Modeling, Recognizing, and Explaining Apparent Personality from Videos

Escalante, Hugo Jair; Kaya, Heysem; Ali Salah, Albert; Escalera, Sergio; Güç;lütürk, Yağmur; Güçlü, Umut ; Baro, Xavier; Sukma Wicaksana, Achmadnoer ; Liem, Cynthia C.S.; More Authors

DOI
10.1109/TAFFC.2020.2973984

Publication date
2020

Document Version
Final published version

Published in
IEEE Transactions on Affective Computing

Modeling, Recognizing, and Explaining Apparent Personality from Videos

Hugo Jair Escalante*, Heysem Kaya*, Albert Ali Salah*, Sergio Escalera, Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio C. S. Jacques Junior, Meysam Madadi, Stephane Ayache, Evelyne Viegas, Furkan Gürpınar, Achmadnoer Sukma Wicaksana, Cynthia C. S. Liem, Marcel A. J. van Gerven, Rob van Lier

Abstract—Explainability and interpretability are two critical aspects of decision support systems. Despite their importance, it is only recently that researchers are starting to explore these aspects. This paper provides an introduction to explainability and interpretability in the context of apparent personality recognition. To the best of our knowledge, this is the first effort in this direction. We describe a challenge we organized on explainability in first impressions analysis from video. We analyze in detail the newly introduced data set, evaluation protocol, proposed solutions and summarize the results of the challenge. We investigate the issue of bias in detail. Finally, derived from our study, we outline research opportunities that we foresee will be relevant in this area in the near future.

Index Terms—Explainable computer vision, First impressions, Personality analysis, Multimodal information, Algorithmic accountability.

1 Introduction

Explainability and interpretability are critical features of any decision support system [16]. The former focuses on mechanisms that can tell what is the rationale behind the decision or recommendation made by a system or model. The latter focuses on revealing which part(s) of the model structure influences its recommendations. Both aspects are decisive when applications can have serious implications, most notably, in health care, security and education scenarios.

There are models that are explainable and interpretable by their nature, e.g., consider Bayesian networks and decision trees. In fact, explainable and interpretable models have been available for a while for some applications within Artificial Intelligence (AI) and machine learning. However, in affective computing this aspect is only recently receiving proper attention. This is in large part motivated by the developments on deep learning and its clear dominance across many tasks and domains. Although such deep models have succeeded at reaching impressive recognition rates in diverse tasks, they are *black box models*, as one cannot say too much on the way these methods make recommendations or on the structure of the model itself.

This paper comprises a comprehensive study on explainability and interpretability in the context of affective computing. In particular, we focus on those mechanisms in the context of first impressions analysis. The contributions of this paper are as follows. We review concepts and the state of the art on the subject. We describe a challenge we organized on explainability in first impressions analysis from video. We analyze in detail the newly introduced data set, the evaluation protocol, issues of bias, and we summarize the results of the challenge. Finally, derived from our study, we outline research opportunities that we foresee will be decisive in the near future for the development of the explainable computer vision field.

2 Related work

We approach the problem of estimating the apparent personality of people and related variables. Personality and

Initial submission: August 7, 2019. *This work was partially supported by CONACyT - grant No. CB-241306, Spanish Ministry grants No. TIN2016-74946-P and RTI2018-095232-B-C22 (MINECO/FEDER, UE), and CERCA Programme / Generalitat de Catalunya. A.A. Salah was supported by the BAGEP Award of the Science Academy.*

