Multimodal Personality Trait Analysis for Explainable Modeling of Job Interview Decisions

Heysem Kaya and Albert Ali Salah[†]

Abstract Automatic analysis of job interview screening decisions is useful for establishing the nature of biases that may play a role in such decisions. In particular, assessment of apparent personality gives insights into the first impressions evoked by a candidate. Such analysis tools can be used for training purposes, if they can be configured to provide appropriate and clear feedback. In this chapter, we describe a multimodal system that analyzes a short video of a job candidate, producing apparent personality scores and a prediction about whether the candidate will be invited for a further job interview or not. This system provides a visual and textual explanation about its decision, and was ranked first in the ChaLearn 2017 Job Candidate Screening Competition. We discuss the application scenario and the considerations from a broad perspective.

Key words: explainable machine learning, job candidate screening, multimodal affective computing, personality trait analysis

---

[†] This is the uncorrected author proof. Please cite as: Kaya, H., A.A. Salah, Multimodal Personality Trait Analysis for Explainable Modeling of Job Interview Decisions, H. J. Escalante et al. (eds.), Explainable and Interpretable Models in Computer Vision and Machine Learning, The Springer Series on Challenges in Machine Learning, Springer Nature Switzerland, Pages 255-275, 2018.

Heysem Kaya
Department of Computer Engineering, Namik Kemal University, Corlu, Tekirdag, TURKEY
e-mail: hkaya@nku.edu.tr

Albert Ali Salah
Department of Computer Engineering, Bogazici University, Istanbul, TURKEY
Future Value Creation Research Center, Nagoya University, Nagoya, JAPAN
e-mail: salah@boun.edu.tr

1 Introduction

Affective and social computing applications aim to realize computer systems that are responsive to social signals of people they interact with. Under this research program, we find robots and virtual agents that engage their users in affect-sensitive interactions, educational monitoring tools, systems that track user behavior for improved prediction capabilities and better services. With the increase of real-time capabilities of such systems, new application areas are becoming feasible, and such technologies are becoming more widespread. It is not uncommon now to have a camera that automatically takes a picture when the people in the frame are smiling. Yet with more widespread use, and more integration of such smart algorithms, there arises the need to design accountable systems that explain their decisions to their users, particularly for cases where these decisions have a major impact on the lives and wellbeing of other people. In this chapter, we describe one such application scenario, and discuss related issues within the context of a solution we have developed for this specific case.

Job interviews are one of the primary assessment tools for evaluating job seekers, and for many corporations and institutions, an essential part of the job candidate selection process. These relatively short interactions with individuals have potentially life-changing impact for the job seekers. In 2017, the ChaLearn Job Candidate Screening (JCS) Competition[4] was organized at CVPR, to investigate the value of automatic recommendation systems based on multimedia CVs (Escalante et al, 2017).

A system that can analyze a short video of the candidate to predict whether the candidate will be invited to a job interview or not is valuable for multiple reasons. For the recruiter, it can help visualize the biases in candidate selection, and assist in the training of the recruitment staff. For the job seeker, it can be a valuable tool to show what impression the candidate is giving to the recruiter, and if properly designed, could even suggest improvements in attitude, speaking style, posture, gaze behavior, attire, and such.

At this point, we caution the reader. It may be tempting to use such a system to automatically screen candidates when the job application figures are overwhelming. If the system approximates the human recruiter's behavior sufficiently well, it may even have a result very similar to the human recruiter's selection. However, what the system is evaluating is the first impression caused by the candidate, and this is not a sound basis to judge actual job performance. For example, overweight people are shown to be more negatively rated in job interviews compared to people with average weight, and were seen as less desirable, "less competent, less productive, not industrious, disorganized, indecisive, inactive, and less successful" (Larkin and Pines, 1979). These stereotypes that the human recruiters have will be learned by the automatic system that relies on human annotations for its supervision. Subsequently, the system will also exhibit such biases. Therefore, it is necessary both to investigate

[4] It was officially called a co-opetition, as it promoted sharing code and results between participants.

any systematic biases in the system, and to design mechanisms where the system gives an explanation about its particular decision, by looking at its own decision process. This resembles endowing the system with a meta-cognitive module. If the output of such a module can be fed back into the system for removing biases in its learning, we will be on our way for much smarter systems.

In this chapter, we first report some related work on apparent personality estimation and evaluation of video resumes for job interviews. We describe the Job Candidate Screening Challenge briefly, and then describe an end-to-end system that officially participated in the Challenge. We report our experimental results, and then investigate both the biases inherent in the annotations, and in the ensuing system. We also describe the meta-cognitive part of the system, namely, the module that explains its decisions. We discuss our findings, the contributions of the challenge to our understanding of the problem, our shortcomings, and what the future looks like for this research area.

2 Related Work

From a physchological perspective, personality is observed as a long term summary of behaviors, having a complex structure that is shaped by many factors such as habits and values. Analysis of personality is difficult, and requires psychological testing on the subject for obtaining a ground truth. Researchers in the field also analyze the "apparent personality," i.e. the *impressions* a subject leaves on other people (the annotators), instead of the actual personality (Gürpınar et al, 2016b; Lopez et al, 2016; Junior et al, 2018). This is easier to annotate, as only external evaluations are required for annotations, and the actual subject is not involved. Both real and apparent personality are typically assessed along the "Big Five" personality traits, namely, Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (commonly abbreviated as OCEAN), respectively (Valente et al, 2012).

Modeling and predicting apparent personality is studied from different modalities, particulary speech acoustics (Schuller et al, 2012; Valente et al, 2012; Madzlan et al, 2014), linguistics (Alam et al, 2013; Gievska and Koroveshovski, 2014; Nowson and Gill, 2014) and visual input (Fernando et al, 2016; Qin et al, 2016). In the literature, short segments of audio or video are used for automatic predictions (Kaya and Salah, 2014; Celiktutan and Gunes, 2016). Furthermore, multimodal systems that benefit from complementary information are increasingly studied (Alam and Riccardi, 2014; Farnadi et al, 2014; Sarkar et al, 2014; Sidorov et al, 2014; Gürpınar et al, 2016a; Barezi et al, 2018).

Deep learning based classifiers have been shown to work well for predicting apparent personality ratings from visual input (Lopez et al, 2016; Zhang et al, 2016; Güçlütürk et al, 2016, 2017; Kaya et al, 2017a; Escalante et al, 2018; Barezi et al, 2018). However, the need for large amounts of training data and high mem-

ory/computational complexity of training deep network models are some of the disadvantages for deep learning based methods.

