.

3.2.2 Modeling

The action unit and head pose & gaze features are modeled using two different models such as Long- and Short-term Time-series Network (LSTNet) [54] and Recurrent Convolutional Neural Networks (RCNN) [55], which combines the benefits of Long Short-Term Memory (LSTM) with CNN. Average mean absolute error of the five traits is minimized to train the models.

LSTNet model used in this study is a modified version of the original architecture [54]. We opt for LSTNet since the literature indicates that it achieves significantly better performance than various other time series models [54]. LSTNet extracts short term patterns and local dependencies via convolution through temporal dimension. Output of the convolution layer is fed to the recurrent layer and the recurrent-skip layer. In recurrent and recurrent-skip layers, Gated Recurrent Unit (GRU) is used. Normally, GRU fails to capture very long-term dependencies due to gradient vanishing. Recurrent-skip layer captures long-term and periodical information by processing the sequence with N skips, where a recurrent layer processes consecutive inputs with 1 skip-length. Output of recurrent and recurrent-skip layers are then concatenated and fed to a linear layer. Skip-length parameter is set to the number of frames per second, which is 30. Dropout with a rate of 0.2 is applied after convolution, recurrent, and recurrent-skip layers. Hidden dimension of the convolution and recurrent layers are set to 100. There are 5 different recurrent-skip components employed, therefore, a 150D vector is obtained via their concatenation. At the penultimate layer, we concatenate this vector with the last hidden state of the recurrent layer and obtain a 250D feature vector. In contrast to [54], we do not use the autoregressive component of LSTNet. We also do not use the tanh activation function at the output since our target problem is regression. We train the network through optimizing the L1 loss via Adam optimizer with a learning rate of 0.001.

RCNN has been proposed in [55] for text classification. It uses Bidirectional Long Short Term Memory (BiLSTM) networks followed by max-pooling through temporal dimension. The output of the max-pooling layer is fed to the linear layer. Our RCNN's recurrent module consists of two BiLSTM layers. We set the dimension of all hidden layers of both backward and forward LSTMs as 256. Hidden dimension of linear output layer is set to 64.

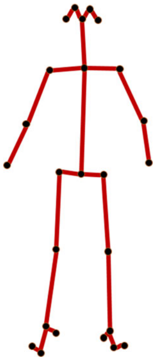

Fig. 3 The tracked body landmarks by OpenPose

3.3 Body pose

Input videos are first processed using OpenPose [16] to track 25 landmarks on the joints (e.g., wrist and elbow), neck, and face as shown in Fig. 3. 2D coordinates of the tracked landmarks are used as posture features to represent the general pose and structure of subjects' body, e.g. how they sit and move while answering questions and watching videos. Note that apart from other visual modalities, this is the only one where we focus not on the face, but the body/posture of the participant. 50-dimensional body features are then modeled with LSTNet as described in Sect. 3.2.2 so as to minimize the average mean absolute error of the five traits.

3.4 Voice

To represent the characteristics of voice, we compute a 34-dimensional feature vector from audio data of videos, including MFCC, Chroma vector, energy and entropy related features using pyAudioAnalysis framework [31]. Voice features are extracted for each 50 milliseconds of videos with 50% overlap, i.e. with a step size of 25 milliseconds. Consequently, these features form the multivariate time series for describing the voice. Details of the feature extraction process can be found in [31]. Obtained features are modeled by LSTNet architecture (as described in Sect. 3.2.2) by minimizing the average mean absolute error of the five traits.

3.5 Transcribed speech

As reported in the literature [35], use of language as an additional modality enhances the estimation reliability of personality traits. In order to model language-based cues for

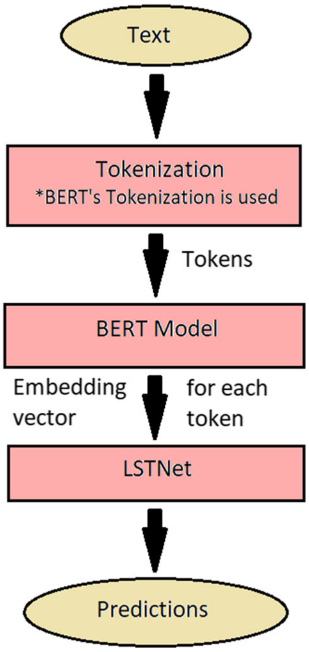

Fig. 4 Flow of modeling the transcribed speech

personality analysis, we first transcribe the subjects' speech in videos using Google's Speech to Text API [30]. Since there may be more than one language spoken in the dataset, language is automatically detected by Google's API. To make our model generalizable, an embedding obtained by a large corpora is used. We employ pytorch-transformers' implementation [44] of pretrained multilingual BERT model [24] to generate embeddings for each token in the transcripts.

BERT model is applied to the whole transcript of each video, separately. BERT model infers the embeddings of each word in the transcript considering its context. Flow of our transcribed speech model is shown in Fig. 4. BERT embeddings are then modeled by LSTNet for the multitask regression task (for details please see Sect. 3.2.2). Similar to the modeling of aforementioned modalities, average mean absolute error of the five traits is minimized during training.

3.6 Fusion

To combine the information obtained from different modalities, we employ different strategies, namely, *early fusion*, *hybrid fusion*, and *late fusion*. By early fusion, it is meant to combine extracted features from different modalities. Late fusion indicates fusing predicted (regression) scores obtained using individual modalities. Finally, hybrid fusion is meant to use extracted features/scores from different layers (including regression layer) of models of individual modalities, for fusion. As for the early fusion strategy, we use three different methods, namely, concatenation, modality attention, and feature attention. Linear Support Vector Regressor (LSVR) is employed as a late fusion strategy. Finally, under the hybrid fusion strategy, CentralNet [79] is used.

In early fusion, we extract features from the penultimate layer of each modality. To combine these features, we employ either concatenation or attention-based fusion. For the concatenation case, obtained feature vector is fed to a Multilayer Perceptron (MLP) with 2-hidden layers to predict regression scores.

In the attentional fusion case, we employ a separate MLP model with 2-hidden layers for each modality to obtain features that have the same dimensionality for all modalities. The only difference of these models for different modalities is the number of neurons in their first hidden layers which is due to the various feature dimensions of different modalities. Then, the obtained representations are fed to an attention layer that assigns weights to each feature. Finally, weighted representation is fed to a linear regression layer to obtain regression predictions. These models are trained in an end-to-end manner using L1-loss. Let h_i denote the hidden representation of ith modality and α_i denote its attention weight. Then, the weighted representation c can be computed as:

$$c = \sum_{i=1}^{K} \alpha_i h_i \ , \tag{1}$$

$$a_i = \frac{exp(\sigma(Wh_i + b))}{\sum_{k=1}^{K} exp(\sigma(Wh_k + b))} \ , \tag{2}$$

where W and b denote weight and bias of the attention layer, respectively. $\sigma(.)$ is the sigmoid function, K is the number of modalities and α_i is a scalar. This mechanism will be referred to as modality attention.

For the feature attention method, the weighted representation c becomes:

$$c = [c^1, c^2, \ldots, c^D]^\top \text{ , where} \tag{3}$$

$$c^d = \sum_{i=1}^{K} \alpha_i^d h_i^d \ . \tag{4}$$

In Eqs. 3 and 4, D denotes the feature dimension, c^d is a scalar for the dth dimension of c, and α_i is a vector. Note that Eq. 2 does not change except the dimensions of W and b. In other words, attention is applied to feature dimensions separately in feature attention, as opposed to Eq. 1, where attention is applied to all features of the corresponding modality.

In late fusion, we first concatenate the predicted scores obtained from different modalities. These score vectors are then modeled by an LSVR. A separate LSVR model is employed for each of the five traits.

For hybrid fusion, we utilize CentralNet [79]. To this end, representation obtained for each modality is fed to a separate 2-hidden layer MLP model with batch normalization, dropout and ReLU activation function between the layers. Then the computed $1st$ hidden layer representations of these MLP models for different modalities, are fused as in modality attention mechanism, forming the central joint representation. The central joint representation (for the $1st$ hidden layer) is then fused together with the $2nd$ hidden layer representations of all modalities in a similar manner. The obtained representation is fed to a regressor. All the models are jointly optimized in an end-to-end manner.

For the fusion methods that employ MLP models, we use ReLU activation function and dropout. Considered hyperparameters of MLP models are given in Table 2. Note that the considered set of values for these hyperparameters are dynamically determined in the given intervals per fusion method based on minimum validation error. The resolution of values is increased when a performance improvement is observed. For LSVR, we use the default hyperparameters (regularization parameter $C = 1.0$ and stopping tolerance is 0.0001).

4 Datasets

In our experiments, we employ two datasets, namely Self-presentation and Induced Behavior Archive for Personality Analysis, a new personality dataset that has been collected during this study, and the ChaLearn LAP First Impressions Dataset [65]. Below, these datasets will be described in detail.

4.1 Self-presentation and induced behavior archive for personality analysis

One of the goals of this study is to investigate whether the personality traits can be estimated from induced audio-visual behavioral characteristics. To this end, we have video-recorded participants while they watch a set of videos, where each video clip has been chosen to be associated with one of the Big Five personality traits. In addition, we have video-recorded their answers to three questions. Self-presentation and Induced Behavior Archive for Personality Analysis (SIAP) includes recordings of 60 participants (37 females, 23 males) from 5 countries. Ages of the participants vary between 18-35 years. The dataset includes self-assessed and observed scores for each the Big Five traits.

To minimize the differences between sessions so as to obtain similar experience for different participants, we have developed and used a computer software rather than employing an interviewer during the data collection. Before beginning the data collection, each participant has been informed of the experimental protocol and the use of our software, and signed a consent form. After that the whole experiment and the data acquisition have been conducted automatically. Following sections will provide further details of SIAP.

