Image and Vision Computing 110 (2021) 104163

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Multimodal assessment of apparent personality using feature attention and error consistency constraint

Süleyman Aslan, Uğur Güdükbay, Hamdi Dibeklioğlu *

Department of Computer Engineering, Bilkent University, Ankara, Turkey

ARTICLE INFO

Article history:
Received 22 August 2020
Accepted 19 March 2021
Available online 24 March 2021

Keywords:
Deep learning
Apparent personality
Multimodal modeling
Information fusion
Feature attention
Error consistency

ABSTRACT

Personality computing and affective computing, where the recognition of personality traits is essential, have gained increasing interest and attention in many research areas recently. We propose a novel approach to recognize the Big Five personality traits of people from videos. To this end, we use four different modalities, namely, ambient appearance (scene), facial appearance, voice, and transcribed speech. Through a specialized subnetwork for each of these modalities, our model learns reliable modality-specific representations and fuse them using an attention mechanism that re-weights each dimension of these representations to obtain an optimal combination of multimodal information. A novel loss function is employed to enforce the proposed model to give an equivalent importance for each of the personality traits to be estimated through a consistency constraint that keeps the trait-specific errors as close as possible. To further enhance the reliability of our model, we employ (pre-trained) state-of-the-art architectures (i.e., ResNet, VGGish, ELMo) as the backbones of the modality-specific subnetworks, which are complemented by multilayered Long Short-Term Memory networks to capture temporal dynamics. To minimize the computational complexity of multimodal optimization, we use two-stage modeling, where the modality-specific subnetworks are first trained individually, and the whole network is then fine-tuned to jointly model multimodal data. On the large scale ChaLearn First Impressions V2 challenge dataset, we evaluate the reliability of our model as well as investigating the informativeness of the considered modalities. Experimental results show the effectiveness of the proposed attention mechanism and the error consistency constraint. While the best performance is obtained using facial information among individual modalities, with the use of all four modalities, our model achieves a mean accuracy of 91.8%, improving the state of the art in automatic personality analysis.

1. Introduction

Personality and emotions have a strong influence on people's lives and they affect behaviors, cognitions, preferences, and decisions. Emotions have distinct roles in decision making, such as providing information about pleasure and pain, enabling rapid choices under time pressure, focusing attention on relevant aspects of a problem, and generating commitment concerning decisions [1]. Additionally, research suggests that human decision-making process can be modeled as a two-system model, consisting of rational and emotional systems [2]. Accordingly, emotions are part of every decision-making process instead of simply affecting on these processes. Likewise, personality also has an important effect on decision making and it causes individual differences in people's thoughts, feelings, and motivations. It can be observed that there are significant relationships among attachment styles, decision-making styles, and personality traits [3]. Besides, personality relates to individual differences in preferences, such as the use of music in everyday life [4,5], and user preferences in multiple entertainment domains including books, movies, and TV shows [6]. Due to the fact that emotion and personality have an essential role in human cognition and perception, there has been a growing interest in recognizing the human personality and affect and integrating them into computing to develop artificial emotional intelligence, which is also known as "affective computing" [7], in combination with "personality computing" [8]. Hence, it becomes essential to recognize the personality and emotion of humans precisely. We present a novel multimodal framework to recognize the apparent personality of individuals from videos to address this problem.

1.1. Personality traits

Personality can be defined as the psychological factors that influence an individual's patterns of behaving, thinking, and feeling that differentiate the individual from others [9,10]. The most mainstream and widely accepted framework for personality among psychology researchers is

* Corresponding author.
E-mail addresses: suleyman.aslan@bilkent.edu.tr (S. Aslan), gudukbay@cs.bilkent.edu.tr (U. Güdükbay), dibeklioglu@cs.bilkent.edu.tr (H. Dibeklioğlu).

https://doi.org/10.1016/j.imavis.2021.104163
0262-8856/

the Five-Factor Model (FFM) [10,11]. FFM is a model based on descriptors of human personality in five dimensions as a complete description of personality. Various researchers identified the same five factors within independent works in personality theory [11,12,13]. Therefore, it is reliable to define personality with FFM.

Based on the work by Costa, McCrae, and John [10,11], the five factors are defined as follows:

- *Openness (O):* Appreciation of experience and curiosity of the unfamiliar.
- *Conscientiousness (C):* Level of organization and being dependable.
- *Extraversion (E):* Social activity and interpersonal interaction.
- *Agreeableness (A):* Tendency to work cooperatively with others and avoiding conflicts.
- *Neuroticism (N):* Emotional instability and being prone to psychological distress.

These five factors lead to bipolar characteristics that can be seen in individuals that score low and high on each trait, as shown in Table 1. The factors are often represented by the acronym *OCEAN*.

Our proposed model estimates the level of traits based on four modalities, namely, facial appearance, ambient appearance (scene), voice, and transcribed speech. To this end, a subnetwork is designed for each modality where we exploit state-of-the-art (pre-trained) deep architectures as the backbone of our model and complement them with Long Short-Term Memory (LSTM) networks to leverage temporal information. To effectively fuse multimodal representations, we design and employ a feature attention layer. Furthermore, an error consistency constraint is introduced in the loss function to prevent overfitting to some of the traits. For effective modeling, two-stage training is employed, where we first train the modality-specific networks individually, and after combining these subnetworks, the whole model is fine-tuned in a multimodal manner. State-of-the-art results are obtained using the proposed approach on the ChaLearn First Impressions V2 challenge dataset [14].

1.2. Contributions

Main contributions of this study to the area of automatic (apparent) personality analysis can be listed as follows: (1) An accurate multimodal personality estimation model is proposed, which outperforms the state-of-the-art methods; (2) For effective fusion of multimodal information, a feature attention mechanism is introduced; (3) To prevent the proposed model from overfitting to some of the five personality traits during joint (multi-task) training, an error consistency term is included in the loss function; (4) Informativeness of different modalities and the reliability of fusing different combinations of them are investigated.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents the proposed method that effectively learns modality-specific representations and fuses them for estimating the levels of apparent personality traits. Section 4 evaluates the proposed method in terms of different quality aspects and presents the results. Finally, Section 5 concludes the paper and provides possible future research directions.

Table 1
The characteristics of personality traits.

Personality trait	Low scorer	High scorer
Neuroticism	Calm, secure	Nervous, sensitive
Extraversion	Quiet, reserved	Talkative, sociable
Openness	Cautious, conventional	Inventive, creative
Agreeableness	Suspicious, uncooperative	Helpful, friendly
Conscientiousness	Careless, negligent	Organized, reliable

2. Related work

Personality computing benefits from methods aimed towards understanding, predicting, and synthesizing human behavior [8]. The effectiveness of analyzing such important aspects of individuals is the main reason behind the growing interest in this topic. Automatic recognition of apparent personality is a part of many applications such as human-computer interaction, computer-based learning, automatic job interviews, autonomous agents, and crowd simulations [14,15,16,17,18,19]. Similarly, emotion is incorporated into adaptive systems in order to improve the effectiveness of personalized content and bring the systems closer to the users [20]. As a result, personality and emotion-based user information is used in many systems, such as effective e-learning [21], conversational agents [22], crowd simulations [23], and recommendation systems [24,25,26]. Overall, personality is usually relevant in any system involving human behavior. Rapid advances in personality computing and affective computing led to the releases of novel datasets for apparent personality and emotional states of people from various sources of information such as physiological responses or video blogs [14,27]. One of the latest problems is recognizing the five personality traits (OCEAN) automatically from videos of people speaking in front of a camera.