Deep learning models for personality analysis typically look at the face or the facial behavior of the person to determine what stereotypes it will activate in the viewers. An advantage of deep neural networks for analysing facial images is that the earlier layers of the network learn good internal representations for faces, regardless of the facial analysis task targeted by the supervised learning process. Since it is relatively easy to collect large amounts of face images together with identity labels (e.g. famous persons) from the Internet, it is possible to train a deep neural network for a face recognition task with millions of samples. Once this is done, the resulting deep (convolutional) neural network can serve as a pre-trained model to enable efficient and effective transfer learning on other tasks, such as emotional expression recognition (Kaya et al, 2017b).

There are different approaches to transfer learning (Pan and Yang, 2010). The approach we use in this work is one where we start from a model pre-trained with a very large database, and fine-tune the model for a different task using a smaller database. This approach is ideal if the there are not sufficiently many samples for training a deep model in the target task, but when the task shares structural similarities (i.e. analysis of faces in our case) with a task that does have such large data for training (e.g. face recognition).

3 Job Candidate Screening Challenge

The CVPR 2017 Job Candidate Screening Challenge was organized to help both recruiters and job candidates using multi-media CVs (Escalante et al, 2017). The challenge relied on a publicly available dataset[5] that contains more than 10 000 clips (average duration 15 seconds) from more than 3 000 videos collected from YouTube (Escalante et al, 2016). These are annotated via Amazon Mechanical Turk annotators for apparent personality traits, as well as a variable that measured whether the candidate would be invited to a job interview, or not. Basic statistics of the dataset partitions are provided in Table 1. The detailed information on the Challenge and the corpus can be found in (Lopez et al, 2016).

Table 1: Dataset summary

	Train	**Val**	**Test**
#Clips	6,000	2,000	2,000
#YouTube videos	2,624	1,484	1,455
#Given frames	2.56M	0.86M	0.86M
#Detected frames	2.45M	0.82M	0.82M

[5] The dataset can be obtained from `http://chalearnlap.cvc.uab.es/dataset/24/description/`

The apparent personality annotations were made through a single question asked per dimension. The annotators saw a pair of candidates, and assigned an attribute to one of the videos (with an option of not assigning it to any video). The attributes used to measure the "Big Five" personality traits were as follows: Friendly vs. Reserved (for Extraversion), Authentic vs. Self-interested (for Agreeableness), Organized vs. Sloppy (for Conscientiousness), Uneasy vs. Comfortable (for Neuroticism), Imaginative vs. Practical (for Openness to Experience). Previously, the ChaLearn Looking at People 2016 First Impression Challenge was organized to develop systems that can predict these apparent personality ratings (Lopez et al, 2016). Additionally, the question of "Who would you rather invite for a job interview?" was posed to obtain a ground truth for the job candidate screening task. These annotations were post-processed to produce cardinal scores for each clip (Escalante et al, 2018).

The Challenge itself was composed of two stages: a quantitative challenge to predict the "invite for interview" variable, and a qualitative challenge to justify the decision with verbal/visual explanations, respectively. The participants were encouraged to use the personality trait dimensions in prediction (quantitative) and explanation (qualitative) stages.

4 Proposed Method

The prediction problem we focus in this paper is based on assessing a short input video for the "Big Five" personality traits and the "invite for interview" variable. The available modalities for analysis include the facial image of the candidate, the acoustics of his or her voice, and the features that can be extracted from the background, which we call the scene. Inspired from the winning system of ICPR 2016 ChaLearn Apparent Personality Challenge that was organized with the same corpus and protocol (Gürpınar et al, 2016b), we implement a multimodal system that evaluates audio, scene, and facial features as separate channels, and use Extreme Learning Machine classifiers to produce intermediate results for each channel. These first-level predictions are then combined in a second modeling stage to produce the final predictions.

The second stage of the competition required the submitted systems to produce explanations for the decisions of the system. It is possible to investigate the system dynamics, the learned features, the weights of the individual classifiers in the system, etc., and follow the path of a decision from the input to the output. This would generate a lot of information, and might make interpretation difficult. We choose a simple approach, where the first-level predictions are treated as a black-box, and no insights are generated for these predictions. However, the final prediction, which is based on the intermediate apparent personality trait estimations of the system, is generated with a tree-based classifier to enable the generation of an explanation. We describe all the components of this system in this section.

The pipeline of the proposed system for the quantitative challenge is illustrated in Figure 1. The input is represented on the left hand side, which consists of a video and its associated audio track. The face is detected, and two sets of features are extracted from the facial image. These are combined via feature-level fusion in the first kernel ELM classifier in the Modeling part. The scene features and the audio features are combined in another, similar classifier. On the right hand side, there is a stacked random forest classifier to give the final predictions, and it is this classifier that the system uses to generate an explanation about its behavior.

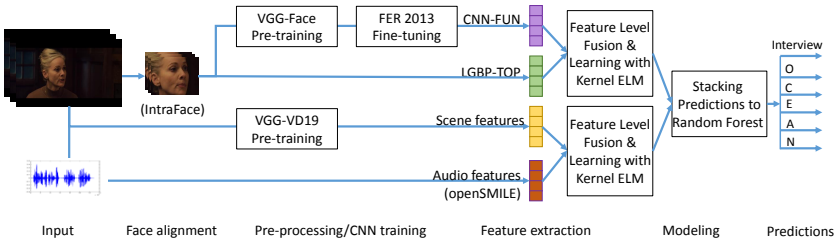


Fig. 1: Flowchart of the proposed method (Kaya et al, 2017a).

We now briefly describe the main steps of our pipeline, namely, face alignment, feature extraction, and modeling, respectively. We refer the reader to (Gürpınar et al, 2016b; Kaya et al, 2017a) for more technical details.

4.1 Visual Feature Extraction

The system detects faces, and locates 49 facial landmarks on each faces using the Supervised Descent Method (SDM) (Xiong and De la Torre, 2013). These points are extremely important to align facial images, so that a comparative analysis can be performed. The roll angle of the face is estimated from the eye corners to normalize the facial image. The distance between the two eyes is called the *interocular distance*, and it is frequently used to normalize the scale of the facial image. Our system adds a margin of 20% of the interocular distance around the outer landmarks to crop the facial image. Each such cropped image is resized to 64×64 pixels. These images are processed in two ways.

The first way uses a deep neural network. We start with the pre-trained VGG-Face network (Parkhi et al, 2015), which is optimized for the face recognition task on a very large set of faces. We change the final layer (originally a 2 622-dimensional recognition layer), to a 7-dimensional emotion recognition layer, where the weights are initialized randomly. We then fine-tune this network with the softmax loss function using more than 30K training images of the FER-2013 dataset (Goodfellow

et al, 2013). We choose an initial learning rate of 0.0001, a momentum of 0.9 and a batch size of 64. We train the model only for 5 epochs. The final, trained network has a 37-layer architecture (involving 16 convolution layers and 5 pooling layers). The response of the 33^{rd} layer is used in this work, which is the lowest-level 4 096-dimensional descriptor.

