Multi-modal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs

Heysem Kaya[1], Furkan Gürpınar[2], and Albert Ali Salah[2]
[1]Department of Computer Engineering, Namık Kemal University, Tekirdağ, Turkey
[2]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
hkaya@nku.edu.tr, furkan.gurpinar@boun.edu.tr,salah@boun.edu.tr

Abstract

We describe an end-to-end system for explainable automatic job candidate screening from video CVs. In this application, audio, face and scene features are first computed from an input video CV, using rich feature sets. These multiple modalities are fed into modality-specific regressors to predict apparent personality traits and a variable that predicts whether the subject will be invited to the interview. The base learners are stacked to an ensemble of decision trees to produce the outputs of the quantitative stage, and a single decision tree, combined with a rule-based algorithm produces interview decision explanations based on the quantitative results. The proposed system in this work ranks first in both quantitative and qualitative stages of the CVPR 2017 ChaLearn Job Candidate Screening Coopetition.

1. Introduction and Related Work

The applications of affective computing are rapidly growing, thanks to the developments in signal processing and machine learning, as well as through inter-disciplinary projects involving research in psychology. Job interviews rely on short interactions with individuals, but have potentially life-changing impact for the job seekers. CVPR 2017 ChaLearn Job Candidate Screening (JCS) Coopetition seeks to help both recruiters and job candidates by promoting the development of automatic recommendation systems based on multi-media CVs. This paper proposes an end-to-end system for this purpose.

The challenge is 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 are encouraged to use the personality trait dimensions in prediction (quantitative) and explanation (qualitative) stages.

Automatic prediction of apparent personality is of interest in many applications ranging from computer assisted tutoring systems, forensics and job recommendation systems. Since the complexity of the personality formation makes it very hard to assess it automatically [18, 6], researchers generally work on the *impressions* (apparent personality), instead of the personality itself [21, 15]. The goal of this work is to predict the "invite for interview" variable together with (and in relation to) the apparent personality, using the data and protocol from the ChaLearn Looking at People 2016 First Impression Challenge [21]. The apparent personality is assessed along the "Big Five" personality traits, namely, Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN).

There are several recent approaches for recognizing apparent personality traits from different modalities such as audio [31, 22], text [2, 23, 11] and visual information [10, 25]. To increase the robustness of predictions, multimodal systems are also investigated [27, 1, 9, 29, 14].

In this paper, we use deep learning based classifiers to predict apparent personality ratings. Deep learning is becoming ever more significant in human behavior analysis, in the sense that it provides representations that are robust for many tasks, compared to traditional feature extraction methods. Deep learning has been successfully applied to related vision-based tasks such as face recognition [24], emotion recognition [19, 17] and age estimation [26, 20, 7, 13]. Moreover, deep representations of images are often interchangeable among tasks, enabling transfer learning from pre-trained models. The disadvantages are the relatively high computational requirements for training such systems, the large amount of training data required, and (relatively) poor temporal extension.

The classification problem we tackle in this paper is based on assessing a short input video. The available modalities for analysis include the facial image of the candidate, the sound 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/protocol [15], we use audio, scene, and facial

features as separate channels, and use Extreme Learning Machine classifiers to evaluate each channel. Convolutional neural networks are used for feature extraction.

In this work, we propose a novel stacking framework that combines both feature and score level fusion. The proposed framework enables the recognition of apparent personality traits, as well as of the interview variable by modeling deep features that encode affective cues. The system leverages multi-modality through a multi-level fusion by an ensemble of Decision Trees (Random Forests). The facial features are extracted after fine tuning of a pre-trained Deep Convolutional Neural Network (DCNN) on a corpus with affect variations. The pipeline of the proposed system for the quantitative challenge is illustrated in Figure 1.

The second stage of the coopetition aims to produce explanations for the decisions of the system. For this stage, the final predictions are binarized and modeled using a single decision tree (DT). The reason of using DT is the fact that it is easy to interpret the model and trace the outcome. Moreover, the resulting tree can be converted into a compact set of "if-then" rules for implementation of the decision algorithm.