- *H. J. Escalante is with Computer Science Departments at INAOE, and CINVESTAV-Zacatenco, Mexico E-mail: hugojair@inaoep.mx*
- *H. Kaya and A.A. Salah are with Utrecht University, The Netherlands E-mail:{h.kaya, a.a.salah}@uu.nl*
- *S. Escalera is with Universitat de Barcelona and Computer Vision Center, Spain E-mail: sergio@maia.ub.es*
- *Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven and R. van Lier are with Radboud University, Donders Institute for Brain, Cognition and Behaviour, the Netherlands E-mail: {y.gucluturk,u.guclu,m.vangerven,r.vanlier}@donders.ru.nl*
- *X. Baró is with Universitat Oberta de Catalunya and Computer Vision Center, Spain E-mail: xbaro@uoc.edu*
- *I. Guyon is with UPSud/INRIA Université Paris-Saclay, France, and ChaLearn, Berkeley, California E-mail: guyon@chalearn.org*
- *J. C. S. Jacques Junior is with Universitat Oberta de Catalunya and Computer Vision Center, Spain E-mail: jsilveira@uoc.edu*
- *M. Madadi is with Computer Vision Center, Spain, E-mail: mmadadi@cvc.uab.es*
- *S. Ayache is with Laboratoire Informatique Fondamentale, Aix-Marseille Université, France E-mail: stephane.ayache@lif.univ-mrs.fr*
- *E. Viegas is the Director of Artificial Intelligence Outreach at Microsoft Research, Redmond, U.S.A. E-mail: evelynev@microsoft.com*
- *F. Gürpınar is with Boğaziçi University, Turkey E-mail: furkan.gurpinar@boun.edu.tr*
- *A. Sukma Wicaksana is with Delft University of Technology, The Netherlands E-mail: a.s.wicaksana@student.tudelft.nl*
- *C. C. S. Liem is with the Multimedia Computing Group, Delft University of Technology, The Netherlands E-mail: c.c.s.liem@tudelft.nl*

* *Means equal contribution by the authors.*

conduct variables in general are rather difficult to infer precisely from visual inspection, this holds even for humans. Accordingly, the community has being paying attention to a less complex problem, that of estimating *apparent* personality from visual data [56]. Related topics receiving increasing attention are first impressions analysis, depression recognition and hiring recommendation systems [10], [15], [56], [66], all of them starting from visual information. For a comprehensive review on apparent personality analysis from visual information we refer the reader to [34].

Methods for automatic estimation of apparent personality can have a huge impact, as ... *any technology involving understanding, prediction and synthesis of human behavior is likely to benefit from personality computing approaches...* [68]. One of such application is *job candidate screening*. According to [53], video interviews are starting to modify the way in which applicants get hired. The advent of inexpensive sensors and the success of online video platforms has enabled the introduction of a sort of video-based resumé. In comparison with traditional document-based resumés, video-based ones offer the possibility for applicants to show their personality and communication skills. If these sort of resumés are accompanied by additional information (e.g., paper resumé, essays, etc.), recruitment processes can benefit from automated job screening in some initial stages. But more importantly, assessor bias can be estimated with these approaches, leading to fairer selection. On the side of the applicant, this line of research can lead to effective coaching systems to help applicants present themselves better and to increase their chances of being hired. This is precisely the aim of the speed interviews project that gave birth to the present study.

Efforts on automatic video-based analysis of job interview processes are scarce. In [53] the formation of job-related first impressions in online conversational audiovisual resumés is analyzed. Feature representations are extracted from audio and visual modalities. Then, linear relationships between nonverbal behavior and the organizational constructs of "hirability" and personality are examined via correlation analysis. Finnerty et al. [21] aimed to determine whether first impressions of stress are equivalent to physiological measurements of electrodermal activity (EDA). In their work, automatically extracted nonverbal cues, stemming from both the visual and audio modalities were examined. Stress impressions were found to be significantly negatively correlated with "hirability" ratings. In the same line, Naim et al. [51] exploited verbal and nonverbal behaviors in the context of job interviews from face to face interactions. Their approach includes facial expression, language and prosodic information analysis. The framework is capable of making recommendations for a person being interviewed, so that he/she can improve his/her "hirability" score based on the output of a support vector regression model. It is important to note that hiring decisions should not be based on apparent personality or the first impressions created by the candidate. Consequently, automated analysis tools are not for pre-screening the candidates for a job application, but for providing insights into the decisions of the interviewers, and to highlight possible biases in detail.

2.1 Explainability in the modeling of visual information

Following the great success obtained by deep learning based architectures in recent years, different models of this kind have been proposed to approach the problem of first impression analysis from video interviews/resumés or video blogs [25], [26], [67]. Although very competitive results have been reported with such methods (see e.g., [17]), a problem with such models is that they are often perceived as *black-box techniques*: they are able to effectively model very complex problems, but they cannot be interpreted, nor can their predictions be explained [67]. Because of this, explainability and interpretability have received special attention in different fields, see e.g., [12]. In fact, the interest from the community on this topic is evidenced by the organization of dedicated events, such as thematic workshops [37], [38], [50], [70], [71] and challenges [17]. This is particularly important to ensure fairness and to verify that the models are not plagued with various kinds of biases [57], which may have been inadvertently introduced.