We combine deep facial features with a second set of features. We use a spatio-temporal descriptor called Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) (Almaev and Valstar, 2013) that is shown to be effective in emotion recognition (Kaya et al, 2017b). The LGBP-TOP descriptor is extracted by applying 18 Gabor filters on aligned facial images with varying orientation and scale parameters. The resulting feature dimensionality is 50 112.

Facial features are extracted over an entire video segment and summarized by functionals. The functionals include mean, standard deviation, offset, slope, and curvature. Offset and slope are calculated from the first order polynomial fit to each feature contour, while curvature is the leading coefficient of the second order polynomial. Scene features, however, are extracted from the first image of each video only. The assumption is that videos do not stretch over multiple shots.

In order to use the ambient information in the images to our advantage, we extract a set of features using the VGG-VD-19 network (Simonyan and Zisserman, 2014), which is trained for an object recognition task on the ILSVRC 2012 dataset. Similar to face features, we use the 4 096-dimensional feature from the 39^{th} layer of the 43-layer architecture, hence we obtain a description of the overall image that contains both face and scene. Using a deep neural network originally trained for an object recognition task basically serves to detect high level objects and object-like parts in these images, which may be linked to the decision variables. It is theoretically possible to analyze this part of the system in greater detail, to detect which objects in the scene, if any, are linked to particular trait predictions. However, the number of training samples is small compared to the number of object classes the network is originally trained for, and consequently, such an analysis may be misleading.

It would be really interesting to conduct a more extensive study to see which objects are associated with which personality traits strongly. Obviously, cultural factors should also be considered for this purpose. In our previous work, we have illustrated the effectiveness of scene features for predicting Big Five traits to some extent (Gürpınar et al, 2016a,b). For the Job Candidate Screening task, these features contribute to the final decision both directly (i.e. in a classifier that predicts the interview variable) and indirectly (i.e. over the personality trait predictions that are used in the final classifier for the interview variable).

4.2 Acoustic Features

There are excellent signal processing approaches for using the acoustic features. The open-source openSMILE tool (Eyben et al, 2010) is popularly used to extract acoustic features in a number of international paralinguistic and multi-modal chal-

lenges. The idea is to obtain a large pool of potentially relevant features by passing an extensive set of summarizing functionals on the low level descriptor (LLD) contours (e. g. Mel Frequency Cepstral Coefficients - MFCC, pitch, energy and their first/second order temporal derivatives).

We use the toolbox with a standard feature configuration that served as the challenge baseline sets in INTERSPEECH 2013 Computational Paralinguistics Challenge (Schuller et al, 2013). This set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs). Additionally, there are LLDs that complement these features, such as logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. In our former work, we compared INTERSPEECH 2013 configuration with the other baseline feature sets used in the computational paralinguistics challenges, and found it to be the most effective for the personality trait recognition task (Gürpınar et al, 2016b). Thus, based on the former analyses, here we use the configuration from (Schuller et al, 2013).

4.3 Classification

We use several levels of classifiers to obtain the model predictions. In all levels, we use simple classifiers with few meta-parameters to prevent overfitting. Overfitting typically happens if the number of free parameters in the classifier and the dimensionality of the samples are large with respect to the number of training samples. Since our models base their decisions on many features obtained from different channels, overfitting is a very important issue.

We use kernel extreme learning machines (ELM) in our first tier classification. The ELM classifier is basically a single-layer neural network, but the first layer weights are determined from the training data (in the kernel version), and the second layer weights are analytically computed. Subsequently, it is very fast to train. We have observed in our simulations that its accuracy is good, and the system is robust. We do not detail the classifier here, and refer the reader to (Huang et al, 2004) for technical details. We use a linear kernel, which has only a single parameter (the regularization coefficient), which we optimize with a 6-fold subject independent cross-validation on the training set.

Once the model has generated a number of predictions from multiple modalities via ELM classifiers, these are stacked to a Random Forest (RF) classifier in the second stage of classification. This is the fusion stage, where the classifier learns to give appropriate weights to different modalities, or features. The RF classifier is an ensemble of decision tree (DT) classifiers. Tree based classifiers base their decisions on multiple tests, where each internal node of the tree tests one attribute, or a feature of the input sample, deciding which branch will be taken next. The root note contains the first test, and the leaf nodes of the tree will contain the decision, i.e. the assigned class of the sample. It is possible to trace the decisions from root to branch, and see which attributes have led to the particular decision. Consequently, decision trees are easy to interpret. The random forest introduces robustness to de-

cision trees by randomly sampling subsets of instances with replacement, and by training multiple trees based on these samples (Breiman, 2001).

To increase the interpretability of the final decision on the interview variable, we use the training set mean value of each attribute to binarize each score prediction of the RF as HIGH or LOW. Thus, if the model predicts the agreeableness of a person as higher than the average agreeableness of the training samples, it is labeled as HIGH AGREEABLENESS. The final classifier that decides on the interview variable is a decision tree, which takes the binarized apparent personality scores, predicted by the RF, and outputs the binary interview class (i.e. invited, or not invited).

Once the decision is given, the system converts it into an explicit description using "if-then" rules and a template, by tracing the decision from the root of the tree to the leaf. The template is formed as follows (Kaya et al, 2017a):

- If the invite decision is 'YES' $\rightarrow$ 'This [gentleman/lady] is invited due to [his/her] high apparent {list of high scores on the trace}' [optional depending on path:', although low {list of low scores on the trace} is observed.']
- If the invite decision is 'NO' $\rightarrow$ 'This [gentleman/lady] is not invited due to [his/her] low apparent {list of low scores on the trace}' [optional depending on path: ' , although high {list of high scores on the trace} is observed.']

In the preliminary weighted fusion experiments we have conducted, we have observed that the video modality typically has higher weight in the final prediction. Similarly, in the audio-scene model, the audio features are more dominant. We reflect this prior knowledge in the automatically generated explanations by checking whether the high/low scores of each dimension have the same sign with that of the model trained on facial features. After this check, the system includes some extra information for the leading apparent personality dimension that helped admittance (or caused rejection). The template for this information is:

'The impressions of {list of traits where visual modality has the same sign with the final decision} are primarily gained from facial features.' [optional, depending on existence: 'Furthermore, the impression of {the list of audio-dominant traits} is predominantly modulated by voice.']

Finally, each record is accompanied with the aligned face from the first face-detected frame of the video and with a bar graph of the mean-normalized predicted scores. This helps the decision maker visualize more precisely what the system computed to base its decision. We give several output examples in the next section.