The remainder of this paper is organized as follows. In the next section we provide background and details on the methodology. Then in Section 3, we present the experimental results. Finally, Section 4 concludes the paper with remarks on the proposed approach in context.

2. Methodology

Our proposed approach evaluates a short video clip that contains a single person, and outputs an estimate of the interview variable and the OCEAN dimensions as mentioned earlier. Both visual and audio features are used in the proposed approach. In this section, we describe the main steps of our pipeline, namely, face alignment, feature extraction, and modeling. We provide a comparison of the proposed approach with the state-of-the-art in Section 3.

2.1. Visual Feature Extraction

Facial features are extracted over an entire video segment and summarized by functionals. Scene features, however, are extracted from the first image of each video only. The assumption is that videos do not stretch over multiple shots.

2.1.1 Face Features

Faces are detected on all frames of the video input. For face alignment, we have used the popular Supervised Descent Method (SDM) [32]. 49 landmarks are located on each detected face. The roll angle is estimated from the eye corners to rotate the image accordingly. Then a margin of 20% of the interocular distance around the outer landmarks is added to crop the facial image. Each image is resized to 64×64 pixels.

After aligning the faces, image-level deep features are extracted from a convolutional neural network trained for facial emotion recognition. To prepare this feature extractor, we start with the pre-trained VGG-Face network [24], 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 [12]. 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 compare and combine deep facial features with a spatio-temporal descriptor called Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [3] that is shown to be effective in emotion recognition [17]. 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.

After extracting frame-level features from each aligned face, we summarize the videos by computing functional statistics of each dimension over time. 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.

2.1.2 Scene Features

In order to use ambient information in the images to our advantage, we extract a set of features using the VGG-VD-19 network [30], 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. The effectiveness of scene features for predicting Big Five traits is shown in [14, 15]. For Job Candidate Screening task, these features contribute to the final decision both directly and indirectly over the personality trait predictions.

2.2. Acoustic Features

The open-source openSMILE tool [8] is popularly used to extract acoustic features in a number of international paralinguistic and multi-modal challenges. The idea is to ob-

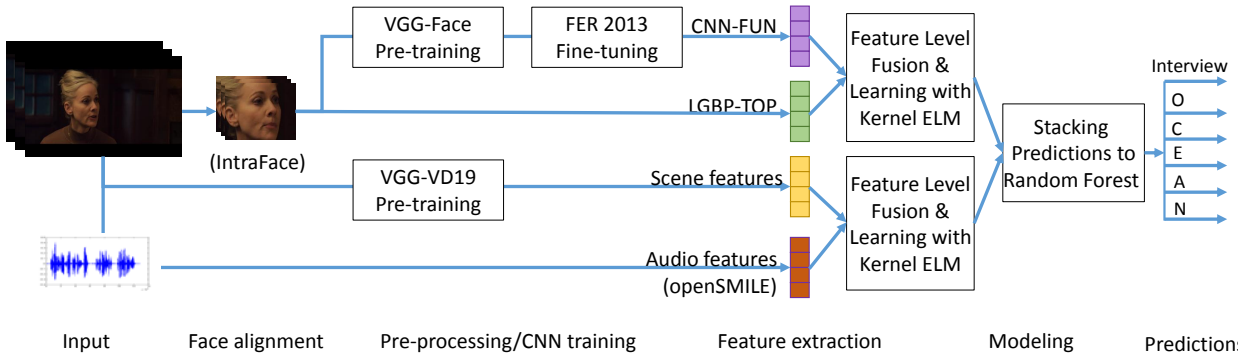


Figure 1: Flowchart of the proposed method.

tain a large pool of potentially relevant features by passing an extensive set of summarizing functionals on the low level descriptor contours (e. g. Mel Frequency Cepstral Coefficients, 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 [28]. This configuration was found to be the most effective acoustic feature set among others for personality trait recognition [15].

2.3. Regression with Kernel ELM

In order to model personality traits from visual features, we used kernel extreme learning machines (ELM), due to the learning speed and accuracy of the algorithm. In the following paragraphs, we briefly explain the learning strategy of ELM.