4.1.1 Data acquisition

The software allows participants to choose their preferred language, either English or Turkish. Once a participant choose his/her preferred language, the software show three videos. In each video, a psychologist ask a question (in the preferred language). The first question asks demographic information of the participant. The second question is about an experience of the participant while having an activity last time which he/she likes. In this way, the participant could specifically talk about a memory without thinking much on a certain one. The last question is about a time that the participant had solved a problem with his/her close other and how they managed it. After watching each question video, the participant is given 60 seconds to answer the corresponding question, with a 15 seconds countdown at the end in order to remind the remaining time. Then the video for the following question starts playing.

Once the participant completes answering the questions, he/she moves on to the second part of the experiment. In this part, to induce behavioral cues of personality traits, the software shows a set of video clips for approximately 15 minutes in total, including three videos for each of the five traits (15 videos in total). Duration of the videos varies approximately 30 to 60 seconds to obtain a proper response from the participant. Notice that a (separate) large set of video clips have been chosen (by a consensus of three psychologists) in order to induce each of the five personality traits, and the software randomly chooses three videos from the corresponding set of videos for each participant.

The software records participants via three cameras. A Logitech C920 webcam is used as a frontal camera to record the facial expressions (which is attached to the monitor). Two wall mount cameras record the participants from right-front and left-rear sides (with respect to the participant) to obtain pose and gesture information. The recording setup and sample frames captured by these three cameras can be seen in Figs. 5 and 6, respectively. While frontal videos have been captured with a resolution of 1920 × 1080 pixels at 30 frames per second, side-view videos have a resolution of 1280 × 720 pixels at 25 frames per second. Audio has been recorded with a sampling rate of 44100 Hz.

In the final stage of the experiment, the participants are asked to complete two ten-item questionnaires on seven-point scale, namely, Ten Item Personality Inventory (TIPI) [33,41] and a questionnaire on close relationships ("Experiences in Close Relationships Inventory-Revised") [71]. All the aforementioned steps of the experiment is handled by our software through an easy to use graphical user interface

Table 2 List of Considered Hyperparameters of MLP

Hyperparameter	Considered Interval
Number of units in the 1st hidden layer	[64, 2048]
Number of units in the 2nd hidden layer	[64, 512]
Initial learning rate	[1e-5, 1e-2]
Weight decay	[1e-4, 1e-2]
Batch size	[64, 256]
Dropout rate	{0.5, 0.6, 0.7}

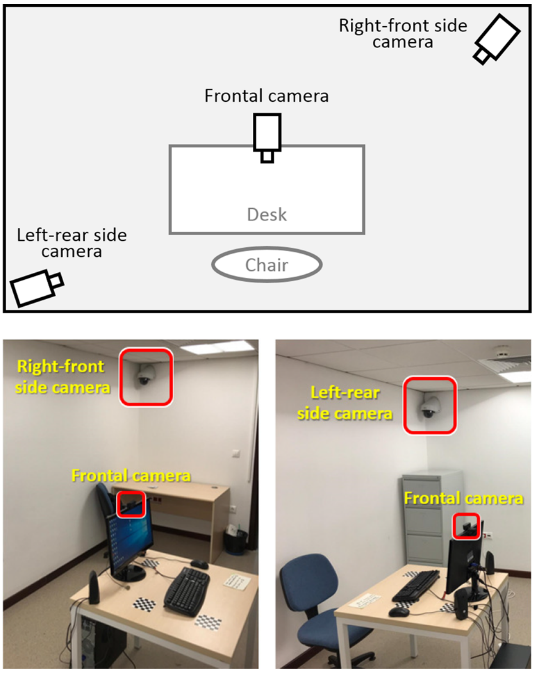

Fig. 5 Recording setup. Positioning of the right-front side, left-rear side, and frontal cameras

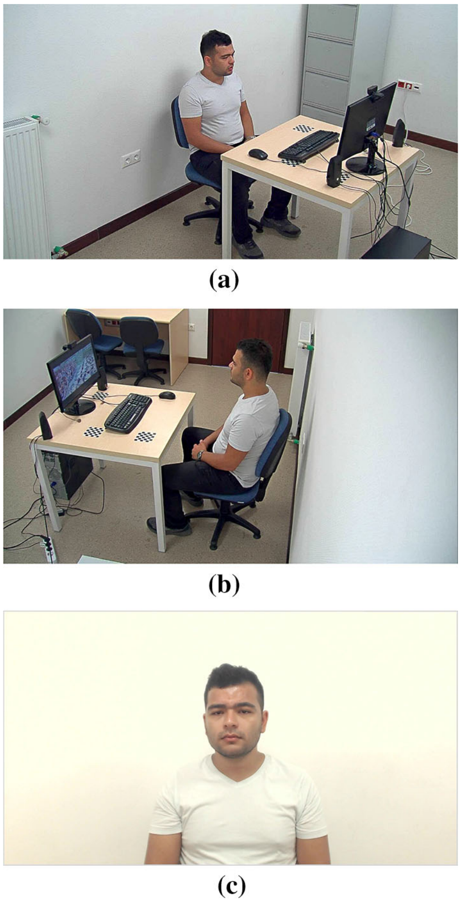

Fig. 6 Sample frames obtained from **a** the right-front side, **b** left-rear side, and **c** frontal cameras

where the only focus is the monitor during the experiment, free of any other distractions.

4.1.2 Choosing video clips to elicit behavioral cues

Picking the correct video that would tap into a specific personality trait is crucial, which might provoke a behavioral mechanism we may catch on. To this end, we investigate the traits and their prominent properties. Many descriptions of Big Five personality traits were used in defining which type of video clips we could pick.

For Openness, we have chosen to focus on the curiosity and intellect aspect of this trait, since they would be easier to express. Recent studies suggest that Openness is related to being a multicultural person, who tends to oversee the racial and ethnic differences of other people [74]. Therefore, we have decided to pick videos that includes activities, where participants presumably would not encounter on daily basis. Such videos would clearly display different cultures or religions (e.g. extreme sports, praying and rituals of different religions).

For Conscientiousness, we have picked videos that would show people’s differences on academic performances and being diligent in each and every manner [42] (e.g. two students preparing for their exam: hardworking versus lazy).

For Extraversion, we have deduced that individuals who have less fun interacting with others, would also be the ones that are introverts. Since individuals who have higher levels of Extraversion tend to be more expressive, we assume that getting a reaction/response based on approval would be easier [68]. So, we have used clips that involve high physical stimulation in other people's presence (e.g. dance parties, performing to a crowd).

For Agreeableness, we have chosen to focus on the social aspect on this trait. Therefore, we have decided to pick videos that would show actions which emphasize social harmony, compassion and empathy with others [34]. Thus, we have used videos that include interpersonal encounters that is either harmonious with others or not (e.g. apologizing, disagreeing to anything without any logical basis).

For Neuroticism, we have employed video clips that would look and make individuals feel like something bad is going to happen, inducing the disturbed thought processes of participants without actually making them feel that way. Therefore, if the corresponding participant is high on Neuroticism continuum, he/she would expect something bad would happen more than others (e.g. a house burning, glass falling down from a table without a reason). In this sense, we expect people to behave in a certain way when viewing these videos. Yet, [68] suggests that people that are high on Neuroticism tend to be less expressive on their affect. So, it might be difficult to catch those expressions if the corresponding participant has high scores on Neuroticism. As explained before, an introvert might be overwhelmed by over stimulation factors contained in the videos, whereas an extrovert might express positive affect with showing behaviors of blending in (e.g. bopping head to an upbeat song). We especially picked some scenes from movies and real life events for Neuroticism that might trigger one's negative affect to further provoke disturbed thoughts.

4.1.3 Annotation

Obtained video recordings of participants were evaluated by three different psychologists in terms of personality traits. The psychologists (after their individual evaluations) discussed each personality trait of each participant until they achieved a 100% consensus on the final score. Participants' postures and facial expressions while watching the aforementioned video clips (e.g. their reactions/responses) were accounted for annotating the personality traits. Trait levels of the participants were annotated based on the relation/correlation between their responses and the target trait of the corresponding video. Seven-point scale was used for the annotation of the scores (1: very low; 7: very high).

For Openness trait, smiling and engagement with the video (more saccadic eye-movement without negative viewing) were pursued in individuals who are high in Openness. On the contrary, disgust-like facial responses were treated as low Openness. For Conscientiousness, high scores were given if the participants become disturbed after individuals being disorganized in their environments. If the participant shows engagement with the opposite type of behavior, he/she was rated low on Conscientiousness. For Extraversion, we again looked for engagement with the extrovert behaviors in the videos. Other than this, we also looked for tapping of foot or swinging with the music when giving high scores. As opposed to these reactions, participants who gave disgust-like reactions or show discomfort were given low rating on Extraversion. For Agreeableness, we have expected individuals with high Agreeableness to remain calm while are shown an individual who is low on this trait. Others, who were showing discomfort, were rated low on this trait. Finally, for Neuroticism, individuals who were watching the video clips with significant amount of discomfort (e.g. squinting eyes and leaning back) even though scenes did not show any discomforting image, were rated high on Neuroticism.

If a participant did not show any signs of being in any given side of the spectrum of reactions while watching the (inducing) video clips, score for him/her was rated as 4 (neutral). If a participant showed any leaning to one side of the spectrum slightly, score for him/her was rated as 3 or 5 accordingly. If the participant showed considerable reaction (e.g. clearly displaying a certain response to the video), we rated his/her score as 2 or 6. Any extreme case of these spectra of behaviors was rated as 1 or 7, accordingly.