5 Experimental Results

The "ChaLearn LAP Apparent Personality Analysis: First Impressions" challenge consists of 10 000 clips collected from 5 563 YouTube videos, where the poses are more or less frontal, but the resolution, lighting and background conditions are not controlled, hence providing a dataset with in-the-wild conditions. Each clip in the

training set is labeled for the Big Five personality traits and an "interview invitation" annotation using Amazon Mechanical Turk. The former is an apparent personality trait, and does not necessarily reflect the actual personality of the person. Similarly, the latter is a decision on whether the person in the video is invited to the interview or not, and signifies a positive or negative general impression.

For brevity, we skip corpus related information here, and refer the reader to (Lopez et al, 2016) for details on the challenge. The performance score in this challenge is the Mean Absolute Error subtracted from 1, which is formulated as follows:

$$1 - \sum_{i}^{N} \frac{|\hat{y}_i - y_i|}{N}, \tag{1}$$

where N is the number of samples, $\hat{y}$ is the predicted label and y is the true label ($0 \leq y \leq 1$). This means the final score varies between 0 (worst case) and 1 (best case).

The competition has a clear experimental protocol, which is followed in this work. The test set labels are sequestered, and limited number of test score submissions were allowed to prevent overfitting. We describe two sets of experiments, by taking a regression and a classification approach, respectively.

5.1 Experimental Results using Regression

The natural way to predict continuous apparent personality traits is via regression. We train our regressors with 6 000 training set instances, using a 6-fold cross-validation (CV) to optimize model hyper-parameters for each feature type and their combinations. Training and validation sets were combined for training the final system for test set predictions.

In Table 2, we report the validation set performances of individual features, as well as their feature-, score- and multi-level fusion alternatives. Here, System 0 corresponds to the top entry in the ICPR 2016 Challenge (Gürpınar et al, 2016b), which uses the same set of features and fuses scores with linear weights. For the weighted score fusion, the weights are searched in the [0,1] range with steps of 0.05. Face, scene, and audio features are used individually, and reported in lines 1-4. These indicate the accuracy of single-modality subsystems. Lines 5-8 are the multimodal fusion approaches.

In general, fusion scores are observed to benefit from complementary information of individual sub-systems. Moreover, we see that fusion of two different types of face features improves over their individual performance. Similarly, the feature level fusion of audio and scene sub-systems is observed to benefit from complementarity. The final score fusion with RF outperforms weighted fusion in all but one dimension (agreeableness), where the performances are equal.

Based on the validation set results, the best fusion system (System 8 in Table 2) is obtained by stacking the predictions from Face feature-fusion (FF) model (Sys-

Table 2: Validation set performance of the proposed framework (System 8) and its sub-systems. FF: Feature-level fusion, WF: Weighted score-level fusion, RF: Random Forest based score-level fusion. INTER: Interview invite variable. AGRE: Agreeableness. CONS: Conscientiousness. EXTR: Extraversion. NEUR: Neuroticism. OPEN: Openness to experience.

SysID	System	INTER	AGRE	CONS	EXTR	NEUR	OPEN	TRAIT AVG
0	ICPR 2016 Winner	N/A	0.9143	0.9141	0.9186	0.9123	0.9141	0.9147
1	Face: VGGFER33	0.9095	0.9119	0.9046	0.9135	0.9056	0.9090	0.9089
2	Face: LGBPTOP	0.9112	0.9119	0.9085	0.9130	0.9085	0.9103	0.9104
3	Scene: VD_19	0.8895	0.8954	0.8924	0.8863	0.8843	0.8942	0.8905
4	Audio: OS_IS13	0.8999	0.9065	0.8919	0.8980	0.8991	0.9022	0.8995
5	FF(Sys1, Sys2)	0.9156	0.9144	0.9125	0.9185	0.9124	0.9134	0.9143
6	FF(Sys3, Sys4)	0.9061	0.9091	0.9027	0.9013	0.9033	0.9068	0.9047
7	WF(Sys5, Sys6)	0.9172	**0.9161**	0.9138	0.9192	0.9141	0.9155	0.9157
8	RF(Sys5, Sys6)	**0.9198**	**0.9161**	**0.9166**	**0.9206**	**0.9149**	**0.9169**	**0.9170**

tem 5) with the Audio-Scene FF model (System 6). This fusion system renders a test set performance of 0.9209 for the interview variable, ranking the first and beating the challenge baseline score (see Table 3). Furthermore, the average of the apparent personality trait scores is 0.917, which advances the state-of-the art result (0.913) obtained by the winner of ICPR 2016 ChaLearn LAP First Impression contest (Gürpınar et al, 2016b).

Table 3: Test set performance of the top systems in the CVPR'17 Coopetition - Quantitative Stage

Participant	INTER	AGRE	CONS	EXTR	NEUR	OPEN	TRAIT AVG
Ours	**0.9209**	**0.9137**	**0.9198**	**0.9213**	**0.9146**	**0.9170**	**0.9173**
Baseline	0.9162	0.9112	0.9152	0.9112	0.9104	0.9111	0.9118
First Runner Up	0.9157	0.9103	0.9138	0.9155	0.9083	0.9101	0.9116
Second Runner Up	0.9019	0.9032	0.8949	0.9027	0.9011	0.9047	0.9013

The test set results of the top ranking teams are both high and competitive. When individual personality dimensions are analyzed, we see that our system ranks the first in all dimensions, exhibiting the highest improvement over the baseline in prediction of Extraversion and the Interview variable. We also observe that the proposed system's validation and test accuracies are very similar: the mean absolute difference of the six dimensions is 0.13%. Therefore, we can conclude that the generalization ability of the proposed system is high.

After the official challenge ended, we have obtained the test set labels from the organizers and analyzed the distribution of the absolute error in our system with respect to the ground truth. Figure 2 shows the scatter plots of the six target variables. The x-axis denotes the ground truth scores, and the V shape we observe in these plots, with a mass centered around the point (0.5, 0.05), means that our least squares based regressor is conservative, trying to avoid extreme decisions. The largest errors are made for high and low value assignments, particularly for Agreeableness. A cumulative distribution analysis shows that over all dimensions, 37.5% of the test set

predictions have MAEs less than 0.05, and 67.3% of predictions have MAEs less than 0.1, with only 5.2% of the predictions having a MAE higher than 0.2.

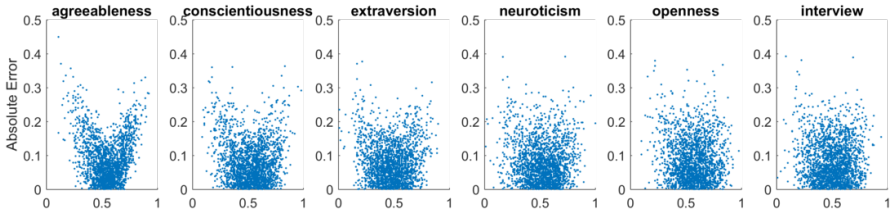


Fig. 2: Absolute error of the test set predictions (y-axis) as a function of ground truth (x-axis).