Initially, ELM is proposed as a fast learning method for Single Hidden Layer Feedforward Networks (SLFN): an alternative to back-propagation [16]. To increase the robustness and the generalization capability of ELM, a regularization coefficient $\mathbf{C}$ is included in the optimization procedure. Therefore, given a kernel $\mathbf{K}$ and the label vector $\mathbf{T} \in \mathbb{R}^{N\times 1}$ where N denotes the number of instances, the projection vector β is learned as follows:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}. \quad (1)$$

In order to prevent parameter over-fitting, we use the linear kernel $\mathbf{K}(x, y) = x^T y$, where x and y are the original feature vectors after min-max normalization of each dimension among the training samples. With this approach, the only parameter of our model is the regularization coefficient C, which we optimize with a 5-fold subject independent cross-validation on the training set.

2.4. Classifier Fusion with Random Forests

The predictions of the multi-modal ELM models are stacked to a Random Forest (RF), which is an ensemble of decision trees (DT) grown with a random subset of instances (sampled with replacement) and a random subset of features. The randomness in both features and samples allow diversity of the base learners and help avoid overfitting [5]. Furthermore, sampling with replacement leaves approximately one third of the training set instances *out-of-bag*, which are used to cross-validate the models and optimize the hyper-parameters at the training stage. This is a very important aspect of the method as far as the challenge conditions are concerned, as cross validation gives an unbiased estimate of the expected value of prediction error [4].

2.5. Qualitative Stage: Explaining the Model and Decisions

For the qualitative stage, the final predictions from the RF model are binarized by thresholding each score with its corresponding training set mean value. The binarized predicted OCEAN scores are mapped to the binarized ground truth interview variable using a DT classifier. The use of a DT is motivated by the fact that the resulting model is self-explanatory and can be converted into an explicit recommender algorithm using "if-then" rules. The proposed approach for decision explanation uses the trace of each decision from the root of the tree to the leaf. The template of the base description is formed as follows.

- 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.']

If the directly predicted interview outcome and the classification results of the decision tree do not match, we start the description with the following explanation:

"The directly predicted interview score and the classification based on traits are not consistent, the [gentleman/lady] may be re-evaluated. Following explanation is based on predicted traits."

In the preliminary weighted fusion experiments we have conducted, we 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, we include 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.']

We finally accompany each record with the aligned face from the first face-detected frame of the video and with a bar graph of the mean-normalized predicted scores.

3. Experiments

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 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 [21] for details on the challenge. The performance score in this challenge is the Mean Absolute Error subtracted from 1, which is formulated as follows:

$$1 - \sum_{i}^{N} \frac{|\hat{y}_i - y_i|}{N}, \tag{2}$$

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 score is then averaged over five tasks. This means the final score varies between 0 (worst case) and 1 (best case).

Both challenges are composed of learning and test phases. In the former, the coopetitors use the training set to learn and optimize a model and predict validation set instances, whose labels are unknown. The coopetitors had multiple submission options (up to 10 per day and up to 700 in total) for this phase. At the end of this phase, the coopetitors share their codes to reproduce the results along with a factsheet, and these are shared publicly by the organizers after some checks. The final evaluation on the test set is done as a single submission; if the coopetitors submit multiple prediction sets, only the last one is used for evaluation. Combined with the fact that the coopetitor does not learn the outcome of the test stage immediately after submission, this ensures that the submitted systems do not overfit the test set.

3.1. Experimental Results for the Quantitative Stage

For the learning stage, we used 6 000 training set instances, using a 6-fold cross-validation (CV) to optimize model hyper-parameters for each feature type and their combinations. Subsequently, the whole training set is used in model learning with corresponding optimal hyper-parameters to predict validation set instances. A similar procedure, but with 8-fold CV and on the combination of the training and validation sets, is carried out to predict the test instances.