4.1.4 Data partitions

As described above, SIAP has recordings for durations of speaking (question answering) and for durations of watching video clips to induce cues of personality traits. These partitions of our dataset will be referred to as SIAP-Interview and SIAP-Induction, respectively, in the remainder of the paper. The interview partition includes 180 sessions (60 participants × 3 questions), and the induction partition has 180 sessions for the induction of each trait, yielding 900 sessions in total (60 participants × 3 inducing videos for each trait × 5 traits). Notice that there are three synchronized videos, namely one frontal, and two side views, for each session.

4.2 ChaLearn LAP first impressions dataset

ChaLearn LAP First Impressions Dataset (FID) [65] contains 10,000 video clips, split to training (6,000 clips), validation (2,000 clips) and test (2,000 clips) subsets. Subjects in the videos are looking at the camera and speaking in English, with varying environmental conditions. These clips have been extracted from over 3,000 different YouTube videos and labeled via Mechanical Turk, where the annotators rate the personality scores of each subject in terms of Big Five personality traits, according to their movements, gestures, voice and appearance.

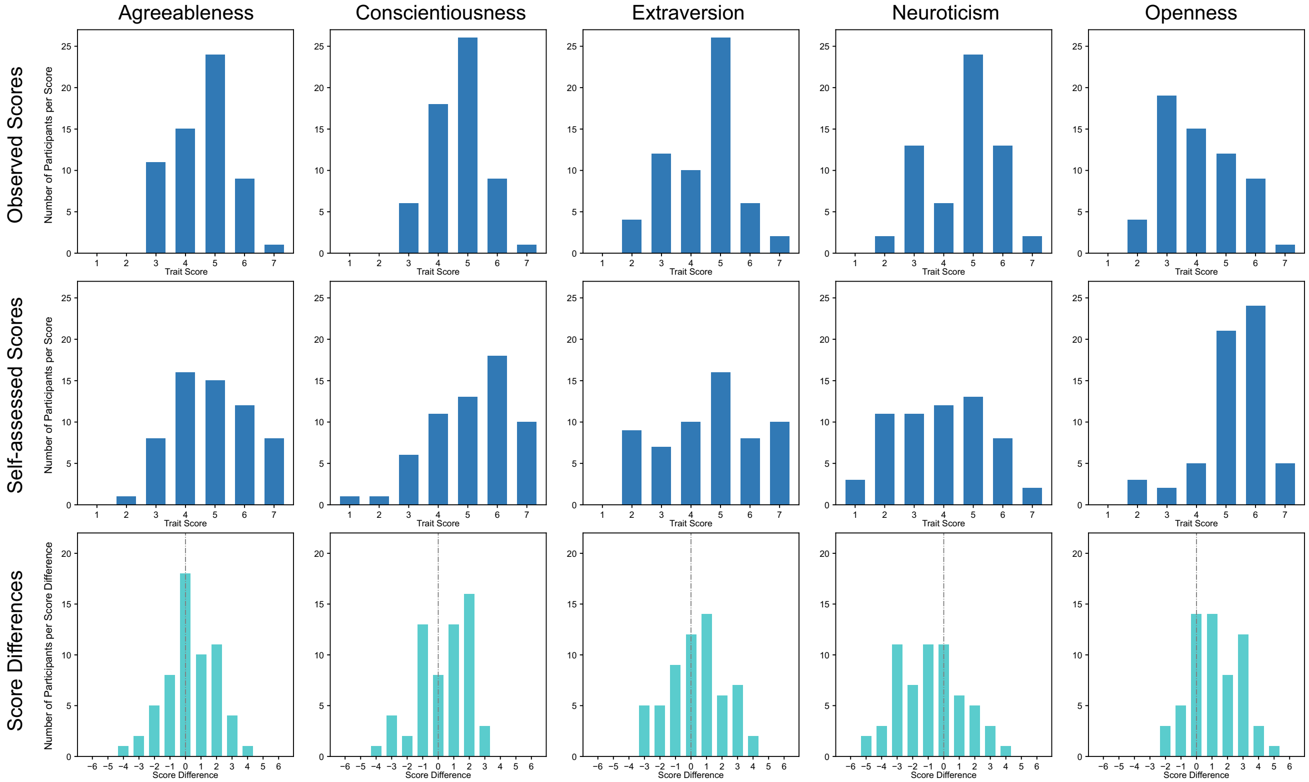

Fig. 7 Histograms of the observed and self-assessed scores, and their differences per participant ("self-assessed score" − "observed score") for the personality traits of 60 participants in SIAP. Positive values for score differences indicate overestimation of the trait by participants compared to the observation, and vice versa

5 Experimental results with discussions

5.1 Experimental setup

In our experiments, two datasets are used, i.e., SIAP and FID. For the experiments on SIAP, a 10-fold cross-validation scheme is used with randomly selected folds (each having 6 participants) ensuring that each subject appears only in one fold. In this way, we guarantee that there is no subject overlap between the training, validation and test sets.

At each fold, while eight parts are used for training, one part is used for validation, and the remaining part is employed for the test. Average of the 10-fold cross-validation results is taken to provide a single result for each experiment. Observed (expert) scores are used for modeling in our experiments on SIAP. In our experiments on FID, we use the predefined training, validation and test sets of FID. Hyperparameters of the models are optimized on the validation set. Test results are reported in terms of mean absolute error (MAE). Notice that seven-point scale scores for each trait is normalized to the range of [0, 1] in our experiments. Therefore, presented MAEs are also in the range of [0, 1].

5.2 Observed versus self-assessed personality scores

One of our goals is to analyze/reveal the relations between/within the observed (expert-annotated) and self-assessed personality scores. To this end, we first analyze the correlations within scores of different traits. Based on the results, only one moderate correlation is found, which is between Extraversion and Openness ($r = 0.57$) in observed scores. Next, the distributions of the observed and self-assessed scores are computed for 60 participants in SIAP , along with their differences per participant ("self-assessed score" − "observed score"), as shown in Fig. 7. Mean and variance of scores for each trait are also computed and reported in Table 3. As the results suggest, subjects tend to rate the traits, which are more socially desirable (e.g. traits that are perceived positive), higher than the observed scores. On the other hand, participants assess traits, which are less socially desirable, lower than observers do. This social desirability effect can especially be seen in the Openness and the Neuroticism. Considering the mean scores, participants overrate their Openness by 39.7% (relative) and underrate their Neuroticism by 21.2% (relative), compared to the observed scores, as reported in Table 3. This finding can be observed in the last row of Fig. 7 as the Openness plot is skewed towards the

Table 3 Mean and Variance of Self-assessed and Observed Scores for the Five Traits in SIAP

	Type	AGR	CON	EXT	NEU	OPE
Mean	Self-assessed	0.643	0.671	0.615	0.479	0.722
	Observed	0.594	0.614	0.567	0.608	0.517
Variance	Self-assessed	1.667	1.837	2.392	2.139	1.206
	Observed	1.012	0.816	1.440	1.494	1.523

Table 4 Inter-rater Reliability Scores between Expert-Annotators across the Big Five Personality Traits

Type	AGR	CON	EXT	NEU	OPE
ICC	0.714	0.754	0.898	0.744	0.872
Krippendorff Alpha	0.712	0.751	0.897	0.742	0.870
Mean Spearman’s Rho	0.728	0.758	0.892	0.737	0.861

positive values, while the Neuroticism plot is skewed towards the negative values.

One-way ANOVA analyses also support our claims in social desirability aspect in self-report scores. There is a significance between annotators’ scores and self-reports in Agreeableness ($F(1, 118) = 44.709, p < .0001$), Openness ($F(1, 118) = 32.887, p < .0001$), and Neuroticism ($F(1, 118) = 9.755' p = .002$). To further investigate social desirability effects, we look at the differences between self-report and observed scores of participants traits across genders by taking absolutes of subtracting participant’s observed score from self-report score. t-tests show a significant difference between genders (p = 0.001) which suggest males (M = 1.91) present themselves more extraverted than females (M = 1.86). Additionally, females report themselves as more agreeable (M = 1.95) than males (M = 1.89) according to t-tests (p = 0.004). It is expected to see gender differences on different traits in terms of social desirability. Previous studies usually show that women score themselves higher on most of the sub traits of that are linked with Conscientiousness and Agreeableness than man. Such as tender-mindedness, dutifulness, self-discipline [22,29]. On the other hand, men show high scores on assertiveness and excitement seeking, which are sub traits that are linked with Extraversion [22,29]. Our results show similar trend of gender differences on Agreeableness and Extraversion compared to previous studies, revealing tendency of high self-evaluation on Agreeableness by women and on Extraversion by men (previous studies reveal women self-evaluate higher on some Extraversion sub traits as well [84]).

Having looked at the social desirability traits, inter-rater reliability scores between the expert-annotators is also crucial for reliability of our results. If the expert-annotators do not agree on which participant scored low or high in a trait, then differences between observed and self-reported scores would be less meaningful. To test this, intraclass correlation coefficients (ICC) [9], Krippendorff alpha coefficients [53] and Spearman’s Rho [75] are computed. According to the results, levels of inter-rater reliability for each trait are found to be consistent and high as reported in Table 4.

5.3 Assessment of different modalities

In this set of experiments, we evaluate the reliability of different modalities such as the facial appearance, facial action units, head pose & gaze, body pose, voice, and transcribed speech on SIAP-Interview and FID for assessing the levels of personality traits. While results on FID are obtained using models that are trained on FID (training set), two set of test results are provided for SIAP-Interview: (i) training on SIAP-Interview, (ii) fine-tuning on SIAP-Interview with an initialization using weights that are learned on FID. Note that the sample size of SIAP-Interview is significantly lower than that of FID. Each session of SIAP includes three videos recorded from different views (see Sect. 4.1). For the evaluation of body pose modality on SIAP-Interview, we use the right-front side videos, yet, in all other experiments on SIAP, frontal videos are employed.