5.2 Experimental Results using Classification

For improved interpretability, the prediction problem can be handled as a binary classification into LOW and HIGH values, which we investigate in this section. Additionally, we analyze how well a parsimonious system can do by looking at a single frame of the video, instead of face analysis in all frames.

To adapt the problem for classification, the continuous target variables in the [0,1] range are binarized using the training set mean statistic for each target dimension, separately. For the single-image tests, we extracted deep facial features from our fine-tuned VGG-FER DCNN, and accompanied them with easy-to-extract image descriptors, such as Local Binary Patterns (LBP) (Ojala et al, 2002), Histogram of Oriented Gradients (HOG) (Dalal and Triggs, 2005) and Scale Invariant Feature Transform (Lowe, 2004).

Hyper-parameter optimization and testing follow similar schemes as the previous section. The test set classification performances of the top systems for single- and multi-modal approaches are shown in Table 4. As expected, we see that the audio-visual approach also performs best in the classification task (77.10% accuracy on the interview variable). This is followed by the video-only approach using facial features (76.35%), and the fusion of audio with face and scene features from the first image (74%). Although this is relatively 4.6% lower compared to the best audio-visual approach, it is higly motivating, as it uses only a single image frame to predict the personality impressions and interview invitation decision, which the annotators gave by watching the whole video. It shows that without resorting to costly image processing and DCNN feature extraction for all images in a video, it is possible to achieve high accuracy, comparable to the state-of-the-art.

The dimension that is the hardest to classify is agreeableness, whereas accuracy for conscientiousness was consistently the highest (see Figure 3). Among the conventional image descriptors, HOG was the most succesful, with an average validation set recognition accuracy (over traits) of 70%, using only a single facial image.

On the other hand, the fusion of scene and face features from the first video frame outperform acoustic features on both the development and test sets by 3%.

Table 4: Test set classification accuracies for the top single and multimodal systems. The scene feature is extracted from the first video frame. FF: Feature Fusion, EF: Equal weighted score fusion.

Sys.	Modality	Features/Fusion	Interview	Trait Avg.
1	Audio + Video	EF(FaceSys,AudioSceneSys)	**77.10**	**75.63**
2	Video (Face Seq.)	FF(VGGFER33,LGBPTOP)	76.35	74.45
3	Audio + Scene + First Face	FF(IS13,VGGFER33,VGGVD19,LBP)	74.00	72.31
4	Audio + Scene	FF(IS13,VGGVD19)	71.95	70.47
5	First Face + Scene	FF(VGGFER33,VGGVD19)	71.15	69.97
6	Audio	IS13 Functionals	69.25	67.93

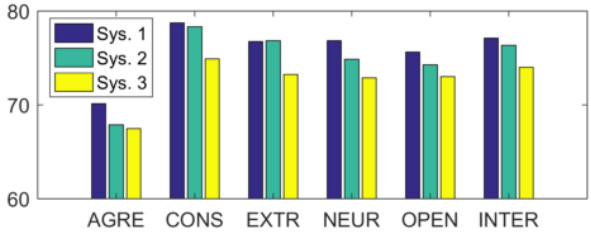


Fig. 3: Test set classification performance of top three fusion systems over personality traits and the interview variable. Sys. 1: Audio-Video system, Sys. 2: Video only system, Sys. 3: Audio plus a single image based system. NEUR refers to non-Neuroticism as it is used throughout the paper.

6 Explainability Analysis

We now turn to the explainability analysis, which was tackled in the qualitative part of the ChaLearn competition. We use the final score outputs of the quantitative stage, as well as the classifiers themselves to produce readable explanations of the decisions.

To make the scores more accessible, we binarize them (as LOW-HIGH) by thresholding each dimension at corresponding training set mean, and feed them to a decision tree classifier, as explained in Section 4. In the preliminary experiments, we tried grouping the scores into more than two levels, using the mean and variance statistics. However, the final classification accuracy suffered, and this was abandoned.

The decision tree trained on the predicted Big Five personality dimensions gives a classification accuracy of 94.2% for the binarized interview variable. A visual illustration of the decision tree (DT) is given in Figure 4.

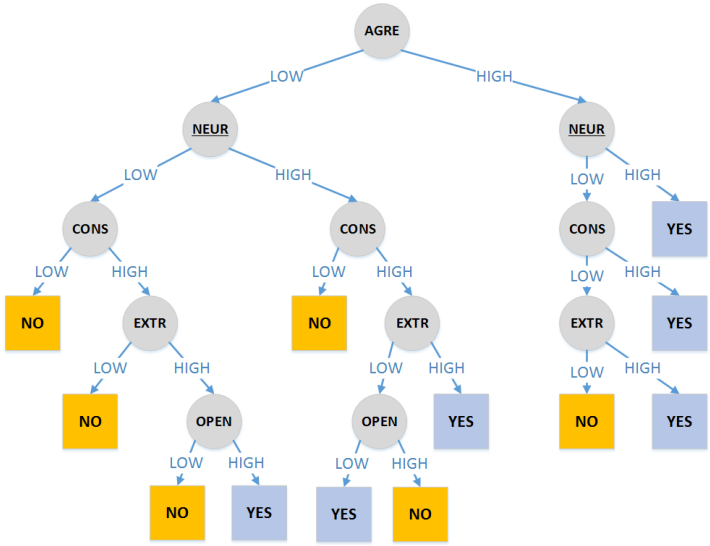


Fig. 4: Illustration of the trained decision tree for job interview invitation. NEUR represents non-Neuroticism, as explained in the text.

The learned model is intuitive in that the higher scores of traits generally increase the chance of interview invitation. As can be seen from the figure, the DT ranks the relevance of the predicted Big Five traits from the highest (Agreeableness) to the lowest (Openness to Experience) with respect to information gain between corresponding trait and the interview variable. The second most important trait for job interview invitation is Neuroticism, which is followed by Conscientiousness and Extraversion. Neuroticism is the only trait which correlates negatively with the Interview variable, so it was represented with its opposite (i.e. non-Neuroticism) during annotations, to ensure sign consistency. Throughout this paper, we use non-Neuroticism as a feature. If the Openness score is high, then having a high score in any of the non-Neuroticism, Conscientiousness or Extraversion variables suffices for invitation. Chances of invitation decrease if Agreeableness is low: only three out of eight leaf nodes are "YES" in this branch. In two of these cases, one has to have high scores in three out of four remaining traits.

There is an interesting rule related to Openness. In some cases high Openness leads to "invite", whereas in others it leads to "do not invite". If Agreeableness is low, but non-Neuroticism and Extraversion are high, then the Openness should be low for interview invitation (a high Openness score results in rejection). This may be due to an unwanted trait combination: someone with a low Agreeableness, Extraversion, and Neuroticism, but high Openness may be perceived as insincere and arrogant.