There are notable approaches to recognizing personality traits. By analyzing the audio from spoken conversations [28] and based on the tune and rhythm aspects of speech [29], it is possible to annotate and recognize the personality traits or predict the speaker's attitudes automatically. These approaches demonstrate that audio information is important for personality detection.

Some studies utilize the status text of users on social networks for the recognition of personality traits [30]. It is also possible to explore the projection of personality, especially extraversion, through specific linguistic factors across different social contexts using transcribed video blogs and dialogs [31]. These studies indicate that there is a strong correlation between users' behavior on social networks and their personality [32]. Additionally, the exploitation of images and words used in public profiles in social networks is a way of obtaining an effective personality trait model, as shown in [33,34].

There are methods for recognition based on combinations of speaking style and body movements. Pianesi et al. [35] automatically detect personality traits in social interactions from acoustic features encoding specific aspects of the interaction and visual features such as head, body, and hands fidgeting. Likewise, Batrinca et al. [36] automatically detect five-factor personality traits in a short self-presentation based on the effectiveness of acoustic and visual non-verbal features such as pitch, acoustic intensity, hand movement, head orientation, posture, mouth fidgeting, and eye-gaze. They later extend their work by extracting these features from multimodal data in human-machine and human-human interaction scenarios [37]. Body gestures, head movements, facial expressions, and speech based on naturally occurring human affective behavior lead to effective assessment of personality and emotion [38].

Facial physical attributes from ambient face photographs are important in modeling trait factor dimensions underlying social traits [39]. Some studies utilize the facial attributes to infer personality traits. Qin et al. [40] perform experiments to evaluate personality traits and intelligence from facial morphological features. Gürpnar et al. [41] use videos depicting faces to predict personality impressions. Fernando et al. [42] use facial features to identify the personality traits from a face image. Yan et al. [43] use mid-level facial features to establish a relationship between facial appearance and personality impression. Ventura et al. [44] performed a study to investigate why convolutional neural networks (CNN) are very successful in automatically recognizing the personality traits of people speaking to a camera. The study shows that face provides the most discriminative information for this task and CNNs primarily inspect crucial parts of the face, such as mouth, nose, and eyes. Combining video appearance and motion gives impressive results for emotion recognition as well [45].

Some studies support the idea that it is human behavior to evaluate individuals by their faces concerning their personality traits and intelligence since self-reported personality traits can be predicted reliably from a facial image [40]. Gürpnar et al. [41] show that the impressions that influence people's behavior towards others can be accurately predicted from videos. Additionally, predicting personality factors for personality-based sentiment classification is beneficial in the analysis of public sentiment implied in user-generated content [46]. Because the personality affects various modalities, personality traits can be automatically recognized by combining multiple features by exploiting a multimodal approach [35,41,47,48].

According to [49], attributes and features such as audio-visual, text, demographic, and sentiment features are essential parts of a personality recognition system. Overall, different modalities provide useful information to infer apparent personality traits such as appearance facial features, low-level acoustic features (pitch, intensity, frequencies), body motion, and lexical features [50]. Although researchers commonly use multimodal approaches to recognize personality traits, the studies using a multimodal approach are relatively limited. Some notable examples are using deep residual networks for impression prediction using a combination of audiovisual and language modalities [51] and providing explainability of multimodal information in the context of first impressions analysis [52]. Moreover, exploring the variability in impressions under varying situational context and different observed modalities are essential to obtain a complete assessment of the observed individuals' personality [53].

3. Proposed framework

In our framework, we take a video clip of a single person as input and predict the personality traits associated with that person. The proposed framework initially learns the modality-specific personality features, then fuse and model those learned high-level features to obtain a final prediction of personality traits. To this end, four modality-specific subnetworks are employed, namely focusing on *ambient appearance*, *facial appearance*, *voice*, and *speech transcription*.

Modeling is performed in two stages. In the first stage, as shown in Fig. 1(a), each subnetwork is trained separately to learn modality-specific representations. In the second stage, the learned models are used as feature extractors, and the representations obtained for the ambient appearance, facial appearance, voice, and speech transcription are fused by concatenation and fed to the regressors enabling feature attention to jointly model these modalities to estimate the scores/levels of the five personality traits (i.e., Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism; see Fig. 1(b)). The details of the proposed framework will be given in the following sections.

3.1. Architecture and modality-specific modeling

3.1.1. Ambient appearance subnetwork

One of the modalities used in the proposed framework is the ambient appearance observed in the videos, such as surrounding objects, lighting, and clothing. The intuition behind employing ambient features

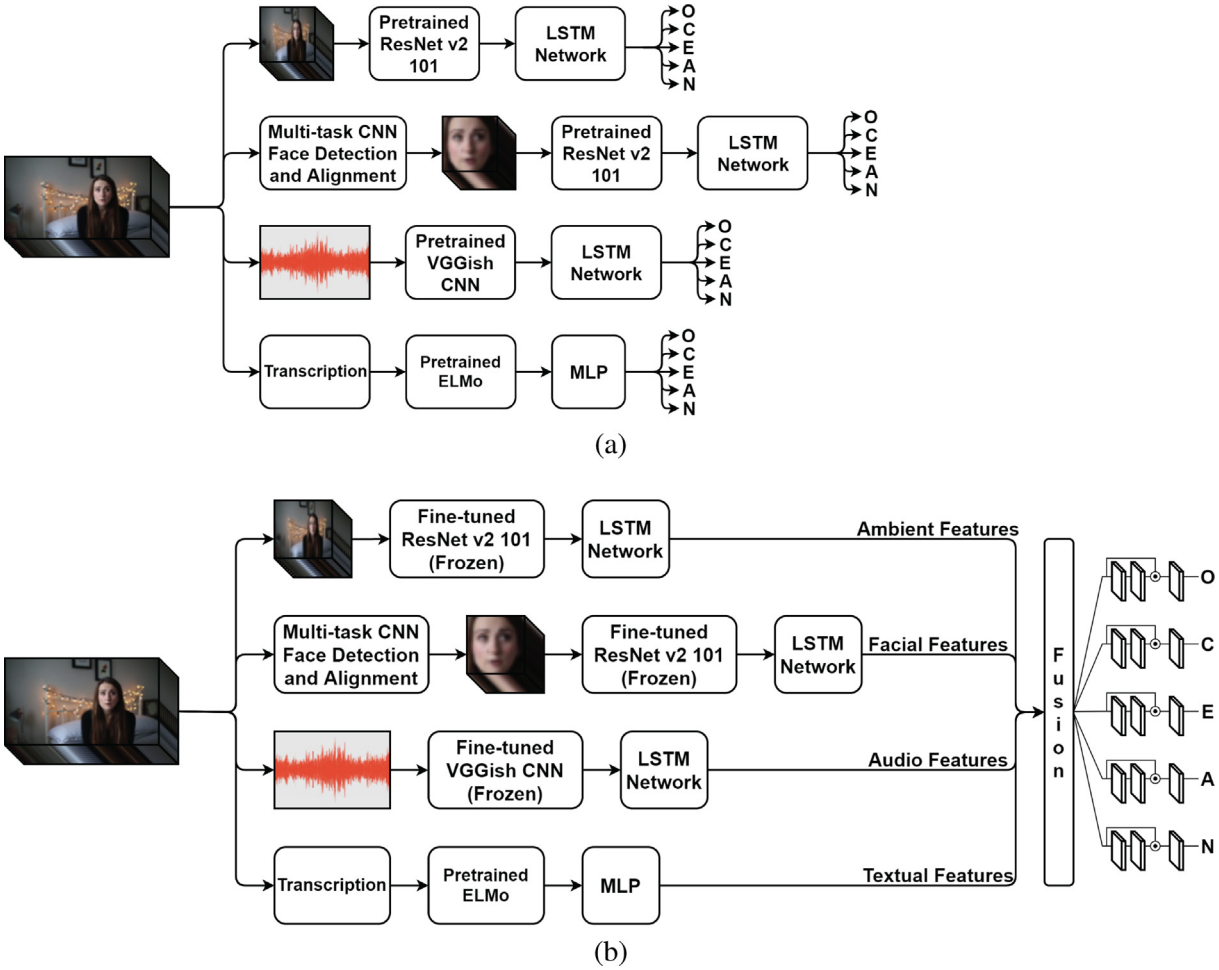

Fig. 1. Overview of the proposed model: (a) In the first stage, subnetworks are separately modeled to estimate personality scores to learn modality-specific representations; (b) In the second stage, modality-specific representations obtained through the learned models are concatenated and fed to the attention module followed by regressors to estimate the level of each personality trait.