5.3.1 Facial appearance

As described in Sect. 3.1, we employ two different architectures, i.e., 3D-ResNext-101 (3D-ResNext) and CNN-GRU, for modeling facial appearance. Both models are evaluated on FID and SIAP-Interview, and obtained MAE results are given in Table 5. On FID, CNN-GRU architecture provides an average MAE of 0.101, which is 5.2% worse than the visual baseline result (0.096) provided in [35]. On the other hand, 3D-ResNext provides the most promising results on FID among all modalities used with an average MAE of 0.088. Notice that the state-of-the-art method [47] on FID provides a MAE of 0.083. Success of the 3D ResNext on FID may rely on the 3D temporal convolutions and its high regularity with random temporal sampling.

In contrast to the results for FID, the lowest average MAE (0.155) on SIAP-Interview is achieved using pretrained version of CNN-GRU, among facial appearance models. It performs better than 3D-ResNext both with and without pretraining in validation set as well as the test set. Our results for SIAP-Interview indicate that pretraining on FID is useful although the structure of the datasets are different. Notice that FID has recordings extracted mostly from YouTube video blogs, while SIAP-Interview includes recordings of answers to specific questions (self-presentation). Yet, in both setups people, facing a camera, are talking on some topics,

Table 5 MAEs for the Use of Facial Appearance

Dataset	Model	AGR	CON	EXT	NEU	OPE	AVG
FID	3D-ResNext	0.085	0.089	0.088	0.091	0.085	0.088
	CNN-GRU	0.097	0.105	0.101	0.102	0.101	0.101
SIAP-Interview	3D-ResNext	0.169	0.129	0.217	0.181	0.196	0.179
	3D-ResNext*	0.147	0.118	0.176	0.166	0.180	0.157
	CNN-GRU	0.149	0.132	0.183	0.178	0.192	0.167
	CNN-GRU*	0.146	0.120	0.161	0.171	0.175	0.155

Note: * denotes pretraining on FID

Table 6 MAEs for the Use of Facial Action Units and Head Pose & Gaze on FID

Model	Features	AGR	CON	EXT	NEU	OPE	AVG
LSTNet	Facial AU	0.102	0.117	0.106	0.111	0.106	0.108
	Head pose & Gaze	0.106	0.125	0.122	0.123	0.117	0.119
RCNN	Facial AU	0.100	0.113	0.102	0.106	0.102	0.104
	Head pose & Gaze	0.105	0.121	0.120	0.121	0.115	0.116

therefore, similar behavioral patterns are expected. Clearly, obtained MAEs for SIAP-Interview are significantly higher than those of FID. This finding can be explained by the fact that SIAP-Interview is a relatively small dataset. Results may also suggest that during answering questions, facial cues of personality would be less visible compared to expression characteristics displayed in video blogs.

5.3.2 Facial action units and head pose & gaze

In this experiment, we first assess the reliability of using facial action units and head pose & gaze on FID. Mean absolute errors of LSTNet and RCNN models on FID are given in Table 6. Individual use of facial action units provides lower MAE than the individual use of head pose, transcribed speech or voice modalities on FID. On the other hand, using action units could not perform as well as the facial appearance modality on FID. As shown by the results, MAEs for the using action units through LSTNet and RCNN models are 0.108 and 0.104, respectively. Yet, the use of head pose & gaze performs 12% (absolute) worse on average than using facial action units. Better performance of using facial action units compared to that of gaze and head-pose is expected since facial expressions are more capable of displaying mental state and emotion [48]. Consequently, on SIAP-Interview we evaluate only the use of action units with and without a pretraining on FID.

As shown in Table 7, we obtain the lowest MAE (0.152) on SIAP-Interview through LSTNet model with pretraining on FID. Interestingly, MAE of RCNN is increased by 0.2% (absolute), when it is pretrained on FID. On SIAP-Interview, LSTNet provides 1.0% (absolute) MAE improvement compared to the RCNN when the models are pretrained on FID. Accordingly, we use only LSTNet for the further experiments.

Table 7 MAEs for the Use of Facial Action Units on SIAP-Interview

Model	AGR	CON	EXT	NEU	OPE	AVG
LSTNet	0.147	0.134	0.169	0.162	0.169	0.156
LSTNet*	0.139	0.122	0.163	0.157	0.176	0.152
RCNN	0.154	0.132	0.172	0.166	0.177	0.160
RCNN*	0.144	0.128	0.179	0.181	0.178	0.162

Note: * denotes pretraining on FID

Table 8 MAEs for the Use of Body Pose on SIAP-Interview

Model	AGR	CON	EXT	NEU	OPE	AVG
LSTNet	0.171	0.163	0.190	0.218	0.191	0.187

5.3.3 Body pose

For the evaluation of using body pose for estimating personality traits, only SIAP-Interview is employed since the videos in FID do not show the whole body of subjects. To this end, we use the videos recorded from the right-front side in our experiment. Table 8 shows that the body pose provides the worst MAEs among all different modalities on SIAP-Interview. Yet, with a large amount of data and powerful fusion strategies, body pose information would be expected to be useful.

Table 9 MAEs for the Use of Voice

Dataset	Model	AGR	CON	EXT	NEU	OPE	AVG
FID	LSTNet	0.102	0.114	0.110	0.110	0.104	0.108
SIAP-Interview	LSTNet	0.138	0.142	0.177	0.184	0.183	0.165
	LSTNet*	0.160	0.139	0.172	0.170	0.179	0.164

Note: * denotes pretraining on FID

Table 10 MAEs for the Use of Transcribed Speech

Dataset	Model	AGR	CON	EXT	NEU	OPE	AVG
FID	LSTNet	0.103	0.117	0.118	0.118	0.112	0.114
SIAP-Interview	LSTNet	0.150	0.127	0.172	0.167	0.186	0.161
	LSTNet*	0.140	0.131	0.168	0.161	0.172	0.154

Note: * denotes pretraining on FID

5.3.4 Voice

The use of voice for modeling personality traits through LSTNet is evaluated on FID and SIAP-Interview. Table 9 shows the results of the voice modality. On FID, the voice modality provides lower MAEs than the transcribed speech and the body pose. Yet, it performs worse than the facial appearance. On SIAP-Interview, MAEs for the voice modality are lower than the body pose modality. On the other hand, MAEs for voice on SIAP-Interview are significantly higher than the MAE on FID. Lastly, the LSTNet model pretrained on FID performs better for the voice modality compared to the randomly initialized LSTNet model.

5.3.5 Transcribed speech

In this experiment, we assess the discriminative power of transcribed speech for estimating personality traits. For a fair comparison, the speech in both SIAP-Interview and FID videos are automatically transcribed using Google's Speech to Text API [30] as indicated in Sect. 3.5. Next, the extracted transcriptions are used for language processing. Since, SIAP-Interview has recordings in two languages, i.e., English and Turkish, multilingual embedding models are used in our experiment (both for SIAP and FID).

As shown in Table 10, similar to the results of using other modalities, the transcribed speech provides higher MAEs on SIAP-Interview compared to FID. According to our results on FID, the transcribed speech is the worst-performing modality. This may be caused due to our pipeline based approach to model the transcribed speech. Recall that we use the Google's service [30] to transcribe the speech (automatically). On SIAP-Interview, transcribed speech performs better than the body pose and the voice. As expected, pretraining LSTNet on FID improves the accuracy on SIAP-Interview.

5.4 Combined use of modalities

To assess the performance of combined use of modalities, we combine different modalities using five different methods under three fusion categories as described in Sect. 3.6, namely by early fusion (concatenation, modality attention, and feature attention), hybrid fusion (CentralNet), and late fusion (LSVR).

In fusion experiments on SIAP-Interview, we use the models pretrained on FID for each modality, except the body pose, because the pretraining mostly improves the results as shown in Sect. 5.3. Body pose analysis on FID is not applicable since videos display only the upper body. In early fusion experiments that employ the concatenation method, combined uses of all possible combinations of facial appearance, action unit, body pose (only for SIAP-Interview), voice, and transcribed speech modalities are evaluated and the best performing set of modalities in terms of validation accuracy, is selected automatically (see Table 11). In the remainder of fusion experiments, results are obtained using all modalities (experiments on FID does not include body pose modality). To obtain the modality-specific features, the following models are used: 3D-ResNext and CNN-GRU for facial appearance, LSTNet for facial action units (AUs), voice, transcribed speech, and body pose. For all fusion techniques, weights/parameters of the modality-specific models are kept frozen during the training of the fusion models due to computational complexity of joint optimization. Except the late fusion with LSVR, all fusion methods are optimized with Adam. Quadratic optimization is employed for LSVR. Learning rate scheduler with a factor of 0.8 and a patience of 10 epochs is applied for early fusion and hybrid fusion strategies during optimization. Notice that the LSVR is trained on the validation set since the individual modalities have already learned to regress the training data. The results of all fusion experiments are presented in Table 11.