In Table 1, 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 [15], 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. In general, fusion scores are observed to benefit from complementary information of individual sub-systems. Moreover, we see that fusion of face features improve 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 1) is obtained by stacking the predictions from Face feature-fusion (FF) model (System 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 2). 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

Table 1: 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	MEAN TRAITS
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**

ICPR 2016 ChaLearn LAP First Impression contest [15].

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.

3.2. Experimental Results for Qualitative Stage

As mentioned earlier, the final outputs of the quantitative stage serve as the inputs of the qualitative stage. These predictions are binarized (0/1 corresponding to low/high scores) by thresholding each dimension at corresponding training set mean. In the preliminary experiments, we tried grouping the scores into more than two levels, using the mean and variance statistics. However, the final classification accuracies of these alternatives were found to be lower compared to the mean value based simple discretization used in this work. The decision tree trained on the predicted Big Five personality dimensions gives a classification accuracy of 94.2% for binarized interview variable. The illustration of the decision tree (DT) is given in Figure 2.

On the overall, the 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 relevance of the predicted Big Five traits from highest (Agreeableness) to 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. The high/low scores of these top four traits are correlated with target variable and are observed to be consistent throughout the DT. If the Openness score is high, then having a high score in any of the Neuroticism, Conscientiousness or Extraversion variables suffices for invitation. Chances of invitation decrease if Agreeability 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 Agreeability is low, but Neuroticism and Extraversion are high, then the Openness should be low for interview invitation (a high Openness score results in rejection). While it is an interesting psychological case to study, this may be due to an unwanted trait combination: someone with low Agreeableness and Extraversion but high Neuroticism and 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 gender annotations, which could have been useful in explanatory sentences. For this purpose, we have manually annotated 4 000 development set (training + validation) videos using the first face-detected frames, then trained a gender prediction model based on the audio and video features used in the apparent personality trait recognition. The ELM based gender 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 casted predictions on the remaining 6 000 (validation + test set) instances.

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 analyzed the results, we observed that individually processed clips cut from different places of a single input video have very similar scores, and the exactly same reasons for

Table 2: Test set performance of the top systems in the CVPR 2017 Coopetition - Quantitative Stage

Participant	INTER	AGRE	CONS	EXTR	NEUR	OPEN	MEAN TRAITS
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

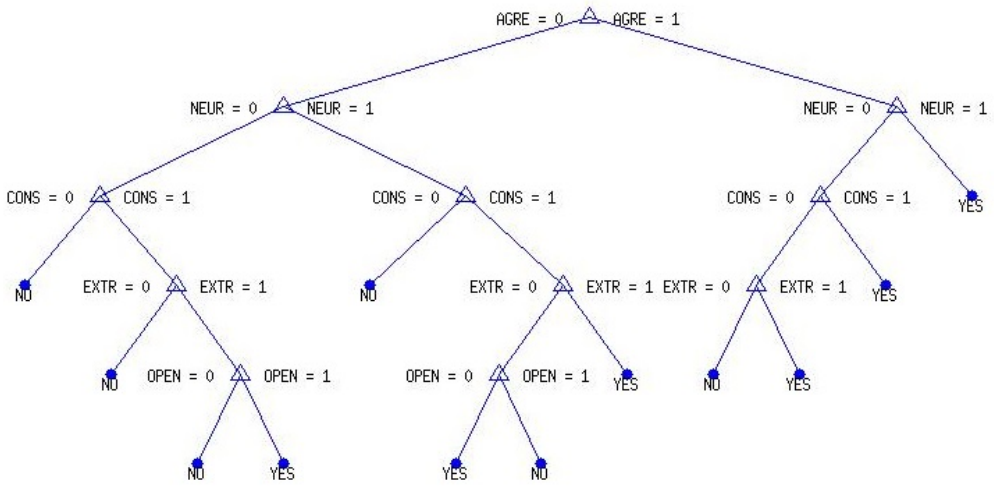


Figure 2: Illustration of the trained decision tree for job interview invitation.

invitation decision, showing the consistency of the proposed approach. Figure 3 illustrates 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?