is that the relation between environmental setup and the apparent personality of the target subject as well as the influence of ambient properties on observers' perception. Findings in the literature demonstrate that environmental elements such as surroundings, colors, and lighting affect the mood and perception [54]. Moreover, such features provide additional information about the preferences and characteristics of the analyzed subject. It is important to note that, ambient appearance network also analyzes face and pose, implicitly since face and body regions are included in the input frames.

To reduce the computational complexity and temporal noise during modeling ambient features, we use subsampled frames from the input videos. To this end, using a non-overlapping sliding window, one frame is sampled for each one-second interval. In this way, with uniform sampling, effective modeling can be achieved without losing significant information. Next, each of the obtained frames is resized to 224×224 pixels. Color information is retained and all of the frames are defined in RGB color space.

To model spatio-temporal ambient characteristics in videos, we employ a CNN in combination with an LSTM network. ResNet-v2-101 [55] is chosen as the CNN backbone and connected to a multilayer LSTM network. The ResNet-v2-101 is initialized with the pre-trained weights [55] (on ILSVRC-2012-CLS image classification dataset [56]). Particularly, we remove the last (classification) layer of the ResNet-v2-101 and obtain 2048-dimensional features for each frame. On top of it, six LSTM layers are placed including residual connections and dropout layers. Between the input and output of each of the first, third, and fifth LSTM layers, a residual connection is added and the output of the ResNet is fed to the 1st LSTM layer. Dropout is employed at each of the second, fourth, and sixth LSTM layers. We set the dimensions of the hidden state of the LSTM cells as 2048, 512, 512, 128, 128, and 64 for the first, second, third, fourth, fifth, and sixth layers, respectively. The output of the sixth LSTM layer of the last time step (last frame) is connected to a fully-connected layer with five neurons, where each neuron outputs the score prediction for one of the five personality traits.

3.1.2. Facial appearance subnetwork

As shown in the literature, personality can be assessed from facial appearance and expressions [57]. Facial symmetry is also associated with five-factor personality traits [58]. Hence, it is crucial to employ facial information in the assessment of personality. Although the facial texture is implicitly analyzed through the ambient appearance subnetwork, in facial appearance subnetwork, faces are cropped and analyzed as the sole input of the model.

First, the target face is detected in each of the temporally subsampled frames (one frame for each one-minute interval; see Section 3.1.1) of the input video using a state-of-the-art method, namely, Multi-task CNN (MTCNN) [59]. Next, the detected faces are cropped and scaled to 224×224 pixels. Notice that we do not employ an explicit facial alignment, but only centralize the detected faces without transforming their pose as shown in Fig. 2. In this way, not only inner facial dynamics but its relations with head pose can be jointly analyzed. For modeling, we employ the same CNN-LSTM architecture (initialized with pre-trained ResNet weights) used in ambient appearance subnetwork (see Section 3.1.1).

3.1.3. Voice subnetwork

The third modality used in the proposed model is voice. We first extract input features from the audio waveforms, which are then fed to our voice network. Particularly, we first compute a log mel-scale spectrogram for the input audio, and using a sliding window, a sequence of successive non-overlapping frames of 960 milliseconds for each audio waveform [60] is generated. For the time/frequency decomposition, a short-time Fourier transform with a step size of 10 milliseconds is applied on 25 milliseconds windows. Using 64 mel-spaced frequency bins, spectrogram frames with a size of 96×64 pixels are obtained.

These 2D log mel-scale spectrogram frames are fed to a CNN to obtain high-level embeddings. Based on our preliminary experiments, we opt for using VGGish architecture [61] as the backbone of our voice subnetwork. For a "warm start", we initialize the VGGish model with the weights learned on a large YouTube dataset that is a preliminary version of YouTube-8M [62]. The VGGish model outputs 128-dimensional embeddings for each log mel-scale spectrogram frame, which is connected to a six layered LSTM network including residual connections and dropout layers. The only differences between the employed LSTM network and the one described in Section 3.1.1, are the dimensions of the hidden state of the LSTM cells, which are set as 128 for the first three layers and 64 for the last three layers. Then, a fully-connected layer with five neurons (used as a regressor to estimate the personality traits) is connected to the output of the sixth LSTM layer of the last time step.

3.1.4. Speech transcription subnetwork

The last modality used in the proposed method is the transcription of the speech of people in the videos. Psychological research has shown that personality influences the way a person writes or talks and word use and expressions are associated with personality [63]. For example, individuals that score high in extraversion prefer complex, long writings and conscientious people tend to talk more about achievements and work [64]. These studies indicate that people with similar personality factors are likely to use the same words and choose similar sentiment expressions. Therefore, this information must be analyzed to make an accurate prediction of personality traits.

To analyze transcribed speech, we first encode the text into high dimensional representation vectors. To this end, we employ a state-of-the-art method, namely Embeddings from Language Models (ELMo) [65]. ELMo computes contextualized word representations using deep

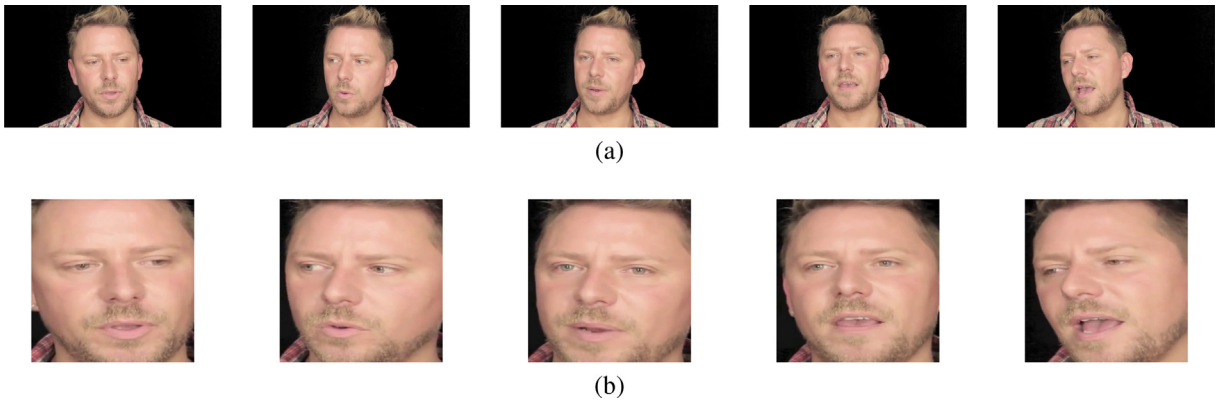

Fig. 2. (a) Sample input frames, and (b) the corresponding cropped/centralized faces.

bidirectional LSTM units, which is trained on one billion word benchmark [66]. In our subnetwork, we use the pre-trained weights of ELMo. This model outputs a 1024-dimensional vector containing a fixed mean-pooling of all contextualized word representations. In our network, ELMo (with frozen weights) is followed by three fully-connected layers with 256, 128, 64 neurons, respectively. ReLU activation is applied after each of these layers. At the last hidden layer, dropout is used. As in the aforementioned modality subnetworks, an additional fully-connected layer with five neurons is used as a score regressor for the five personality traits.