Table 11 MAEs of Different Fusion Methods

Dataset	Fusion Type	Method	AGR	CON	EXT	NEU	OPE	AVG
FID	Early	Concatenation[a]	0.087	0.084	0.084	0.088	0.088	0.086
		Modality Attn.	0.088	0.086	0.086	0.090	0.090	0.088
		Feature Attn.	0.088	0.086	0.086	0.090	0.090	0.088
	Hybrid	CentralNet	0.091	0.086	0.086	0.091	0.090	0.089
	Late	LSVR	0.087	0.082	0.083	0.086	0.087	0.085
SIAP-Interview	Early	Concatenation[b]	0.142	0.126	0.169	0.164	0.180	0.156
		Modality Attn.	0.141	0.135	0.165	0.160	0.174	0.155
		Feature Attn.	0.143	0.123	0.158	0.157	0.183	0.153
	Hybrid	CentralNet	0.148	0.132	0.176	0.194	0.177	0.165
	Late	LSVR	0.154	0.135	0.180	0.190	0.187	0.169

Note: [a] Automatically selected modalities on FID: Facial Appearance (3D-ResNext) + Facial AUs + Transcribed Speech.
[b] Automatically selected modalities on SIAP-Interview: Facial Appearance (CNN-GRU) + Facial AUs + Transcribed Speech

Inspecting the results on FID in Table 11, one cannot observe an enormous difference between the MAEs of different fusion strategies. This may be due to the fact that individual use of facial appearance with 3D-ResNext performs much better than all other modalities on FID. Remember that the second best performing modality on FID, namely facial appearance with CNN-GRU, provides a 12.9% (relative) higher MAE. Utilization of LSVR and concatenation (with modality selection) increase the accuracy of estimations considerably. Best performing fusion strategy on FID is found to be the late fusion using LSVR model. LSVR based fusion reduces the MAE of the best performing individual modality (facial appearance using 3D-ResNext) by 3.4% (relative). This result may be due to the simplicity and effectiveness of focusing solely on the score vectors, rather than learning from the high-dimensional representations of different modalities.

Although feature attention performs better than modality attention for SIAP-Interview, its validation error is higher. Therefore, modality attention should be considered as the winning method in this case. Yet, none of the fusion strategies on SIAP-Interview could reach the performance of solely using facial action units (LSTNet; see Table 7).

Interestingly, hybrid fusion (CentralNet) cannot reach its competitors. While late fusion performs best on FID, early fusion using concatenation (with modality selection) provides the best results on SIAP-Interview. This suggest that the structure of the dataset would easily influence the reliability of fusion methods.

Table 12 Relative Importance Rates (%) for Different Modalities on FID and SIAP-Interview

Target Modality	AGR	CON	EXT	NEU	OPE	AVG
Results on FID						
Face (3D-ResNext)	13.6	23.1	18.0	11.6	12.7	15.7
Face (CNN-GRU)	−1.9	−0.2	−1.7	−1.0	−1.7	−1.3
Facial AU	0.1	0.2	−0.2	−0.3	−0.1	−0.1
Voice	0.8	0.1	0.5	0.6	−0.4	0.3
T. speech	0.3	1.3	−0.1	−0.1	−0.2	0.3
Results on SIAP-Interview						
Face (3D-ResNext)	3.9	−0.6	2.0	−3.8	−3.2	−0.6
Face (CNN-GRU)	−1.3	−3.8	−0.2	−6.2	−11.5	−5.0
Facial AU	−2.5	−7.2	−0.8	−3.5	−9.7	−4.9
Voice	4.8	−6.6	0.2	−5.4	−9.1	−3.6
T. speech	2.0	−4.2	−3.2	−2.8	−6.3	−3.1
Body pose	2.1	0.3	2.0	−3.0	−8.5	−1.9

Note: "Face" indicates facial appearance modality

5.5 Relative importance of modalities in fusion

In this section, we systematically investigate how much contribution is provided by each modality to the performance of fusion. To this end, we combine all modalities except one of them at a time, using early fusion with concatenation (without modality selection based on minimum validation error). This procedure is repeated for each modality, where the corresponding/target modality is excluded from fusion. Let E_{all} and E_{target} denote the MAE of using all modalities and the MAE of using all modalities except the target modality to be evaluated, respectively. Then, the relative importance rate for the target modality can be calculated as $(E_{\text{target}} - E_{\text{all}})/E_{\text{all}}$.

The relative importance rates of different modalities on FID and SIAP-Interview, are reported in Table 12. While the positive (high) values indicate that including the target modality in fusion improves the performance of the model (reducing the MAE), the negative (lower) importance rates suggest that having the target modality in fusion causes a performance drop.

According to the results, facial appearance (3D-ResNext) contributes to fusion accuracy very highly on FID. Voice

Table 13 MAEs of Different Methods on FID

Method	Representation	Fusion Type	AGR	CON	EXT	NEU	OPE	AVG
Proposed	Facial appearance (ResNext + CNN-GRU), Facial action units (LSTNet), Voice (LSTNet), Transcribed speech (LSTNet)	Late	0.087	0.082	0.083	0.086	0.087	0.085
Kaya *et al.* (2017) [47]*	Facial appearance (VGG-Face + LGBP-TOP), Scene (VGG-VD19), Voice (openSMILE)	Hybrid	0.086	0.080	0.079	0.085	0.083	0.083
Gurpinar *et al.* (2016) [39]	Facial appearance (VGG-Face + LGBP-TOP), Scene (VGG-VD19), Voice (openSMILE)	Late	0.093	0.085	0.082	0.089	0.086	0.087
Wei *et al.* (2017) [83]	Facial appearance (DAN+), Voice (LSTM)	Late	0.087	0.083	0.087	0.090	0.088	0.087
Subramaniam *et al.* (2016) [76]	Facial appearance (CNN), Voice (pyAudioAnalysis)	Early	0.088	0.088	0.085	0.090	0.088	0.088
Gucluturk *et al.* (2017) [35]	Facial appearance (ResNet-18), Voice (ResNet-18), Transcribed speech (Skip-thought Vectors)	Early	0.089	0.085	0.089	0.090	0.089	0.088
Bekhouche *et al.* (2017) [25]	Facial appearance (PML-BSIF + PML-LPQ)	Early	0.090	0.086	0.085	0.092	0.090	0.089
Gucluturk *et al.* (2016) [36]	Facial appearance (ResNet-18), Voice (ResNet-18)	Early	0.090	0.087	0.089	0.091	0.089	0.089
Wicaksana *et al.* (2017) [85]	Facial appearance (wMEI + AU), Transcribed speech (NLTK)	Late	0.103	0.120	0.113	0.115	0.110	0.112

Note: * denotes the combined use of training and validation sets for training

and transcribed speech modalities also have importance for fusion, however, their importance levels account for only 2% of that of the facial appearance using 3D-ResNext. Interestingly, each of the modalities has a negative importance on SIAP-Interview. Therefore, we can claim that either the non-linear relations between representations extracted from different modalities on SIAP-Interview are highly confusing, or they have high levels of redundancy. On the other hand, relative importance of the facial appearance modality using 3D-ResNext is clearly the highest one among other modalities in fusion both on FID and SIAP-Interview. Yet, on both datasets, the lowest relative importance is also observed for the facial appearance modality, however, using CNN-GRU. This may suggest that the facial representation learned from 3D-ResNext displays a more compatible latent structure, yielding better interactions with other modalities in fusion. Therefore, having only more compatible one of the facial appearance representations in the fusion would be more effective since there is a high level of feature redundancy between them.

5.6 Comparison to other methods

In this section, we compare our best performing multimodal model (with the minimum validation error) to eight recent studies, which provides results on FID for apparent personality estimation. MAEs of these methods and ours are given in Table 13, sorted in an ascending order based on their average MAEs. As seen, our proposed method outperforms all methods except the state-of-the-art proposed by Kaya *et al.*[47]. Still, we provide comparable MAEs to those of [47]; average MAE of our method is only 0.24% (absolute) higher. On the other hand, it is important to note that while our method has been trained solely on the training set (for consistency with other studies), [47] includes the validation set in the training set once the hyperparameters are optimized on the validation set. In this way, they employ 33.3% more data samples in the training.

5.7 Analysis of induced behavior

Assessment of personality traits from induced behavior is another goal of this study. To this end, as described in Sect. 4.1 we have recorded 60 subjects while they are watching short video clips. This dataset in SIAP is referred to as SIAP-Induction. Each of the aforementioned video clips target inducing the behavioral cues of (at least) one of the five personality traits. For a detailed analysis of induced behavior, we split SIAP-Induction into five subsets, each of which includes recordings during the elicitation by displaying video clips targeting one of the Big Five traits. Each of these subsets has 180 recordings (60 participants × 3 unique video clips shown for each trait). Consequently, we train and evaluate different unimodal models (facial appearance, facial AUs, and body pose)

Table 14 MAEs for the Individual Use of Facial Appearance, Facial Action Units, and Body Pose on SIAP-Induction and on Its Subsets

Modality	Induction	MAE					
	Subset	AGR	CON	EXT	NEU	OPE	AVG
Facial Appearance	AGR	0.146	0.115	0.163	0.166	0.188	0.156
	CON	0.150	0.124	0.158	0.172	0.191	0.159
	EXT	0.145	0.119	0.164	0.171	0.174	0.155
	NEU	0.160	0.126	0.172	0.161	0.172	0.158
	OPE	0.150	0.120	0.164	0.173	0.179	0.157
	AVG	0.150	0.121	0.164	0.169	0.181	0.157
	ALL	0.149	0.118	0.165	0.164	0.181	0.155
Facial AUs	AGR	0.145	0.139	0.175	0.167	0.181	0.161
	CON	0.152	0.130	0.166	0.164	0.178	0.158
	EXT	0.142	0.126	0.166	0.165	0.170	0.154
	NEU	0.147	0.126	0.168	0.156	0.172	0.154
	OPE	0.144	0.127	0.168	0.158	0.159	0.151
	AVG	0.146	0.130	0.168	0.162	0.172	0.156
	ALL	0.142	0.127	0.166	0.161	0.174	0.154
Body Pose	AGR	0.147	0.143	0.177	0.174	0.183	0.165
	CON	0.168	0.137	0.172	0.171	0.191	0.168
	EXT	0.141	0.131	0.170	0.164	0.169	0.155
	NEU	0.144	0.130	0.170	0.159	0.171	0.155
	OPE	0.140	0.133	0.165	0.165	0.171	0.155
	AVG	0.148	0.135	0.171	0.166	0.177	0.159
	ALL	0.144	0.128	0.172	0.161	0.174	0.156

Note: Whole set of SIAP-Induction is denoted as “ALL”

on SIAP-Induction and on each of its five subsets to evaluate the informativeness of induced behavior for estimating personality traits. Notice that voice and transcribed speech modalities are not included in SIAP-Induction. Besides, the use of head pose & gaze modality is discarded in the current evaluation due to its low accuracy in our previous experiments. Similarly, based on the findings of earlier experiments, 3D-ResNext is used for modeling facial appearance while LSTNet is used for modeling facial action units and body pose. Training of each model on SIAP-Induction is initiated using weights that are learned on FID.