Among the efforts for making models more explainable/interpretable, visualization has been seen as a powerful technique to understand how deep neural networks work [41], [47], [69], [73], [75]. These approaches primarily seek to understand what internal representations are formed in the black box model. Although visualization by itself is a convenient formulation to understand model structure, approaches going one step further can also be found in the literature [13], [32], [39], [44], [60], [61]. We refer the reader to [16] for a compilation on recent progress explainability and interpretability in the context of Machine Learning and Computer Vision.

2.2 Explaining and interpreting first impressions

Methods for first impressions analysis developed so far are limited in their explainability and interpretability capabilities. The question of why a particular individual receives a positive (or negative) evaluation deserves special attention, as such methods will influence our lives strongly, once they become more and more common. Recent studies, including those submitted to a workshop we organized - ChaLearn: Explainable Computer Vision Workshop and Job Candidate Screening Competition at CVPR2017[1], sought to address this question. In the remainder of this section, we review these first efforts on explainability and interpretability for first impressions and "hirability" analyses.

Güçlütürk et al. [25] proposed a deep residual network, trained on a large dataset of short YouTube video blogs, for predicting first impressions and whether persons seemed *suitable* to be invited to a job interview. In their work, they use a linear regression model that predicts the interview annotation ("invite for an interview") as a function of personality trait annotations in the five dimensions of the Big-Five personality model. The average "bootstrapped" coefficients of the regression are used to assess the influence of the various traits on hiring decisions. The trait annotations were highly predictive of the interview annotations ($R^2 = 0.9058$), and the predictions were significantly above chance

1. http://openaccess.thecvf.com/CVPR2017_workshops/CVPR2017_W26.py

level ($p \ll 0.001$, permutation test). Conscientiousness had the largest and extroversion had the smallest contributions to the predictions ($\beta > 0.33$ versus $\beta < 0.09$, respectively). For individual decisions, the traits corresponding to the two largest contributions to the decision are considered "explanations". In addition, a visualization scheme based on representative face images was introduced to visualize the similarities and differences between the facial features of the people that were attributed the highest and lowest levels of each trait and interview annotation.

In [26], the authors identified and highlighted the audiovisual information used by their deep residual network through a series of experiments in order to explain its predictions. Predictions were *explained* using different strategies, based either on the visualization of representative face images [25], or using an audio/visual occlusion based analysis. The later involves systematically masking the visual or audio inputs to the network while measuring the changes in predictions as a function of location, predefined region or frequency band. This approach marks the features to which the decision is sensitive (parts of the face, pitch, etc.)

Ventura et al. [67] presented a deep study on understanding why CNN models are performing surprisingly well in automatically inferring first impressions of people talking to a camera. Although their study did not focus on "hirability" systems, results show that the face provides most of the discriminative information for personality trait inference, and the internal CNN representations mainly analyze key face regions such as eyes, nose, and mouth.

Kaya et al. [35] described an end-to-end system for explainable automatic job candidate screening from video interviews. In their work, audio, facial and scene features are extracted. Then, these multiple modalities are fed into modality-specific regressors in order to predict apparent personality traits and "hirability" scores. The base learners are stacked to an ensemble of decision trees to produce quantitative outputs, and a single decision tree, combined with a rule-based algorithm produces interview decision explanations based on quantitative results. Wicaksana and Liem [65] presented a model to predict the Big Five personality trait scores and interviewability of vloggers, explicitly targeting explainability of the system output to humans without technical background. In their work, multimodal feature representations are constructed to capture facial expression, movement, and linguistic information.