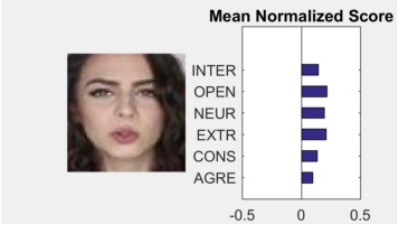
The test set scores of the winners for this stage are shown in Table 3. Our team ranks the first in terms of the overall mean score. However, since the first runner up has better Creativity scores and the mean scores are not significantly different, both teams are designated as winners.

Table 3: Qualitative stage test stage winner teams' scores

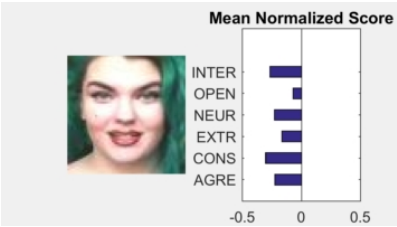
Participant	Our Team	First Runner Up
Clarity	4.31±0.54	3.33±1.43
Explainability	3.58±0.64	3.23±0.87
Soundness	3.40±0.66	3.43±0.92
Interpretability	3.83±0.69	2.40±1.02
Creativity	2.67±0.75	3.40±0.8
Mean Score	**3.56**	3.16

4. Conclusions

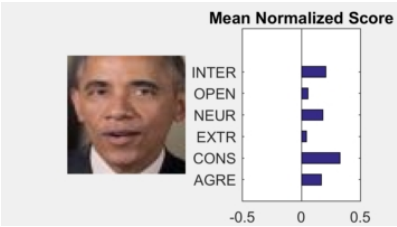
In this paper, we propose a multi-level fusion framework by stacking multi-modal ELM predictions to an ensemble of decision trees (DT). The quantitative results on apparent Big Five trait estimation outperform the state-of-the-art methods that use the same set of audio and video features. The proposed approach leverages the accurate prediction of apparent personality by using them as high level features to predict the interview invitation variable. For the explanability stage, the final trait predictions are mapped to interview variable also using a DT. The DT model eases interpretability and implementation. Thanks to DTs used for fusion and explanation, the proposed systems rank the first in both quantitative and qualitative stages of the official Challenge.



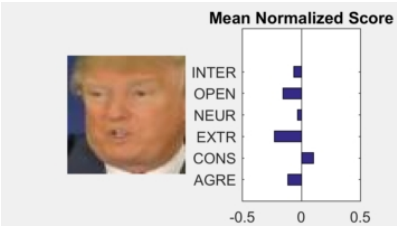
This lady is invited for an interview due to her high apparent agreeableness and neuroticism impression. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.



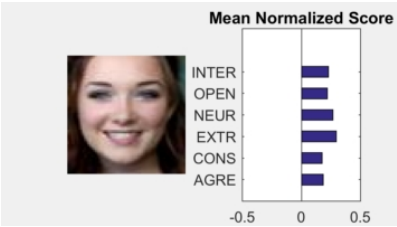
This lady is not invited due to her low apparent agreeableness, neuroticism, conscientiousness, extraversion and openness scores. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.



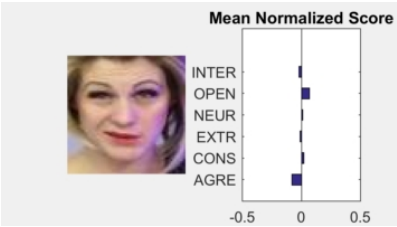
This gentleman is invited for an interview due to his high apparent agreeableness and neuroticism impression. The impressions of agreeableness, conscientousness, extraversion, neuroticism and openness are primarily gained from facial features.



This gentleman is not invited due to his low apparent agreeableness, neuroticism, extraversion and openness scores. The impressions of agreeableness, conscientousness, 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, conscientousness, 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 conscientousness is predominantly modulated by voice.

Figure 3: Sample verbal and visual explanations from qualitative stage.

5. Reproducibility

The scripts to reproduce the results reported for the quantitative and qualitative stages can be accessed over `https://github.com/frkngrpnr/jcs` and `https://github.com/frkngrpnr/jcs_qual`, respectively.