Notice that unlike all other subnetworks, no LSTM layer or any other variation of Recurrent Neural Networks (RNN) is employed for modeling the transcribed speech since the information related to the sequences of words is already encoded into the embeddings through the bidirectional LSTM units in ELMo.

3.1.5. Modality-specific modeling

In the first stage of modeling, each of the aforementioned subnetworks are separately trained through minimizing average Mean Absolute Error (MAE) of predicting the five personality traits, which is defined as:

$$\ell_{\text{MAE}} = \frac{1}{5n}\sum_{i}^{n}\sum_{j\in S}\left|y_i^j - \hat{y}_i^j\right|, \tag{1}$$

where, $y_i^j, \hat{y}_i^j$ denote the actual and predicted personality scores of the ith subject in terms trait j, respectively. S is the set of employed personality traits, i.e., $S = \{\text{Openness, Conscientiousness, Extraversion Agreeableness, Neuroticism}\}$. n represents the number of training samples. Adam optimizer is employed for training each of the modality-specific networks, where the learning rate is chosen based on minimum validation error.

3.2. Multimodal fusion & attention-based modeling

Once all modality-specific networks are trained individually (pre-trained), we freeze their weights and remove the regression layer (the last fully-connected layer) of each subnetwork. Representation vectors of the four modalities, obtained before the corresponding regression layers, are concatenated to form a multimodal representation vector and fed to an attention module. With the proposed module, we aim to capture the feature importance based on the complex relations between features. Feature attention module is composed of two fully-connected layers, and computes an attention weight for each dimension of the (concatenated) multimodal representation. Let the obtained multimodal representation vector be $\mathbf{F} \in \mathbb{R}^{256}$, then the attention weight vector for $\mathbf{F}$, namely, $\mathbf{a}$ can be computed as:

$$\mathbf{a} = \tanh\left(\mathbf{V}\tanh\left(\mathbf{W}\mathbf{F} + \mathbf{b}\right) + \mathbf{c}\right), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{d\times 256}$ denotes the transformation matrix and $\mathbf{b}$ is the bias term for the first fully-connected layer. $\mathbf{V} \in \mathbb{R}^{256\times d}$ is the weight matrix of the second fully-connected layer. Here d denotes the dimension of hidden representation and it is set to 64. Notice that the use of *tanh* constraints the attention weight to lie in the interval $[-1, 1]$. The obtained attention vector a is then used to re-weight each dimension of the multimodal representation vector $\mathbf{F}$ by an element-wise multiplication as follows:

$$\hat{\mathbf{F}} = \mathbf{F}\odot\mathbf{a}, \tag{3}$$

where, $\hat{\mathbf{F}}$ is the output of the feature attention module and it is fed to a final fully-connected layer with one neuron, which acts as a regressor for predicting the score of a personality trait. For each of the five personality traits, we employ a separate set of attention module and regressor (see Fig. 1(b)), so as to effectively learn trait-specific weights.

To train the resulting multimodal model, we propose a MAE-based loss function that includes an additional term for error consistency that enforces trait-specific errors to be similar with each other. Similar to the Eq. (1), let ℓ_{MAE}^j denote the MAE for the prediction of the jth personality trait as:

$$\ell_{\text{MAE}}^j = \frac{1}{n}\sum_{i}^{n}\left|y_i^j - \hat{y}_i^j\right|. \tag{4}$$

Then, our multimodal loss function can be defined as:

$$\mathcal{L} = \ell_{\text{MAE}} + \frac{1}{2}\sqrt{\sum_{j\in S}\left(\ell_{\text{MAE}}^j - \ell_{\text{MAE}}\right)^2}. \tag{5}$$

Recall that ℓ_{MAE} is the average MAE (see Eq. (1)) and S denotes the set of five personality traits. In Eq. (5), the second term enforces the error consistency and it is defined as the standard deviation of MAEs for the five personality traits. In this way, an effective regularization can be achieved. In other words, the error consistency term prevents having high levels of error for some personality traits while the model overfits to other traits. During training, Adam optimizer is employed, and we determine the learning rate based on minimum validation error.

4. Experimental results and evaluation

In this section, we present the experiments that are carried out for the proposed method, the dataset which the model is trained on, and the experimental results. The proposed approach and various other alternatives are experimented with and compared to each other, and the best performing method is compared to the state-of-the-art. The results demonstrate that the proposed method outperforms the current state-of-the-art.

4.1. Dataset

To assess the reliability and accuracy of our proposed method, we employ the ChaLearn First Impressions V2 (CVPR'17) challenge dataset [14], which is publicly available [67]. This challenge aims to automatically recognize apparent personality traits according to the five-factor model. The dataset for this challenge consists of 10,000 videos of people facing and speaking to a camera. Videos are extracted from YouTube, they are mostly in high-definition (1280×720 pixels), and in general, they have an average duration of 15 seconds with 30 frames per second. In the videos, people talk to the camera in a self-presentation context and there is a diversity in terms of age, ethnicity, gender, and nationality. The videos are labeled with personality factors using Amazon Mechanical Turk (AMT), so the ground truth values are obtained by using human judgment. The database has predefined training, validation, and test sets with 6000, 2000, and 2000 videos, respectively. Fig. 3 shows some examples of videos.

In the data collection process, AMT workers have compared pairs of videos and evaluated the personality factors of people in the videos by choosing which person is likely to have more of an attribute than the other person for each personality factor [14]. Multiple votes per video, pairwise comparisons, and labeling small batches of videos have been used to address the problem of bias for the labels. The final scores have been obtained from the pairwise scores by using a Bradley-Terry-Luce (BTL) model [68], while addressing the problem of calibration of workers and worker bias [69]. While the level of five personality traits are defined by seven-point scale scores, the provided personality trait scores are the normalized ones, in the range of [0, 1].