For verbal explanations, we converted the DT structure into a compact set of "if-then" rules in the form mentioned earlier. The metadata provided by the organizers

do not contain sex annotations, which could have been useful in explanatory sentences. For this purpose, we have initially annotated 4 000 development set (training + validation) videos using the first face-detected frames, then trained a sex prediction model based on the audio and video features used in the apparent personality trait recognition. The ELM based sex predictors gave 97.6% and 98.9% validation set accuracies using audio (openSMILE) and video (CNN-FUN) features, respectively. We fused the scores of audio and video models with equal weight and obtained a validation set accuracy of 99.3%, which is close to perfect. We then used all annotated data for training with the optimized hyper-parameters and cast predictions on the remaining 6 000 (validation + test set) instances. After the challenge, we annotated the whole set of 10.000 videos for apparent age, sex, and ethnicity.

The verbal explanations are finally accompanied with the aligned image from the first face-detected frame and the bar graphs of corresponding mean normalized scores. When we analyze the results, we observe that individually processed clips cut from different places of a single input video have very similar scores, and the same reasons for the invitation decision, showing the consistency of the proposed approach. Figure 5 illustrates some automatically generated verbal and visual explanations for this stage.

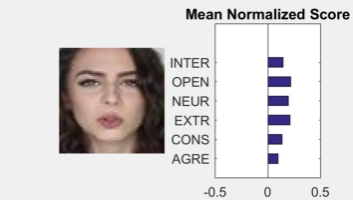
The test set of the quantitative challenge was based on the accuracy (1-MAE) of the interview variable. In the qualitative stage, the submissions (one for each team) were evaluated by a committee based on the following criteria:

- **Clarity**: Is the text understandable / written in proper English?
- **Explainability**: Does the text provide relevant explanations to the hiring decision made?
- **Soundness**: Are the explanations rational and, in particular, do they seem scientific and/or related to behavioral cues commonly used in psychology.
- **Model interpretability**: Are the explanation useful to understand the functioning of the predictive model?
- **Creativity:** How original / creative are the explanations?

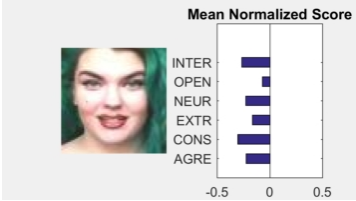
The test set scores of the official competition for this stage are shown in Table 5. Our team ranked the first in terms of the overall mean score. However, since the first runner up has better Creativity scores and the mean scores are not significantly different, both teams are designated as winners.

Table 5: Qualitative stage test stage winner teams' scores

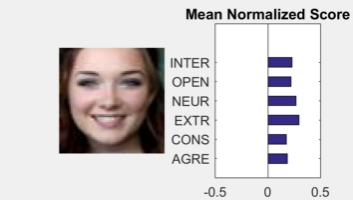
Participant	**Our Team**	First Runner Up
Clarity	4.31±0.54	3.33±1.43
Explainability	3.58±0.64	3.23±0.87
Soundness	3.40±0.66	3.43±0.92
Interpretability	3.83±0.69	2.40±1.02
Creativity	2.67±0.75	3.40±0.8
Mean Score	**3.56**	3.16



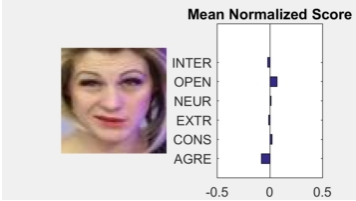
This lady is invited for an interview due to her high apparent agreeableness and neuroticism impression. The impressions of agreeableness, conscientiousness, extraversion, neuroticism and openness are primarily gained from facial features.



This lady is not invited due to her low apparent agreeableness, neuroticism, conscientiousness, extraversion and openness scores. The impressions of agreeableness, conscientiousness, extraversion, neuroticism and openness are primarily gained from facial features.



This lady is invited for an interview due to her high apparent agreeableness and neuroticism impression. The impressions of agreeableness, conscientiousness, extraversion, neuroticism and openness are primarily gained from facial features.



This lady is not invited for an interview due to her low apparent agreeableness and extraversion impressions, although predicted scores for neuroticism, conscientiousness and openness were high. It is likely that this trait combination (with low agreeableness, low extraversion, and high openness scores) does not leave a genuine impression for job candidacy. The impressions of agreeableness, extraversion, neuroticism and openness are primarily gained from facial features. Furthermore, the impression of conscientiousness is predominantly modulated by voice.

Fig. 5: Sample verbal and visual explanations automatically generated by the system.

6.1 The Effect of Ethnicity, Age, and Sex

Automatic machine learning approaches that rely on human-supplied labels for supervised learning are prone to learn the biases inherent in these labels. To investigate potential biases in job interview screening, 10 000 videos of the ChaLearn corpus are annotated for apparent ethnicity, age, and sex in (Escalante et al, 2018). It is shown that people who originally annotated the corpus for the interview variable

are negatively biased toward African-Americans, while being positively biased towards Caucasians, both in terms of personality traits and the interview variable.

When biases for age and sex are investigated, they are found to be strongly correlated. As can be expected, the prior probability of job interview invitation is lower than 0.5 for people who are outside the working-age group, i.e. not in the age range of [18, 60]. Within the working-age group, the prior probability of job invitation is positively (and strongly) correlated with age of male candidates, while it is negatively correlated with the age of female candidates. In other words, the annotators prefer younger female candidates and older male candidates for invitation to a job interview.

We have analysed how the proposed explanation system varies with respect to apparent age group and sex combinations. To preserve simplicity, we thresholded the working-age group at the age of 33, thus having a younger working age group with range [18, 32] and an older age group with range [33, 60]. With two age groups and two different sexes, we trained four decision trees. The results are shown in Figure 6. We observe that while all trees are different in structure, they all have Agreeableness at their root node, which indicates the importance of this cue for invitation to interview. Moreover, the importance ordering of the variables (i.e. apparent personality traits) imposed by the DTs for females are the same as those obtained from the whole dataset (given in Figure 4).

7 Discussion and Conclusions

In this chapter, we have discussed an automatic system for the multimedia job candidate screening task. The proposed multi-level fusion framework uses multimodal fusion followed by a decision tree (DT), in order to produce text-based explanations of its decisions. These decisions are largely based on apparent personality predictions, which the system reports as intermediate results, but beyond that, the internal dynamics are not investigated for explainability. The proposed system ranked the first in both quantitative and qualitative stages of the official Challenge.