2.3 A word of caution

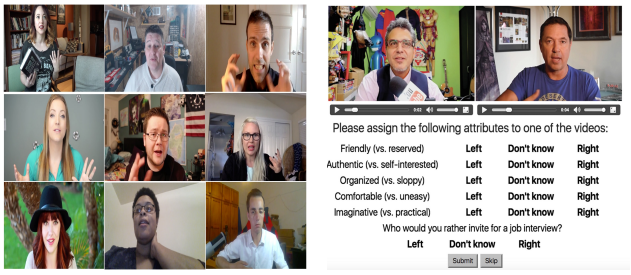
Researchers have made a great progress in different areas of the so called, *looking at people* (LaP) field, as a result of which, human-level performance has almost been achieved on a number of tasks (e.g., face recognition) for controlled settings and adequate training conditions. However, most progress has concentrated on *obviously visual* problems (e.g., gesture recognition). More recently, LaP is targeting problems that deal with subjective assessments, such as first impression estimation. Such systems can be used for understanding and avoiding bias in human assessment, for implementing more natural behaviors, and for training humans in producing adequate social signals. Any task related to social signals in which computers partake in the decision process will benefit from accurate, but also explainable models. Subsequently, this line of research *should not be conceived of implementing systems that may (in some dystopic future) dislike a person's face and deny them a job interview*, but rather look at the face and explain why the biased human assessor denied the job interview.

3 The Job Candidate Screening Coopetition

With the goal of advancing research on explainable models in computer vision, we organized an academic *coopetition* on explainable computer vision and pattern recognition to assess "first impressions" on personality traits. It is called a "coopetition," rather than a competition, because it promoted code sharing between the participants. The 2017 ChaLearn challenge at CVPR was framed in the context of Job Candidate screening. More concretely, the main task of the challenge was to guess the *apparent* first impression judgments on people in video blogs, and whether they would be considered to be invited to a job interview.

3.1 Overview

The challenge relied on a novel data set that we made publicly available recently[2] [15], [56]. The so-called first impressions data set comprises 10,000 clips (with an average duration of 15s) extracted from more than 3,000 different YouTube high-definition videos of people facing a camera and speaking in English. People in videos have different gender, age, nationality, and ethnicity (see Section 5). Figure 1(a) shows snapshots of sample videos from the data set.



(a) (b)

Fig. 1. a) Snapshots of samples from the First Impressions data set [56]. b) Snapshots of the interface for labeling videos [56]. The "big five" traits are characterized by adjectives: Extroversion = Friendly (vs. Reserved); Agreeableness = Authentic (vs. Self-interested); Conscientiousness = Organized (vs. Sloppy); (non-)Neuroticism = Comfortable (vs. Uneasy); Openness = Imaginative (vs. Practical).

In the coopetition, we challenged the participants to provide predictive models with explanatory mechanisms. The recommendation that models had to make was on whether a job candidate should be invited for an interview or not, by using short video clips (see Sec. 3.2). Since this is a decisive recommendation, we thought explainability would be extremely helpful in a scenario in which human resources personnel wants to know what are the reasons of the model for making a recommendation. We assumed that the candidates have already successfully passed technical

2. http://chalearnlap.cvc.uab.es/dataset/24/description/

screening interview steps e.g. based on a CV review. We addressed the part of the interview process related only to **human factors**, complementing aptitudes and competence, which were supposed to have been separately evaluated. Although this setting is simplified, the challenge was a real and representative scenario where explainable computer vision and pattern recognition is highly needed: *a recruiter needs an explanation for the recommendations made by a machine*.

The challenge was part of a larger project on speed interviews[3], whose overall goal is to help both recruiters and job candidates by using automatic recommendations based on multi-media CVs. Also, this challenge was related to two previous 2016 competitions on first impressions that were part of the contest programs of ECCV2016 [56] and ICPR2016 [15]. Both previous challenges focused on predicting the apparent personality of candidates in video. In this version of the challenge, we aimed at predicting **hiring recommendations** in a candidate screening process, i.e. whether a job candidate is worth interviewing. More importantly, we focused on the explanatory power of techniques: *solutions have to "explain" why a given decision was made.* Another distinctive feature of the challenge is that it incorporates a collaboration-competition scheme by rewarding participants who share their code during the challenge, weighting rewards with the usefulness of their code.

3.2 Data Annotation

Videos were labeled both with apparent personality traits and a"*job-interview variable*". The considered personality traits were those from the Five Factor Model (also known as the "Big Five" or OCEAN traits) [48], which is the dominant paradigm in personality research. It models human personality along five dimensions: ***O**penness, **C**onscientiousness, **E**xtroversion, **A**greeableness, **N**euroticism*, respectively. Thus, each clip has ground truth labels for these five traits. Because "Neuroticism" is the only negative trait, we replaced it by its opposite (non-Neuroticism) to score all traits in a similar way on an positive scale. Additionally, each video was labeled with a variable indicating whether the subject should be invited to a job interview or not (the "*job-interview variable*").