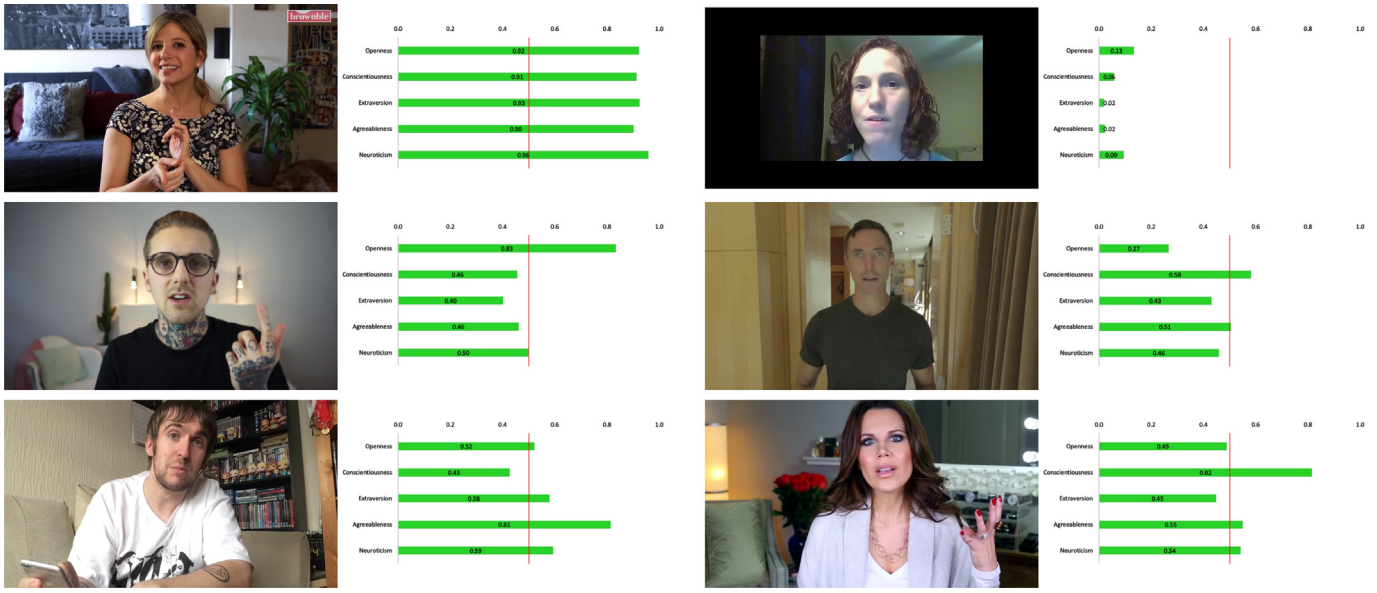

Fig. 3. Sample videos from the the ChaLearn First Impressions dataset depicting various cases of how personality traits are perceived by humans.

4.2. Experimental setup

The ChaLearn First Impressions V2 dataset is employed to evaluate the proposed method. In our experiments, we use the predefined training, validation, and test sets of the dataset. Adam optimizer is used for training the models, where its learning rate is chosen on the validation set. Particularly, the learning rate is optimized in the range of $[10^{-3}, 10^{-6}]$. During our experiments, best learning rates (with the minimum validation errors) are found as 10^{-4} and 10^{-5} for voice subnetwork and other subnetworks, respectively. Due to the computational complexity (based on the large sample size of the dataset), we use only the first six seconds of the videos (while videos can last up to 15 seconds), where the whole duration of the audio is employed. A minibatch size of 8 videos is used. The dropout probability is set to 0.5.

For ambient appearance and facial appearance subnetworks, we perform data augmentation during training, i.e. adjusting the brightness, saturation, hue, and contrast of RGB images by random factors. Besides, the RGB values of pixels are scaled to the range of [0,1].

We evaluate our method in terms of the evaluation metric introduced in the ChaLearn First Impressions Challenge [14], namely, the accuracy. Particularly, it is defined as 1 − MAE. Notice that our model is trained to output continuous values for the five target personality traits in the range of [0,1]. These values are produced separately for each trait, therefore there are five predicted values to be evaluated/reported. Consequently, we report the mean accuracy, i.e., the average of the per-trait accuracy values, as a summary measure. In some experiments, we also provide the per-trait accuracies for a detailed analysis. In the experiments, test accuracies are reported, unless otherwise indicated.

4.3. Informativeness of different modalities

We employ four different modalities, namely, ambient appearance, facial appearance, voice, and transcribed speech for estimating apparent personality. First, we evaluate the performance of the sole use of these modalities and compare them with that of their joint use. In this experiment, we do not use feature attention or error consistency constraint since we focus on investigating the informativeness of different modalities without the influence of such methods. Particularly, we evaluate the modality-specific models (see Section 3.1.5). For a comparison with the joint use of these modalities, we remove the regression layer of each of these modality-specific models and concatenate the last layer outputs before the regression layer. The obtained representation vector is fed to a fully-connected layer with five neurons, which is employed as a regressor for estimating the level of the five personality traits. This multimodal model is initialized with the weights learned from modality-specific training, and re-trained (fine-tuned) to minimize the average MAE for the five traits. Obtained results are given in Table 2.

When we compare the results obtained from the modality-specific models, facial appearance performs best with a mean accuracy of 91.30%. The ambient appearance model closely follows it with a mean accuracy of 91.13%, where the transcribed speech model performs worst and can only reach a mean accuracy of 88.81%. When we analyze the trait-specific accuracies, it is seen that the facial appearance provides the best results for all traits, except for the conscientious where the ambient appearance performs best. For extraversion, both facial appearance and ambient appearance networks provide an accuracy of 91.46%. The lowest accuracy value for each trait is obtained using the transcribed speech model. While the voice modality achieves relatively higher performance compared to the transcribed speech, it cannot reach the accuracy obtained with the use of facial appearance or ambient appearance. These results are in line with the findings in the literature, showing that facial cues, such as appearance, expression, and also head pose, are quite important for personality analysis. It is also important to note that our ambient appearance network leverages information from facial cues since its input frames not only include the background in the videos but also the facial image. This may explain the relatively high performance of the ambient appearance model.

When we look at the performance of the joint use of four modalities, we can say that the multimodal modeling clearly outperforms the sole use of the aforementioned modalities in terms of both mean accuracy

Table 2
Performance of the sole use of the considered modalities and of their fusion in terms of accuracy. Please notice that the error consistency constraint and feature attention are not used in this experiment.

Modality	Mean	Open.	Cons.	Extr.	Agre.	Neur.
Ambient Ap.	0.9113	0.9100	0.9154	0.9146	0.9086	0.9081
Facial Ap.	0.9130	0.9124	0.9135	0.9146	0.9144	0.9101
Voice	0.9045	0.9046	0.9083	0.9066	0.9046	0.8985
Tr. Speech	0.8881	0.8806	0.8811	0.9007	0.8929	0.8852
All	0.9172	0.9156	0.9216	0.9199	0.9153	0.9138

Table 3
Prediction performances of using different combinations of modalities in terms of mean accuracy.

Modality	Accuracy
Ambient Ap. + Facial Ap.	0.9158
Ambient Ap. + Voice	0.9153
Ambient Ap. + Tr. Speech	0.9113
Facial Ap. + Voice	0.9163
Facial Ap. + Tr. Speech	0.9134
Voice + Tr. Speech	0.9046
Ambient Ap. + Facial Ap. + Voice	0.9171
Ambient Ap. + Facial Ap. + Tr. Speech	0.9160
Ambient Ap. + Voice + Tr. Speech	0.9155
Facial Ap. + Voice + Tr. Speech	0.9163
All	0.9172

Table 4
Influence of using error consistency constraint and feature attention in the proposed method in terms of accuracy.

Feature attention	Error consistency	Mean	Open.	Cons.	Extr.	Agre.	Neur.
✓	✓	0.9181	0.9163	0.9223	0.9202	0.9162	0.9153
✘	✓	0.9174	0.9163	0.9199	0.9198	0.9160	0.9151
✓	✘	0.9177	0.9156	0.9212	0.9200	0.9165	0.9151
✘	✘	0.9172	0.9150	0.9219	0.9196	0.9151	0.9143

Table 5
Relative reduction (%) in MAE using error consistency constraint and feature attention in the proposed method (compared to without using these techniques). Please notice that negative percentages show reduction in MAE, while positive percentages indicate vice versa.