The scenario tackled in this chapter and in the related ChaLearn challenge is a limited case, where only passively recorded videos are available, as opposed to dyadic interactions. Subsequently, this scenario is more adequate to investigate first impression judgments, which are known to be very fast in their production, and very influential in behavior (Willis and Todorov, 2006). There is a recent trend to ask job candidates to submit video resumes for job applications, and a widely held belief that such a format, being richer than a paper resume, will give a better leverage for the assessor to judge the personality of the candidate. Apers and Derous (2017) recently reported some results that illustrate that both paper resumes and video resumes are inadequate for judging the real personality of a candidate. But there is no doubt that they influence the recruiter's decisions, so the impact on the first impressions needs to be taken into account.

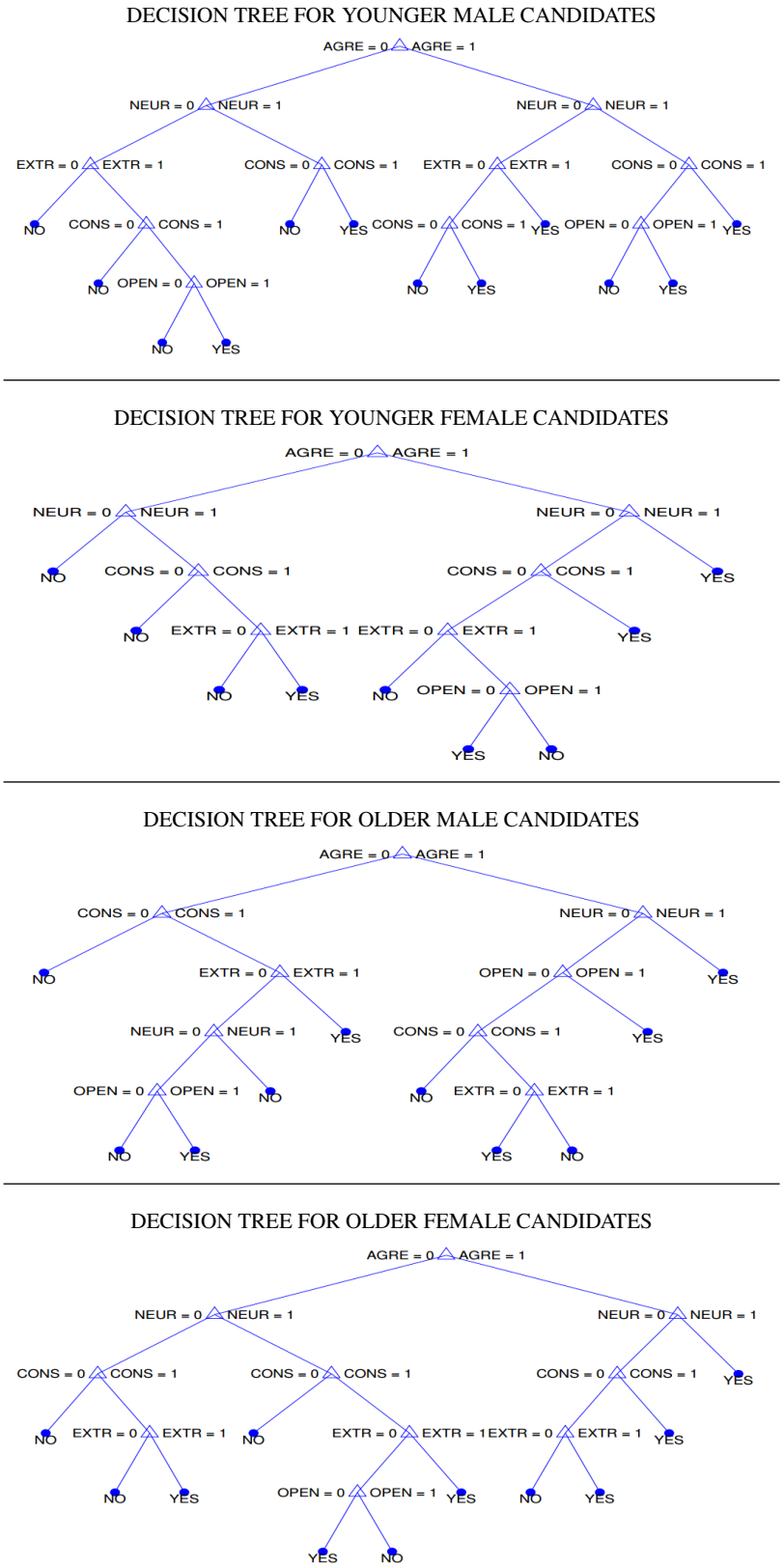


Fig. 6: Visualization of age group and sex dependent decision trees to be used as explanation models. Here, NEUR refers to non-Neuroticism.

There is substantial research on first impressions, linking these to judgments of competence. Research on stereotype judgments put forward that two dimensions, posited as universal dimensions of human social cognition, particularly capture stereotype judgments, namely, warmth and competence (Fiske et al, 2002). These dimensions are, for instance, helpful to describe Western stereotypes against elderly (i.e. high warmth and low competence), or against Asians (i.e. high competence and low warmth). The warmth dimension predicts whether the interpersonal judgment is positive or negative (i.e. the valence of the impression), whereas the competence dimension quantifies the strength of this impression (Fiske et al, 2007). In an interesting study, Agerström et al (2012) investigated 5 636 job applications by Swedish and Arab applicants, and found substantial discrimination where Arab applicants receive fewer invitations to job interviews. The authors used the warmth-competence model to suggest that the Arab applicants need to "appear warmer and more competent than Swedish applicants to be invited equally often," but how exactly this can be achieved is an open question. Automatic analysis tools, if they can properly quantify such perceived qualities, can act as useful training tools.

There is further research on stigmatizing features that give the applicant a distinct disadvantage during a job interview. Examples of such features include obesity (Agerström and Rooth, 2011), physical unattractiveness (Dipboye, 2005), and visible disabilities (Hayes and Macan, 1997). An automatic system that can accurately predict how such biases will effect decisions can be a useful tool in combatting these biases.

One of the limitations of the automatic job assessment task is that it considers only the applicant. However, any biases that exist on the interviewer's side are also essential in assessing the quality of this process (Dipboye et al, 2012). Future work should therefore ideally capture both the interviewer and the applicant during interactions. In particular, both the expertise and the confidence of the interviewer in their hiring decision need to be recorded to properly analyze the strength of the biases in the assessment.

Acknowledgment

This work is supported by Boğaziçi University Project BAP 16A01P4 and by the BAGEP Award of the Science Academy.