As shown in Table 14, the best performing modality is facial action units, followed by facial appearance in SIAP-Induction. These results are in line with the ones on SIAP-Interview. Therefore, we may claim that facial information displays the most discriminative behavioral patterns of different traits. When all the induction subsets are used together, we observe a slight but consistent reduction in MAEs of all modalities. This improvement seems to be due to the increased number of training samples. In this way, a large variety of behavioral patterns can be observed by the models, allowing to develop higher generalization power.

With the use of SIAP-Induction subsets, we aim to explicitly elicit behavioral characteristics of a specific trait. Thus, if the induction is successful, the lowest prediction error is expected to be obtained on a trait-targeted subset when estimating the level of that specific trait. To this end, for each trait, we compare the MAEs obtained on different induction subsets. As expected, with the use of any modality, estimation of Neuroticism consistently obtains the lowest MAE on the Neuroticism-targeted induction sets (compared to the results on other subsets). Similarly, maximum accuracy is achieved for the estimation of Openness and Extraversion levels on their-targeted subsets with the individual use of both facial action units and body pose modalities. However, such results cannot be obtained for Agreeableness and Conscientiousness. To understand the reasons behind this, we analyze the distributions of observed scores for different traits (see Fig.7). Interestingly, the distributions of observed scores for Agreeableness and Conscientiousness are quite imbalanced (especially compared to those for Neuroticism, Openness, and Extraversion). Consequently, the variances of observed scores of Agreeableness and Conscientiousness are much lower than those of Neuroticism, Openness, and Extraversion (see Table 3). Therefore, such unexpected results regarding

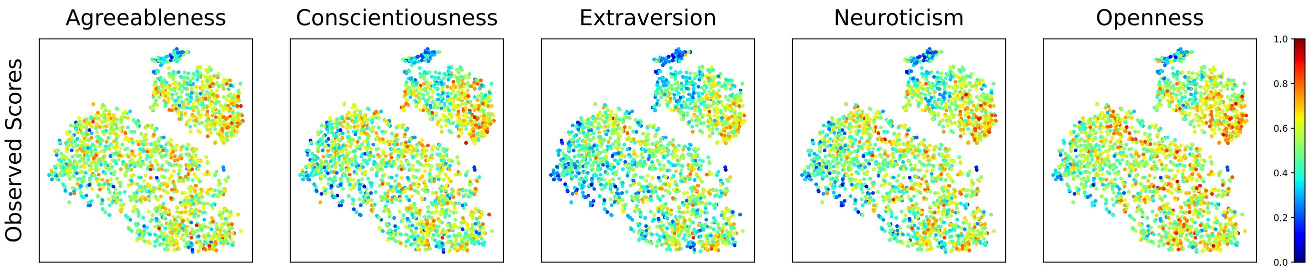

Fig. 8 t-SNE visualization of FID (test set). Each point represents a data sample, where colors indicate the observed scores of traits

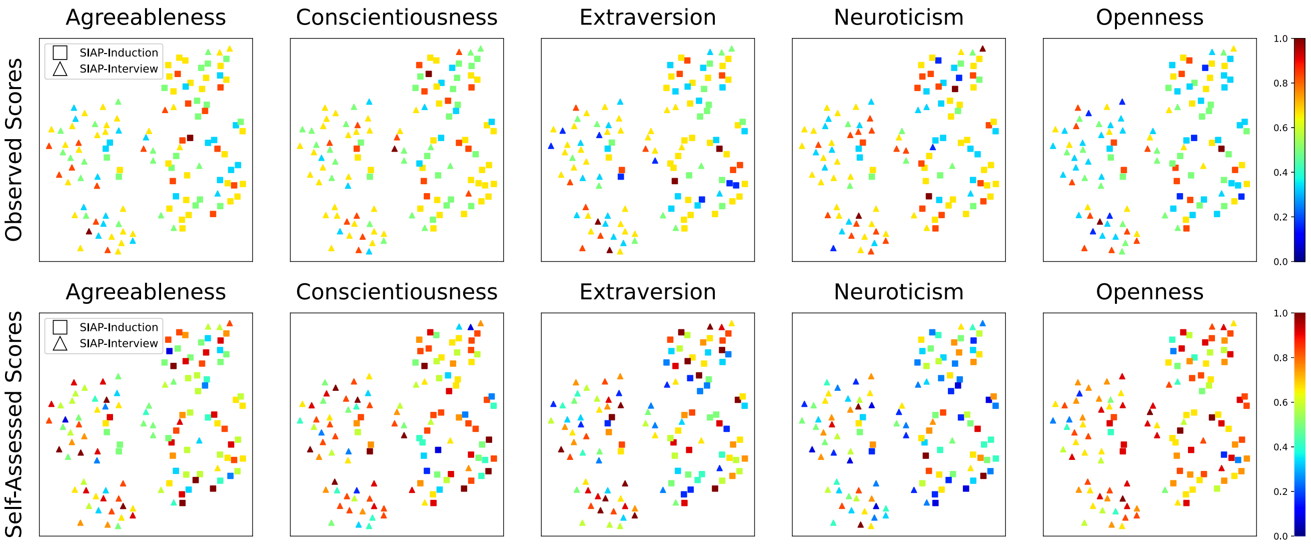

Fig. 9 Joint t-SNE visualization of SIAP-Interview and SIAP-Induction. Each point represents a data sample, where colors indicate the observed scores in the first row, and self-assessed scores in the second row

Agreeableness and Conscientiousness are due to the imbalanced distribution of their observed scores, yielding biased models (biased to the scores that are observed much more frequently in the dataset). These findings may also be valid for the results of SIAP-Interview (see Sect. 5.3). Furthermore, according to our results and findings, we can claim that induced behavior can indeed exhibit characteristics of personality traits.

5.8 Visual analysis of the latent representations

In this section, we visually analyze the latent structure of the datasets used in our experiments. Since our results suggest that facial information is the most powerful one for personality analysis, we use facial features in our analysis. Particularly, latent representations obtained from the penultimate layers of facial appearance models, namely, 3D-ResNext, CNN-GRU, and facial action units (LSTNet) model, are extracted and concatenated for samples in FID test set, SIAP-Interview, and SIAP-Induction. For comparability, the representations of samples in all of the aforementioned datasets, are extracted using the models trained on FID. Dimensionality of the combined representation is then reduced to 2D for visualization purposes. To this end, we employ t-Distributed Stochastic Neighbor Embedding (t-SNE) [58]. Notice that before applying t-SNE, Principal Component Analysis (PCA) is employed as the first step, to reduce the dimensionality maintaining 90% of the variance. Plots of the resulting 2D representations obtained from t-SNE are used for our visual analysis.

As described in Sect. 4.1.4, SIAP includes several sessions for each participant. Since t-SNE tends to cluster data samples based on the identity of participants, we do not treat each session (video) as a separate sample in the analysis. Instead, the representations obtained for all sessions of each participant, are averaged to generate a single representation. Consequently, we obtain 60 samples for each of SIAP-Interview and SIAP-Induction. This procedure is not applied to FID since its samples are of different subjects.

As shown in Fig. 8, FID (test set) samples are distributed in a way that the trait scores change smoothly. In other words, the facial models learned on the training set of FID is able to capture the ordinal transition of behavioral patterns based on trait scores. When the latent structures of SIAP-Interview

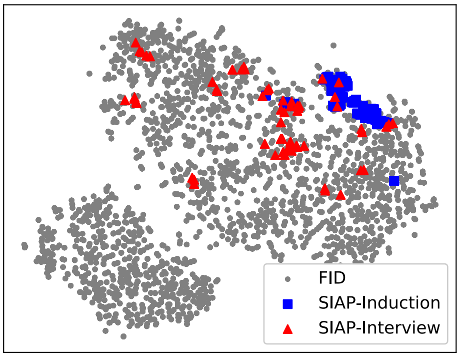

Fig. 10 Joint t-SNE visualization of FID (test set), SIAP-Interview, and SIAP-Induction. Each point represents a data sample, where colors denote the corresponding dataset

and SIAP-Induction are analyzed jointly (see Fig. 9), we observe that these datasets are mostly separated in the latent space forming their own clusters. This finding, indicates the existence of distinguishable differences between facial behavioral patterns displayed during self-presentation and during visual induction. In Fig. 9, while samples with similar observed scores are grouped together (in several clusters), this is not the case for the self-assessed scores. In terms of self-assessed scores, samples with very high and very low trait levels are spread all over the latent space without a trend. This shows the reliability and objectiveness of the expert annotations for SIAP in contrast to the inconsistency of self-assessed scores. Our findings also indicate that people may become extremists during self evaluation. In line with the literature [64], due to coupling of social desirability effects (see Sect. 5.2) and extremity, more prominent differences in Openness, Extraversion and Neuroticism are observed between self-assessed and observed scores.