Feature attention	Error consistency	Mean	Open.	Cons.	Extr.	Agre.	Neur.
✓	✓	−1.09	−1.53	−0.51	−0.75	−1.30	−1.17
✘	✓	−0.24	−1.53	2.56	−0.25	−1.06	−0.93
✓	✘	−0.60	−0.71	0.90	−0.50	−1.65	−0.93

and trait-specific accuracies. The fusion of modalities, even without using the proposed attention mechanism and error consistency constraint, achieves a mean accuracy of 91.72%, where the corresponding MAE is 4.8% (relative) less than that of the sole use of facial appearance. This finding suggests the importance of multimodal analysis for personality recognition.

4.4. Reliability of using different combinations of modalities

As the findings of our previous experiment show, the joint use of the four modalities provides additional information (compared to their sole use) yielding a better accuracy. On the other hand, some modalities may have redundant or noisy information and their usage in the model can degrade the optimal performance. To investigate the influence of interaction between modalities, we train models for all possible combinations of the considered four modalities, namely, ambient appearance, facial appearance, voice, and transcribed speech. To this end, we follow the same approach used in the previous experiment, where the representation vectors of the input modalities are concatenated and connected to a fully-connected layer with five neurons. While the initial parameters of the model are obtained from the modality-specific training, additional training is used to minimize the average MAE of the multimodal model. In this experiment, similar to the previous one, we discard the use of feature attention and error consistency constraint because we aim to purely analyze the influence of interaction between different modalities, on the prediction accuracy.

As shown in Table 3, the best (mean) prediction accuracy is achieved by using all modalities together. The minimal improvement is obtained by including the transcribed speech to the other three modalities with a rate of 0.1% (from 0.9171 to 0.9172), where the improvement on the validation set is 0.2%. While our results show that the transcribed speech has a marginal influence on the accuracy when it is included in the analysis, its contribution is consistent. When we combine two modalities, the best accuracy is obtained for the fusion of facial appearance and voice. Similarly, for the fusion of three modalities, the joint use of voice together with facial appearance and ambient appearance provides the best performance. All these findings are in line with the results presented for the individual use of modalities in Section 4.3. Besides, we could not observe any accuracy reduction in the model, when we include an additional modality. This finding is also valid for the results obtained on the validation set. Since none of the considered modalities cause a negative influence on the prediction accuracy, we will continue using all four modalities in the remainder of our experiments.

4.5. Influence of feature attention and error consistency constraint

One of the main contributions of this study is the joint use of feature attention and error consistency constraint for effective multimodal modeling. To evaluate the influence of them on prediction accuracy, we assess our model for all possible cases with and without using feature attention and error consistency constraint. To this end, four different models are trained and evaluated. The obtained mean accuracies for these models are given in Table 4. Using these results, we also compute the relative reduction in MAE for each case compared to the model that does not use the feature attention mechanism or the error consistency constraint (see Table 5).

As shown in Table 5, when we employ only the error consistency constraint in the model, the average MAE is reduced by 0.24%. The sole use of feature attention reduces the average MAE by 0.6%. If both of them are employed then the reduction rate in the average MAE reaches to 1.09%, yielding a mean accuracy of 91.81%. With the joint use of these methods, we obtain improved predictions for each of the traits. Yet, interestingly, if we use only one of the feature attention mechanism or the error consistency constraint, then the MAE for the conscientiousness trait increases with a rate of 2.56% and 0.9%, respectively. This outcome supports our choice of using these two methods

Table 6
The comparison of different methods from the literature. The best performance for each trait and for mean accuracy is boldfaced.

Method	Modalities	Mean	Open.	Cons.	Extr.	Agre.	Neur.
Proposed method	Facial Ap., Ambient Ap., Voice, Tr. Speech	**0.918**	0.916	**0.922**	0.920	**0.916**	**0.915**
Wei et al. (2018) [48]	Facial Ap., Voice	0.913	0.912	0.917	0.913	0.913	0.910
Kaya et al. (2017) [70]	Facial Ap., Ambient Ap., Voice	0.917	**0.917**	0.920	**0.921**	0.914	**0.915**
Güçlütürk et al. (2017) [51]	Facial Ap., Voice, Tr. Speech	0.912	0.911	0.915	0.911	0.911	0.910
Bekhouche et al. (2017) [71]	Facial Ap.	0.912	0.910	0.914	0.915	0.910	0.908
Güçlütürk et al. (2016) [72]	Facial Ap., Voice	0.911	0.911	0.914	0.911	0.910	0.909
Subramaniam et al. (2016) [73]	Facial Ap., Voice	0.912	0.911	0.914	0.916	0.911	0.910
Gürpınar et al. (2016) [74]	Facial Ap., Ambient Ap., Voice	0.913	0.914	0.915	0.918	0.907	0.911

together. In this way, while we can re-weight the feature values we also prevent our model to overfit to the characteristics of a specific modality. The highest improvement in (trait-specific) MAE is achieved for the openness trait with a rate of 1.53% with the joint use of these methods. The same amount of improvement for the openness trait is also provided with the sole use of the error consistency constraint. This finding indicates that the error consistency constraint helps our model to focus on the openness trait (and other traits) rather than overfitting to the conscientiousness patterns, yielding a better overall performance.

4.6. Comparison to other methods

In this section, we compare our proposed method to seven recent multimodal studies that provide results on the ChaLearn First Impressions dataset. As shown in Table 6, our method outperforms all the competitor methods with a mean accuracy of 91.8%. While the performance of our method seems to be close to that of the state-of-the-art method proposed by Kaya et al. [70], we provide a 1.2% reduction in average MAE. Besides, we need to note that Kaya et al. [70] include the validation set (33% more data) in the training set after optimizing the hyperparameters (on the validation set), yet, our model is trained only on the training set. On the other hand, our method employs an additional modality, i.e., transcribed speech, in comparison to Kaya et al. [70]. When the trait-specific results are analyzed, it is seen that our model achieves the highest accuracy for conscientiousness (92.2%) and agreeableness (91.6%), while Kaya et al. [70] get the best results for openness (91.7%) and extraversion (92.1%). For neuroticism, both our method and theirs achieve the highest accuracy with a rate of 91.5%. To sum up, although the state-of-the-art performance in personality analysis is quite high and leaves limited room for improvement, our proposed method employing feature attention mechanism and error consistency constraint provides a clear enhancement.

5. Conclusions and future work

We propose a novel multimodal approach for the estimation of apparent personality traits. Our method relies on four subnetworks, each of which focuses on a specific modality, namely ambient appearance, facial appearance, voice, and transcribed speech. These subnetworks employ state-of-the-art deep architectures (e.g., ResNet-v2-101, VGGish, ELMo) as backbones, and they are complemented with additional LSTM layers to leverage temporal information. For more effective modeling, first, each of the aforementioned subnetworks has been initialized with the (pre-trained) weight parameters of the corresponding backbone network and trained (fine-tuned) in a modality-specific manner. Then, these subnetworks (after removing their regression layers) have been combined and complemented by feature attention and regression layers. While the parameters of the subnetworks are kept frozen, new layers of the whole network have been trained to minimize the average MAE of predicting the five personality traits as well as keeping the errors for each modality as close as possible to each other using the proposed error consistency constraint. In this way, our model prevents overfitting to some specific traits due to joint multi-task optimization. Although the proposed architecture is end-to-end trainable, we have followed a hierarchical training to minimize computational costs while improving effectiveness.