References

Agerström J, Rooth DO (2011) The role of automatic obesity stereotypes in real hiring discrimination. Journal of Applied Psychology 96(4):790–805

Agerström J, Björklund F, Carlsson R, Rooth DO (2012) Warm and competent Hassan= cold and incompetent Eric: A harsh equation of real-life hiring discrimination. Basic and Applied Social Psychology 34(4):359–366

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, ACM, pp 15–18

Alam F, Stepanov EA, Riccardi G (2013) Personality traits recognition on social network-Facebook. WCPR (ICWSM-13), Cambridge, MA, USA

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, IEEE, pp 356–361

Apers C, Derous E (2017) Are they accurate? recruiters' personality judgments in paper versus video resumes. Computers in Human Behavior 73:9–19

Barezi EJ, Kampman O, Bertero D, Fung P (2018) Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction. arXiv preprint arXiv:180500705

Breiman L (2001) Random forests. Machine learning 45(1):5–32

Celiktutan O, Gunes H (2016) Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. IEEE Transactions on Affective Computing 8(1):29–42, DOI 10.1109/TAFFC.2015.2513401

Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, vol 1, pp 886–893

Dipboye RL (2005) Looking the part: Bias against the physically unattractive as discrimination issue. Discrimination at work: The psychological and organizational bases pp 281–301

Dipboye RL, Macan T, Shahani-Denning C (2012) The selection interview from the interviewer and applicant perspectives: Cant have one without the other. The Oxford handbook of personnel assessment and selection pp 323–352

Escalante HJ, Ponce-López V, Wan J, Riegler MA, Chen B, Clapés A, Escalera S, Guyon I, Baró X, Halvorsen P, et al (2016) Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In: Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, pp 67–73

Escalante HJ, Guyon I, Escalera S, Jacques J, Madadi M, Baró X, Ayache S, Viegas E, Güçlütürk Y, Güçlü U, et al (2017) Design of an explainable machine learning challenge for video interviews. In: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, pp 3688–3695

Escalante HJ, Kaya H, Salah AA, Escalera S, Gucluturk Y, Guclu U, Baro X, Guyon I, Junior JJ, Madadi M, et al (2018) Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos. arXiv preprint arXiv:180200745

Eyben F, Wöllmer M, Schuller B (2010) OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: ACM International Conference on Multimedia, pp 1459–1462

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 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, pp 1–6

Fernando T, et al (2016) Persons personality traits recognition using machine learning algorithms and image processing techniques. Advances in Computer Science: an International Journal 5(1):40–44

Fiske ST, Cuddy AJ, Glick P, Xu J (2002) A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of personality and social psychology 82(6):878–902

Fiske ST, Cuddy AJ, Glick P (2007) Universal dimensions of social cognition: Warmth and competence. Trends in cognitive sciences 11(2):77–83

Gievska S, Koroveshovski K (2014) The impact of affective verbal content on predicting personality impressions in youtube videos. In: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, pp 19–22

Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH, et al (2013) Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing, Springer, pp 117–124

Güçlütürk Y, Güçlü U, van Gerven M, van Lier R (2016) Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop Proceedings, pp 349–358

Güçlütürk Y, Güçlü U, Baro X, Escalante HJ, Guyon I, Escalera S, van Gerven MAJ, van Lier R (2017) Multimodal first impression analysis with deep residual networks. IEEE Transactions on Affective Computing, online DOI 10.1109/TAFFC.2017.2751469

Gürpınar F, Kaya H, Salah AA (2016a) Combining deep facial and ambient features for first impression estimation. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop Proceedings, pp 372–385

Gürpınar F, Kaya H, Salah AA (2016b) Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation. In: ChaLearn Joint Contest and Workshop on Multimedia Challenges Beyond Visual Analysis, Collocated with ICPR 2016, Cancun, Mexico

Hayes TL, Macan TH (1997) Comparison of the factors influencing interviewer hiring decisions for applicants with and those without disabilities. Journal of Business and Psychology 11(3):357–371

Huang GB, Zhu QY, Siew CK (2004) Extreme Learning Machine: a new learning scheme of feedforward neural networks. In: IEEE International Joint Conference on Neural Networks, vol 2, pp 985–990

Junior JCSJ, Güçlütürk Y, Perez M, Güçlü U, Andujar C, Baro X, Escalante HJ, Guyon I, van Gerven MAJ, van Lier R, Escalera S (2018) First impressions: A survey on computer vision-based apparent personality trait analysis. arXiv preprint arXiv:180408046 URL `https://arxiv.org/abs/1804.08046`

Kaya H, Salah AA (2014) Continuous mapping of personality traits: A novel challenge and failure conditions. In: Proceedings of the 2014 ICMI Workshop on Mapping Personality Traits Challenge, ACM, pp 17–24

Kaya H, Gürpınar F, Salah AA (2017a) Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. In: CVPR Workshops, Honolulu, Hawaii, USA, pp 1651–1659

Kaya H, Gürpınar F, Salah AA (2017b) Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing 65:66–75, DOI http://dx.doi.org/10.1016/j.imavis.2017.01.012

Larkin JC, Pines HA (1979) No fat persons need apply: experimental studies of the overweight stereotype and hiring preference. Sociology of Work and Occupations 6(3):312–327

Lopez VP, Chen B, Places A, Oliu M, Corneanu C, Baro X, Escalante HJ, Guyon I, Escalera S (2016) Chalearn lap 2016: First round challenge on first impressions - dataset and results. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop Proceedings, Springer, pp 400–418

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2):91–110

Madzlan N, Han J, Bonin F, Campbell N (2014) Towards automatic recognition of attitudes: Prosodic analysis of video blogs. Speech Prosody, Dublin, Ireland pp 91–94

Nowson S, Gill AJ (2014) Look! who's talking?: Projection of extraversion across different social contexts. In: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, pp 23–26

Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7):971–987

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Transactions on knowledge and data engineering 22(10):1345–1359

Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: British Machine Vision Conference

Qin R, Gao W, Xu H, Hu Z (2016) Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. arXiv preprint arXiv:160407499

Sarkar C, Bhatia S, Agarwal A, Li J (2014) Feature analysis for computational personality recognition using youtube personality data set. In: Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, ACM, pp 11–14

Schuller B, Steidl S, Batliner A, Nöth E, Vinciarelli A, Burkhardt F, Van Son R, Weninger F, Eyben F, Bocklet T, et al (2012) The INTERSPEECH 2012 speaker trait challenge. In: INTERSPEECH, pp 254–257

Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S (2013) The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. In: INTERSPEECH, Lyon, France, pp 148–152

Sidorov M, Ultes S, Schmitt A (2014) Automatic recognition of personality traits: A multimodal approach. In: Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop, ACM, pp 11–15

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Valente F, Kim S, Motlicek P (2012) Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In: INTERSPEECH, pp 1183–1186

Willis J, Todorov A (2006) First impressions: Making up your mind after a 100-ms exposure to a face. Psychological science 17(7):592–598

Xiong X, De la Torre F (2013) Supervised Descent Method and Its Application to Face Alignment. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 532–539

Zhang CL, Zhang H, Wei XS, Wu J (2016) Deep bimodal regression for apparent personality analysis. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop Proceedings, pp 311–324