Acknowledgment

We thank the ChaLearn organization and other contributors of this challenge. This work is supported by Boğaziçi University Project BAP 16A01P4 and by the BAGEP Award of the Science Academy.

References

[1] F. Alam and G. Riccardi. Predicting personality traits using multimodal information. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 15–18. ACM, 2014.

[2] F. Alam, E. A. Stepanov, and G. Riccardi. Personality traits recognition on social network-facebook. *WCPR (ICWSM-13), Cambridge, MA, USA*, 2013.

[3] T. R. Almaev and M. F. Valstar. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 356–361. IEEE, 2013.

[4] L. Breiman. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383, 1996.

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

[6] O. Celiktutan and H. Gunes. Automatic prediction of impressions in time and across varying context: Personality, attractiveness and likeability. *IEEE Transactions on Affective Computing*, 8(1):29–42, 2016.

[7] S. Escalera, M. Torres Torres, B. Martinez, X. Baro, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, and M. Valstar. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–8, June 2016.

[8] F. Eyben, M. Wöllmer, and B. Schuller. OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[9] G. Farnadi, S. Sushmita, G. Sitaraman, N. Ton, M. De Cock, and S. Davalos. A multivariate regression approach to personality impression recognition of vloggers. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, pages 1–6. ACM, 2014.

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

[11] S. Gievska and K. Koroveshovski. 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*, pages 19–22. ACM, 2014.

[12] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International Conference on Neural Information Processing*, pages 117–124. Springer, 2013.

[13] F. Gürpınar, H. Kaya, H. Dibeklioğlu, and A. A. Salah. Kernel ELM and CNN Based Facial Age Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 80–86, Las Vegas, Nevada, USA, June 2016.

[14] F. Gürpınar, H. Kaya, and A. A. Salah. Combining deep facial and ambient features for first impression estimation. In *ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop Proceedings*, pages 372–385, 2016.

[15] F. Gürpınar, H. Kaya, and A. A. Salah. 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, December 2016.

[16] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme Learning Machine: a new learning scheme of feedforward neural networks. In *IEEE International Joint Conference on Neural Networks*, volume 2, pages 985–990, 2004.

[17] H. Kaya, F. Gürpınar, and A. A. Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 2017.

[18] H. Kaya and A. A. Salah. Continuous mapping of personality traits: A novel challenge and failure conditions. In *Proceedings of the 2014 ICMI Workshop on Mapping Personality Traits Challenge*, pages 17–24. ACM, 2014.

[19] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 427–434. ACM, 2015.

[20] X. Liu, S. Li, M. Kan, J. Zhang, S. Wu, W. Liu, H. Han, S. Shan, and X. Chen. Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 16–24, 2015.

[21] V. P. Lopez, B. Chen, A. Places, M. Oliu, C. Corneanu, X. Baro, H. J. Escalante, I. Guyon, and S. Escalera. 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*, pages 400–418. Springer, 2016.

[22] N. Madzlan, J. Han, F. Bonin, and N. Campbell. Towards automatic recognition of attitudes: Prosodic analysis of video blogs. *Speech Prosody, Dublin, Ireland*, pages 91–94, 2014.

[23] S. Nowson and A. J. Gill. Look! who's talking?: Projection of extraversion across different social contexts. In *Proceed-*

ings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition, pages 23–26. ACM, 2014.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[25] R. Qin, W. Gao, H. Xu, and Z. Hu. Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. *arXiv preprint arXiv:1604.07499*, 2016.

[26] R. Rothe, R. Timofte, and L. Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 10–15, 2015.

[27] C. Sarkar, S. Bhatia, A. Agarwal, and J. Li. 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*, pages 11–14. ACM, 2014.

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

[29] M. Sidorov, S. Ultes, and A. Schmitt. Automatic recognition of personality traits: A multimodal approach. In *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*, pages 11–15. ACM, 2014.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] F. Valente, S. Kim, and P. Motlicek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *INTERSPEECH*, pages 1183–1186, 2012.

[32] X. Xiong and F. De la Torre. Supervised Descent Method and Its Application to Face Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.