Lastly, we compare the sample distributions of FID (test set), SIAP-Interview, and SIAP-Induction in the latent space. As visualized in Fig. 10, samples of SIAP-Induction are clustered closely, while SIAP-Interview and FID spread over a large manifold. This shows that the variability of facial expressions (or inner facial movements) displayed during self-presentations are much higher than that of induced behavior. Moreover, it is clear that behavioral characteristics of FID and SIAP-Interview samples resemble each other. Yet, samples of SIAP-Induction are tightly located on the boundary of the data manifold, suggesting that facial patterns of induced behavior are mostly different from those of self-presentations (SIAP-Interview and FID).

6 Conclusion

In this study, we have developed and presented spatio-temporal models that automatically assess the level of the Big Five personality traits (i.e., Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness), from different behavioral modalities such as facial appearance, facial action units, head pose & gaze, body pose, voice, and transcribed speech. State-of-the-art deep architectures have been utilized to this end. For a detailed analysis of personality, we have collected Self-presentation and Induced Behavior Archive for Personality Analysis (SIAP) that consists of speech and induced behavior recordings of 60 participants. Each session in the dataset has three recordings acquired from frontal, right-front side, and left-rear side cameras. In addition, SIAP includes both self-assessed and observed scores for each of the Big Five personality traits.

Using the developed methods, we have systematically evaluated the discriminative power of the aforementioned modalities for the assessment of personality on two different datasets, namely on SIAP and ChaLearn LAP First Impressions (FID) datasets. In our experiments, face-related modalities have been found to be the most reliable ones for personality analysis. The best performing combinations of behavioral modalities have also been analyzed through various fusion strategies. As well as providing baseline results on Self-presentation and Induced Behavior Archive for Personality Analysis, we obtain comparable results to the state-of-the-art on FID. Based on our extensive experiments, we present new findings such that: (1) Induced behavior can display cues of personality; (2) People tend to express themselves as more socially desirable in their self-reports, when compared with the observed ones; (3) Visual modalities can be more effective than audio or speech in personality analysis, yet, fusion of these modalities can further improve the reliability of predictions.

Acknowledgements Authors would like to thank Mr. Alper Dağgez, Ms. Merve Nur Doğanlı and Prof. Gül Günaydin for their contributions to this study.

Funding Not applicable.

Data Availability Statement The collected SIAP dataset will be made available for research purposes upon requests with a signed EULA.

Compliance with ethical standards

Conflict of interest The authors have no conflict of interest.

Ethical approval The current study received the required ethical approval from the Ethics Committee of the university. Each participant of the experiment was informed about the procedure, and signed a consent form.

References

1. Abadi MK, Correa JAM, Wache J, Yang H, Patras I, Sebe N (2015) Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos. In: 2015 11th IEEE international conference and workshops on automatic face and gesture

recognition (FG), Ljubljana, pp 1–8. https://doi.org/10.1109/FG.2015.7163100
2. Abele AE, Wojciszke B (2007) Agency and communion from the perspective of self versus others. J Pers Soc Psychol 93(5):751
3. Alam F, Riccardi G (2014) Predicting personality traits using multimodal information. In: Proceedings of the 2014 ACM multi media on workshop on computational personality recognition, pp 15–18
4. Alam F, Stepanov EA, Riccardi G (2013) Personality traits recognition on social network-facebook. In: International AAAI conference on weblogs and social media
5. Almaev TR, Valstar MF (2013) Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: Humaine association conference on affective computing and intelligent interaction, pp 356–361. IEEE
6. Aran O, Gatica-Perez D (2013) Cross-domain personality prediction: from video blogs to small group meetings. In: Proceedings of the 15th ACM on international conference on multimodal interaction, pp 127–130
7. Ashton MC, Lee K (2007) Empirical, theoretical, and practical advantages of the Hexaco model of personality structure. Personality Soc Psychol Rev 11(2):150–166
8. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) Openface 2.0: facial behavior analysis toolkit. In: Proceedings of the 13th IEEE international conference on automatic face & gesture recognition (FG 2018), pp 59–66. IEEE
9. Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. Psychol Rep 19(1):3–11
10. Biel JI, Aran O, Gatica-Perez D (2011) You are known by how you vlog: personality impressions and nonverbal behavior in youtube. In: Fifth international AAAI conference on weblogs and social media
11. Biel JI, Gatica-Perez D (2012) The youtube lens: crowdsourced personality impressions and audiovisual analysis of vlogs. IEEE Trans Multimedia 15(1):41–55
12. Boyette LL, Korver-Nieberg N, Meijer C, de Haan L (2014) Genetic risk and outcome of psychosis investigators: quality of life in patients with psychotic disorders: impact of symptoms, personality, and attachment. J Nerv Ment Dis 202(1):64–69
13. Boyette LL, Korver-Nieberg N, Verweij K, Meijer C, Dingemans P, Cahn W, de Haan L, Kahn R, de Haan L, van Os J, Wiersma D, Bruggeman R, Cahn W, Meijer C, Myin-Germeys I (2013) Associations between the five-factor model personality traits and psychotic experiences in patients with psychotic disorders, their siblings and controls. Psychiatry Res 210(2):491–497
14. Bromet E, Andrade LH, Hwang I, Sampson NA, Alonso J, De Girolamo G, De Graaf R, Demyttenaere K, Hu C, Iwata N, Karam A, Kaur J, Kostyuchenko S, Lépine JP, Levinson D, Matschinger H, Mora M, Browne M, Posada-Villa J, Viana M, Williams D, Kessler R (2011) Cross-national epidemiology of DSM-IV major depressive episode. BMC Med 9(1):90
15. Butcher JN, Graham JR, Williams CL, Ben-Porath YS (1990) Development and use of the MMPI-2 content scales. University of Minnesota Press, Minneapolis
16. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7291–7299
17. Celiktutan O, Gunes H (2014) Continuous prediction of perceived traits and social dimensions in space and time. In: 2014 IEEE international conference on image processing (ICIP), pp 4196–4200. IEEE
18. Celiktutan O, Gunes H (2015) Automatic prediction of impressions in time and across varying context: personality, attractiveness and likeability. IEEE Trans Affect Comput 8(1):29–42
19. Chandrashekar P (2018) Do mental health mobile apps work: evidence and recommendations for designing high-efficacy mental health mobile apps. Mhealth 4:6
20. Compton MT, Bakeman R, Alolayan Y, Balducci PM, Bernardini F, Broussard B, Crisafio A, Cristofaro S, Amar P, Johnson S, Wan CR (2015) Personality domains, duration of untreated psychosis, functioning, and symptom severity in first-episode psychosis. Schizophr Res 168(1–2):113–119
21. Corr PJ, Matthews G (2009) The Cambridge handbook of personality psychology. Cambridge University Press, New York
22. Costa PT Jr, Terracciano A, McCrae RR (2001) Gender differences in personality traits across cultures: robust and surprising findings. J Pers Soc Psychol 81(2):322
23. Cucurull G, Rodríguez P, Yazici VO, Gonfaus JM, Roca FX, Gonzàlez J (2018) Deep inference of personality traits by integrating image and word use in social networks. arXiv preprint arXiv:1802.06757
24. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
25. Eddine Bekhouche S, Dornaika F, Ouafi A, Taleb-Ahmed A (2017) Personality traits and job candidate screening via analyzing facial videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 10–13
26. Escalante HJ, Kaya H, Salah AA, Escalera S, Güç Y, Güçlü U, Baró X, Guyon I, Jacques JCS, Madadi M, Ayache S, Viegas E, Gurpinar F, Wicaksana AS, Liem C, Van Gerven MA, Van Lier R (2020) Modeling, recognizing, and explaining apparent personality from videos. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2020.2973984
27. Fan Y, Lu X, Li D, Liu Y (2016) Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th ACM international conference on multimodal interaction, pp 445–450. ACM
28. Farnadi G, Sushmita S, Sitaraman G, Ton N, De Cock M, Davalos S (2014) A multivariate regression approach to personality impression recognition of vloggers. In: Proceedings of the ACM multimedia workshop on computational personality recognition, pp 1–6
29. Feingold A (1994) Gender differences in personality: a meta-analysis. Psychol Bull 116(3):429
30. Google cloud speech-to-text. https://cloud.google.com/speech-to-text. Accessed 2019-09-15
31. Giannakopoulos T (2015) pyaudioanalysis: An open-source python library for audio signal analysis. PloS ONE 10(12):e0144610
32. Goldberg LR (1990) An alternative“ description of personality”: the big-five factor structure. J Pers Soc Psychol 59(6):1216
33. Gosling SD, Rentfrow PJ, Swann WB Jr (2003) A very brief measure of the big-five personality domains. J Res Pers 37(6):504–528
34. Graziano WG, Eisenberg N (1997) Agreeableness: a dimension of personality. In: Handbook of personality psychology, pp 795–824. Elsevier
35. Güçlütürk Y, Güçlü U, Baro X, Escalante HJ, Guyon I, Escalera S, Van Gerven MA, Van Lier R (2017) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9(3):316–329
36. Güçlütürk Y, Güçlü U, van Gerven MA, van Lier R (2016) Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition. In: European conference on computer vision, pp 349–358. Springer
37. Guntuku SC, Qiu L, Roy S, Lin W, Jakhetiya V (2015) Do others perceive you as you want them to? modeling personality based on selfies. In: Proceedings of the 1st international workshop on affect & sentiment in multimedia, pp 21–26