Our framework has been thoroughly evaluated on the large-scale ChaLearn First Impressions dataset. The effectiveness and reliability of the proposed feature attention mechanism and the error consistency constraint have been systematically assessed. Besides, the informativeness of different modalities and the added value of their joint use have been investigated. Our results show that the proposed feature attention and error consistency constraint are indeed useful and improve prediction accuracy. With the use of ambient appearance, facial appearance, voice, and transcribed speech modalities, our proposed model achieves a mean accuracy of 91.8%, improving the state of the art. As future research directions, we envision that correlation between personality, body movements, posture, eye-gaze, and emotion can be investigated.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H.-R. Pfister, G. Böhm, The multiplicity of emotions: a framework of emotional functions in decision making, Judgm. Decis. Mak. 3 (1) (2008) 5.
[2] D. Kahneman, P. Egan, Thinking, Fast and Slow, vol. 1, Farrar, Straus and Giroux New York, 2011.
[3] M. Deniz, An investigation of decision making styles and the five-factor personality traits with respect to attachment styles, Educ. Sci. Theor. Pract. 11 (1) (2011) 105–113.
[4] T. Chamorro-Premuzic, A. Furnham, Personality and music: can traits explain how people use music in everyday life? Br. J. Psychol. 98 (2) (2007) 175–185.
[5] P.J. Rentfrow, S.D. Gosling, The do re mi's of everyday life: the structure and personality correlates of music preferences, J. Pers. Soc. Psychol. 84 (6) (2003) 1236.
[6] I. Cantador, I. Fernández-Tobas, A. Bellogín, Relating personality types with user preferences in multiple entertainment domains, CEUR Workshop Proceedings, Shlomo Berkovsky, 2013.
[7] R.W. Picard, Affective Computing, MIT Press, 2000.
[8] A. Vinciarelli, G. Mohammadi, A survey of personality computing, IEEE Trans. Affect. Comput. 5 (3) (2014) 273–291.
[9] D. Cervone, L. Pervin, Personality: Theory and Research, 12th ed. Wiley Global Education, 2013.
[10] R.R. McCrae, O.P. John, An introduction to the five-factor model and its applications, J. Pers. 60 (2) (1992) 175–215.
[11] R.R. McCrae, P.T. Costa, Validation of the five-factor model of personality across instruments and observers, J. Pers. Soc. Psychol. 52 (1) (1987) 81–90.
[12] E.C. Tupes, R.E. Christal, Recurrent personality factors based on trait ratings, J. Pers. 60 (2) (1992) 225–251, https://doi.org/10.1111/j.1467-6494.1992.tb00973.x.
[13] L. Goldberg, The structure of phenotypic personality traits, Am. Psychol. 48 (1) (1993) 26–34.
[14] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H.J. Escalante, S. Escalera, ChaLearn Lap, First round challenge on first impressions-dataset and results, Proceedings of the European Conference on Computer Vision, 2016, Springer 2016, pp. 400–418.
[15] C. Nass, K.M. Lee, Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction, J. Exp. Psychol. Appl. 7 (3) (2001) 171.
[16] A. Tlili, F. Essalmi, M. Jemni, N.-S. Chen Kinshuk, Role of personality in computer based learning, Comput. Hum. Behav. 64 (2016) 805–813.
[17] F. Durupnar, N. Pelechano, J.M. Allbeck, U. Güdükbay, N.I. Badler, How the Ocean personality model affects the perception of crowds, IEEE Comput. Graph. Appl. 31 (3) (2011) 22–31.
[18] A. Bera, T. Randhavane, D. Manocha, Aggressive, tense or shy? identifying personality traits from crowd videos, Proceedings of the International Joint Conference on Artificial Intelligence 2017, pp. 112–118, https://doi.org/10.24963/ijcai.2017/17.
[19] A.E. Başak, U. Güdükbay, F. Durupınar, Using real life incidents for creating realistic virtual crowds with data-driven emotion contagion, Comput. Graph. 72 (2018) 70–81.
[20] M. Tkalčič, B. De Carolis, M. de Gemmis, A. Odić, A. Košir, Introduction to emotions and personality in personalized systems, Emotions and Personality in Personalized Services, Springer 2016, pp. 3–11.
[21] L. Shen, M. Wang, R. Shen, Affective e-learning: using "emotional" data to improve learning in pervasive learning environment, J. Educ. Technol. Soc. 12 (2) (2009) 176–189.
[22] G. Ball, J. Breese, Emotion and personality in a conversational agent, Embodied Convers. Agents (2000) 189–219.
[23] F. Durupınar, U. Güdükbay, A. Aman, N.I. Badler, Psychological parameters for crowd simulation: from audiences to mobs, IEEE Trans. Visual. Comput. Graph. 22 (9) (2016) 2145–2159.
[24] J.A. Recio-Garcia, G. Jimenez-Diaz, A.A. Sanchez-Ruiz, B. Diaz-Agudo, Personality aware recommendations to groups, Proceedings of the ACM Conference on Recommender Systems, ACM 2009, pp. 325–328.
[25] R. Hu, P. Pu, A study on user perception of personality-based recommender systems, Proceedings of the International Conference on User Modeling, Adaptation, and Personalization, Springer 2010, pp. 291–302.
[26] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, J.M. Jose, Integrating facial expressions into user profiling for the improvement of a multimodal recommender system, Proceedings of the IEEE International Conference on Multimedia and Expo, IEEE 2009, pp. 1440–1443.
[27] R. Subramanian, J. Wache, M.K. Abadi, R.L. Vieriu, S. Winkler, N. Sebe, ASCERTAIN: emotion and personality recognition using commercial sensors, IEEE Trans. Affect. Comput. 9 (2) (2018) 147–160, https://doi.org/10.1109/TAFFC.2016.2625250.
[28] F. Valente, S. Kim, P. Motlicek, Annotation and recognition of personality traits in spoken conversations from the AMI Meetings Corpus, Proceedings of the Annual