38. Gürpınar F, Kaya H, Salah AA (2016) Combining deep facial and ambient features for first impression estimation. In: European Conference on Computer Vision, pp. 372–385. Springer
39. Gürpinar F, Kaya H, Salah AA (2016) Multimodal fusion of audio, scene, and face features for first impression estimation. In: Proceedings of the international conference on pattern recognition (ICPR), pp 43–48. IEEE
40. Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3D CNNS retrace the history of 2D CNNS and imagenet? In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 6546–6555
41. Hasan A (2013) The Turkish adaptation of the ten-item personality inventory. Nöro Psikiyatri Arşivi 50(4):312
42. Hassan S, Akhtar N, Yılmaz AK (2016) Impact of the conscientiousness as personality trait on both job and organizational performance. J Manag Sci 10(1):1–14
43. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 770–778. Retrieved from https://repository.upenn.edu/asc_papers/43. Accessed 6 Oct 2020
44. Huggingface: huggingface/pytorch-transformers (2019). https://github.com/huggingface/pytorch-transformers. Accessed 29 July 2019
45. Husain MI, Carvalho AF (2020) The importance of assessing personality traits and disorders in clinical trials of major depressive disorder. Braz J Psychiatry 42(1):3–4
46. Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A (2017) The kinetics human action video dataset. arXiv preprint arXiv:1705.06950
47. Kaya H, Gurpinar F, Ali Salah A (2017) Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVS. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–9
48. Keltner D, Ekman P, Gonzaga GC, Beer J (2003) Facial expressions of emotion. In: Davidson RJ, Scherer KR, Goldsmith HH (eds) Handbook of affective sciences. Oxford University Press, New York, pp 415–432
49. Khan AA, Jacobson KC, Gardner CO, Prescott CA, Kendler KS (2005) Personality and comorbidity of common psychiatric disorders. Br J Psychiatry 186(3):190–196
50. Kim B, Joo YH, Kim SY, Lim JH, Kim EO (2011) Personality traits and affective morbidity in patients with bipolar I disorder: the five-factor model perspective. Psychiatry Res 185(1–2):135–140
51. Kim B, Lim JH, Kim SY, Joo YH (2012) Comparative study of personality traits in patients with bipolar I and II disorder from the five-factor model perspective. Psychiatry Investig 9(4):347
52. Kotov R, Gamez W, Schmidt F, Watson D (2010) Linking “big” personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. Psychol Bull 136(5):768
53. Krippendorff, K (2011). Computing Krippendorff’s alpha-reliability. Departmental Papers (ASC) https://repository.upenn.edu/asc_papers/43. Accessed 6 Oct 2020
54. Lai G, Chang WC, Yang Y, Liu H (2018) Modeling long-and short-term temporal patterns with deep neural networks. In: Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval, pp 95–104. ACM
55. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. In: Proceedings of the 29th AAAI conference on artificial intelligence
56. Lecomte T, Spidel A, Leclerc C, MacEwan GW, Greaves C, Bentall RP (2008) Predictors and profiles of treatment non-adherence and engagement in services problems in early psychosis. Schizophr Res 102(1–3):295–302
57. Lysaker PH, Davis LW (2004) Social function in schizophrenia and schizoaffective disorder: associations with personality, symptoms and neurocognition. Health Qual Life Outcomes 2(1):15
58. van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
59. McCrae RR, John OP (1992) An introduction to the five-factor model and its applications. J Pers 60(2):175–215
60. Mulder RT (2002) Personality pathology and treatment outcome in major depression: a review. Am J Psychiatry 159(3):359–371
61. Odić A, Tkalčič M, Tasič J, Košir A (2013) Personality and social context: impact on emotion induction from movies. In: Extended proceedings of the conference on user modeling, adaptation, and personalization
62. Ohi K, Shimada T, Nitta Y, Kihara H, Okubo H, Uehara T, Kawasaki Y (2016) The five-factor model personality traits in schizophrenia: a meta-analysis. Psychiatry Res 240:34–41
63. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British machine vision conference. British Machine Vision Association
64. Pedregon CA, Farley RL, Davis A, Wood JM, Clark RD (2012) Social desirability, personality questionnaires, and the “better than average” effect. Personality Individ Differ 52(2):213–217
65. Ponce-López V, Chen B, Oliu M, Corneanu C, Clapés A, Guyon I, Baró X, Escalante HJ, Escalera S (2016) Chalearn lap 2016: first round challenge on first impressions-dataset and results. In: Proceedings of the European conference on computer vision, pp 400–418. Springer
66. Qiu L, Lu J, Yang S, Qu W, Zhu T (2015) What does your selfie say about you? Comput Hum Behav 52:443–449
67. Quilty LC, De Fruyt F, Rolland JP, Kennedy SH, Rouillon PF, Bagby RM (2008) Dimensional personality traits and treatment outcome in patients with major depressive disorder. J Affect Disord 108(3):241–250
68. Riggio HR, Riggio RE (2002) Emotional expressiveness, extraversion, and neuroticism: a meta-analysis. J Nonverbal Behav 26(4):195–218
69. Salem H, Ruiz A, Hernandez S, Wahid K, Cao F, Karnes B, Beasley S, Sanches M, Ashtari E, Pigott T (2019) Borderline personality features in inpatients with bipolar disorder: impact on course and machine learning model use to predict rapid readmission. J Psychiatr Pract 25(4):279–289
70. Segalin C, Cheng DS, Cristani M (2017) Social profiling through image understanding: personality inference using convolutional neural networks. Comput Vis Image Underst 156:34–50
71. Selçuk E, Günaydin G, Sümer N, Uysal A (2005) Yetişkin bağlanma boyutları için yeni bir ölçüm: Yakın ilişkilerde yaşantılar envanteri-II’nin Türk örnekleminde psikometrik açıdan değerlendirilmesi. Türk Psikoloji Yazıları
72. Silveira Jacques Junior JC, Güçlütürk Y, Pérez M, Güçlü U, Andujar C, Baró X, Escalante HJ, Guyon I, Van Gerven MA, Van Lier R, Escalera S (2019) First impressions: a survey on vision-based apparent personality trait analysis. IEEE Trans Affect Comput. https://doi.org/10.1109/TAFFC.2019.2930058
73. Smeland OB, Wang Y, Lo MT, Li W, Frei O, Witoelar A, Tesli M, Hinds DA, Tung JY, Djurovic S, Chen CH, Dale AM, Andreassen OA (2017) Identification of genetic loci shared between schizophrenia and the big five personality traits. Sci Rep 7(1):1–9
74. Sparkman DJ, Eidelman S, Dueweke AR, Marin MS, Dominguez B (2019) Open to diversity: openness to experience predicts beliefs in multiculturalism and colorblindness through perspective taking. J Individ Diff 40(1):1–12. https://doi.org/10.1027/1614-0001/a000270
75. Spearman C (1961) The proof and measurement of association between two things. In: Jenkins JJ, Paterson DG (eds) Studies

in individual differences: the search for intelligence. Appleton-Century-Crofts, pp 45–58. https://doi.org/10.1037/11491-005
76. Subramaniam A, Patel V, Mishra A, Balasubramanian P, Mittal A (2016) Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In: European conference on computer vision, pp 337–348. Springer
77. Valente F, Kim S, Motlicek P (2012) Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In: Annual conference of the international speech communication association
78. Ventura C, Masip D, Lapedriza A (2017) Interpreting CNN models for apparent personality trait regression. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 55–63
79. Vielzeuf V, Lechervy A, Pateux S, Jurie F (2018) Centralnet: a multilayer approach for multimodal fusion. In: Proceedings of the European conference on computer vision workshops, pp 575–589
80. Vinciarelli A, Mohammadi G (2014) A survey of personality computing. IEEE Trans Affect Comput 5(3):273–291
81. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W (2016) CNN-RNN: a unified framework for multi-label image classification. In: The IEEE conference on computer vision and pattern recognition (CVPR)
82. Wauthia E, Lefebvre L, Huet K, Blekic W, El Bouragui K, Rossignol M (2019) Examining the hierarchical influences of the big-five dimensions and anxiety sensitivity on anxiety symptoms in children. Front Psychol 10:1185
83. Wei XS, Zhang CL, Zhang H, Wu J (2017) Deep bimodal regression of apparent personality traits from short video sequences. IEEE Trans Affect Comput 9(3):303–315
84. Weisberg YJ, DeYoung CG, Hirsh JB (2011) Gender differences in personality across the ten aspects of the big five. Front Psychol 2:178
85. Wicaksana AS, Liem CC (2017) Human-explainable features for job candidate screening prediction. In: IEEE Conference on computer vision and pattern recognition workshops (CVPRW), pp 1664–1669. IEEE
86. Willis J, Todorov A (2006) First impressions: Making up your mind after a 100-ms exposure to a face. Psychol Sci 17(7):592–598
87. Wu X, He H, Shi L, Xia Y, Zuang K, Feng Q, Zhang Y, Ren Z, Wei D, Qiu J (2019) Personality traits are related with dynamic functional connectivity in major depression disorder: a resting-state analysis. J Affect Disord 245:1032–1042
88. Xie S, Girshick R, Dollár P, Tu Z, He K (2017) Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1492–1500
89. Yan Y, Nie J, Huang L, Li Z, Cao Q, Wei Z (2016) Exploring relationship between face and trustworthy impression using mid-level facial features. In: International conference on multimedia modeling, pp 540–549. Springer
90. Yik MS, Russell JA (2001) Predicting the big two of affect from the big five of personality. J Res Pers 35(3):247–277
91. Zillig LMP, Hemenover SH, Dienstbier RA (2002) What do we assess when we assess a big 5 trait? a content analysis of the affective, behavioral, and cognitive processes represented in big 5 personality inventories. Pers Soc Psychol Bull 28(6):847–858

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.