Conference of the International Speech Communication Association, 2012, INTERSPEECH 2012, pp. 1183–1186.
[29] N.A. Madzlan, J. Han, F. Bonin, N. Campbell, Automatic recognition of attitudes in video blogs—prosodic and visual feature analysis, Proceedings of the Annual Conference of the International Speech Communication Association, 2014, INTERSPEECH 2014, pp. 1826–1830.
[30] F. Alam, E.A. Stepanov, G. Riccardi, Personality traits recognition on social network - Facebook, Proceedings of the International AAAI Conference on Web and Social Media, Cambridge, MA 2013, pp. 6–9 , aAAI Technical Report WS-13-01 Computational Personality Recognition (Shared Task).
[31] S. Nowson, A.J. Gill, Look! who's talking?: Projection of extraversion across different social contexts, Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition 2014, pp. 23–26.
[32] G. Farnadi, G. Sitaraman, S. Sushmita, F. Celli, M. Kosinski, D. Stillwell, S. Davalos, M.-F. Moens, M. De Cock, Computational personality recognition in social media, User Modeling and User-Adapted Interaction, 26, , 02 2016https://doi.org/10.1007/s11257-016-9171-0.
[33] G. Cucurull, P. Rodríguez, V.O. Yazici, J.M. Gonfaus, F.X. Roca, J. Gonzàlez, Deep Inference of Personality Traits by integrating Image and Word use in Social Networks, arXiv preprint arXiv:1802.06757 2018.
[34] C. Segalin, D.S. Cheng, M. Cristani, Social profiling through image understanding: personality inference using convolutional neural networks, Comput. Vis. Image Underst. 156 (2017) 34–50.
[35] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, M. Zancanaro, Multimodal recognition of personality traits in social interactions, Proceedings of the International Conference on Multimodal interfaces, ACM 2008, pp. 53–60.
[36] L.M. Batrinca, N. Mana, B. Lepri, F. Pianesi, N. Sebe, Please, tell me about yourself: Automatic personality assessment using short self-presentations, Proceedings of the International Conference on Multimodal Interfaces, ACM 2011, pp. 255–262.
[37] L. Batrinca, N. Mana, B. Lepri, N. Sebe, F. Pianesi, Multimodal personality recognition in collaborative goal-oriented tasks, IEEE Trans. Multimed. 18 (4) (2016) 659–673, https://doi.org/10.1109/TMM.2016.2522763.
[38] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.
[39] R.J. Vernon, C.A. Sutherland, A.W. Young, T. Hartley, Modeling first impressions from highly variable facial images, Proc. Natl. Acad. Sci. 111 (32) (2014) E3353–E3361.
[40] R. Qin, W. Gao, H. Xu, Z. Hu, Modern physiognomy: an investigation on predicting personality traits and intelligence from the human face, SCIENCE CHINA Inf. Sci. 61 (5) (2018), 058105, .
[41] F. Gürpınar, H. Kaya, A.A. Salah, Combining deep facial and ambient features for first impression estimation, Proceedings of the European Conference on Computer Vision, Springer 2016, pp. 372–385.
[42] K. Ilmini, T. Fernando, Personsâ€™ personality traits recognition using machine learning algorithms and image processing techniques, Adv. Comput. Sci. 5 (1) (2016) 40–44.
[43] Y. Yan, J. Nie, L. Huang, Z. Li, Q. Cao, Z. Wei, Exploring relationship between face and trustworthy impression using mid-level facial features, International Conference on Multimedia Modeling, Springer 2016, pp. 540–549.
[44] C. Ventura, D. Masip, A. Lapedriza, Interpreting CNN models for apparent personality trait regression, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017, pp. 1705–1713, https://doi.org/10.1109/CVPRW.2017.217.
[45] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using cnn-rnn and c3d hybrid networks, Proceedings of the ACM International Conference on Multimodal Interaction, ACM 2016, pp. 445–450.
[46] J. Lin, W. Mao, D.D. Zeng, Personality-based refinement for sentiment classification in microblog, Knowl.-Based Syst. 132 (2017) 204–214.
[47] C.-L. Zhang, H. Zhang, X.-S. Wei, J. Wu, Deep bimodal regression for apparent personality analysis, in: G. Hua, H. Jégou (Eds.),Computer Vision – ECCV 2016 Workshops 2016, pp. 311–324.
[48] X. Wei, C. Zhang, H. Zhang, J. Wu, Deep bimodal regression of apparent personality traits from short video sequences, IEEE Trans. Affect. Comput. 9 (3) (2018) 303–315, https://doi.org/10.1109/TAFFC.2017.2762299.
[49] C. Sarkar, S. Bhatia, A. Agarwal, J. Li, Feature analysis for computational personality recognition using youtube personality data set, Proceedings of the ACM Multi Media on Workshop on Computational Personality Recognition, ACM 2014, pp. 11–14.
[50] J.C. Silveira Jacques Junior, Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H.J. Escalante, I. Guyon, M.A. Van Gerven, R. Van Lier, S. Escalera, First impressions: a survey on vision-based apparent personality trait analysis, IEEE Trans. Affect. Comput. (2019)https://doi.org/10.1109/TAFFC.2019.2930058.
[51] Y. Güçlütürk, U. Güçlü, X. Baro, H.J. Escalante, I. Guyon, S. Escalera, M.A. Van Gerven, R. Van Lier, Multimodal first impression analysis with deep residual networks, IEEE Trans. Affect. Comput. 9 (3) (2017) 316–329.
[52] H.J. Escalante, H. Kaya, A.A. Salah, S. Escalera, Y. Güç, U. Güçlü, X. Baró, I. Guyon, J.C.S. Jacques, M. Madadi, S. Ayache, E. Viegas, F. Gürpınar, A.S. Wicaksana, C. Liem, M.A. Van Gerven, R. Van Lier, Modeling, recognizing, and explaining apparent personality from videos, IEEE Trans. Affect. Comput. (2020)https://doi.org/10.1109/TAFFC.2020.2973984.
[53] O. Çeliktutan, H. Gunes, Automatic prediction of impressions in time and across varying context: personality, attractiveness and likeability, IEEE Trans. Affect. Comput. 8 (1) (2017) 29–42.
[54] J.V. Kasmar, W.V. Griffin, J.H. Mauritzen, Effect of environmental surroundings on outpatients' mood and perception of psychiatrists, J. Consult. Clin. Psychol. 32 (2) (1968) 223.
[55] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, Proceedings of the European Conference on Computer Vision, Springer 2016, pp. 630–645.
[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[57] A.C. Little, D.I. Perrett, Using composite images to assess accuracy in personality attribution to faces, Br. J. Psychol. 98 (1) (2007) 111–126.
[58] B. Fink, N. Neave, J.T. Manning, K. Grammer, Facial symmetry and the 'big-five' personality factors, Personal. Individ. Differ. 39 (3) (2005) 523–529.
[59] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Sig. Process. Lett. 23 (10) (2016) 1499–1503.
[60] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, Q. Tian, Hmm-based audio keyword generation, in: K. Aizawa, Y. Nakamura, S. Satoh (Eds.), Advances in Multimedia Information Processing - PCM 2004, Springer, Berlin Heidelberg, Berlin, Heidelberg 2005, pp. 566–574.
[61] S. Hershey, S. Chaudhuri, D.P. Ellis, J.F. Gemmeke, A. Jansen, R.C. Moore, M. Plakal, D. Platt, R.A. Saurous, B. Seybold, M. Slaney, R.J. Weiss, K. Wilson, CNN architectures for large-scale audio classification, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE 2017, pp. 131–135.
[62] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, S. Vijayanarasimhan, YouTube-8M: A Large-Scale Video Classification Benchmark, CoRR abs/1609.08675 2016.
[63] G. Stemmler, J. Wacker, Personality, emotion, and individual differences in physiological responses, Biol. Psychol. 84 (3) (2010) 541–551.
[64] M.R. Mehl, S.D. Gosling, J.W. Pennebaker, Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life, J. Pers. Soc. Psychol. 90 (5) (2006) 862.
[65] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics, Human Language Technologies 2018, pp. 2227–2237.
[66] C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, T. Robinson, One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling, CoRR abs/1312.3005 2013.
[67] ChaLearn, First Impressions V2 (CVPR'17) Dataset, http://chalearnlap.cvc.uab.es/dataset/24/description, accessed: 2019-04-30 2017.
[68] R.A. Bradley, M.E. Terry, Rank analysis of incomplete block designs: I. the method of paired comparisons, Biometrika 39 (3/4) (1952) 324–345.
[69] B. Chen, S. Escalera, I. Guyon, V. Ponce-López, N. Shah, M.O. Simón, Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits, Proceedings of the European Conference on Computer Vision, Springer 2016, pp. 419–432.
[70] H. Kaya, F. Gurpinar, A. Ali Salah, Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017, pp. 1–9.
[71] S. Eddine Bekhouche, F. Dornaika, A. Ouafi, A. Taleb-Ahmed, Personality traits and job candidate screening via analyzing facial videos, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops 2017, pp. 10–13.
[72] Y. Güçlütürk, U. Güçlü, M.A. van Gerven, R. van Lier, Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition, Proceedings of the European Conference on Computer Vision, Springer 2016, pp. 349–358.
[73] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, A. Mittal, Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features, Proceedings of the European Conference on Computer Vision, Springer 2016, pp. 337–348.
[74] F. Gürpinar, H. Kaya, A.A. Salah, Multimodal fusion of audio, scene, and face features for first impression estimation, Proceedings of the International Conference on Pattern Recognition, IEEE 2016, pp. 43–48.