Amazon Mechanical Turk (AMT) was used for generating the labels. To avoid calibration problems, we adopted a pairwise ranking approach for labeling the videos: each Turker was shown two videos and asked to answer which of the two subjects present individual traits more strongly. Also, annotators were instructed to indicate which of two subjects they would invite for a job interview. In both cases, a neutral, "I do not know" answer was possible. During labeling, different pairs of videos were given to different and unique annotators. Around 2500 annotators labelled the data, and a total of $321,684$ pairs were used [7]. Although this procedure does not allow us to perform the agreement analysis among annotators which labeled the same pairs, below we report an experiment that aims at assessing the consistency of labellings from AMT workers and human annotators in a more controlled scenario. Figure 1(b) illustrates the interface that AMT workers had access to.

In addition to the audio visual information available in the raw clips, we provided transcripts of the audio. In total, this added about $375,000$ transcribed words for the entire data set. The transcriptions were obtained by using a professional human transcription service[4] to ensure maximum quality for the ground truth annotations.

The feasibility of the challenge annotations was successfully evaluated prior to the start of the challenge. The reconstruction accuracy of all annotations obtained by the BTL model was greater than 0.65 (test accuracy of cardinal rating reconstruction by the model [7]). Furthermore, the apparent trait annotations were highly predictive of invite-for-interview annotations, with a significantly above-chance coefficient of determination of 0.91.

3.2.1 Annotation agreement

Since no video pair was viewed by more than 1 person in the original data collection experiment, in order to estimate the consistency of the personality assessments in the dataset we ran a second experiment with 12 participants (6 males and 6 females, mean age = 27.2). This experiment was in most aspects a replication of the original experiment, the main differences being that the same videos were viewed and assessed by all participants and that it was not an online study. The experiment consisted of viewing a subset of 100 video pairs that were randomly drawn from the original dataset, and then making judgements regarding the personality of the people similarly to the original experiment. Since all participants evaluated the same video pairs in this second experiment, we were able to quantify the consistency of their choices amongst each other. To measure consistency, for each video pair, we calculated the entropy of the distribution of the choices of the participants, and averaged the results per each personality trait. Entropy can take values between 0 and 1, where a low average entropy value represents high consistency for that trait and a high average entropy value represents low consistency. Note that the mapping between consistency and entropy of a distribution is not linear. Our analysis revealed that all traits were significantly more consistent than chance-level ($p << 0.05$, permutation test), with organizedness evaluations that represented conscientiousness trait being the most consistent (entropy = 0.72), and imaginativeness evaluations that represented the openness trait being the least consistent (entropy = 0.85) among participants (See Figure 2).

First impression annotation is a very complex and subjective task. Several aspects can influence the way people perceive others, such as cultural aspects, gender, age, attractiveness, facial expression, among others (from the observer point of view as well as from the perspective of the person being observed). It means that, e.g., different individuals can have very distinct impressions of the same person in an image. Moreover, the same individual can perceive the same person differently at different circunstancies (e.g., time intervals, images or videos) due to many reasons. Subjectivity in data labeling, and more specifically in first impression, is a very challenging task which has attracted a lot of attention

3. http://gesture.chalearn.org/speed-interviews

4. http://www.rev.com

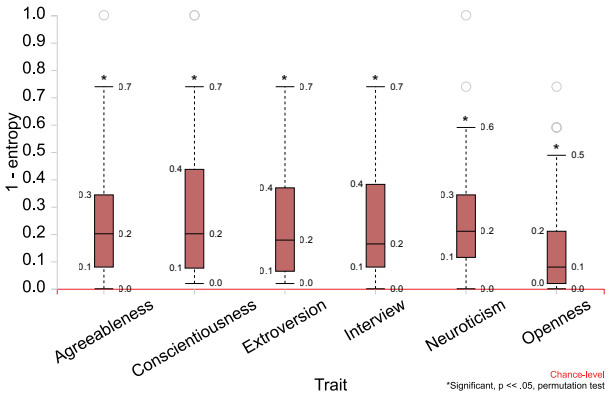


Fig. 2. Consistency estimates of the dataset per each personality trait. For clarity, 1 - entropy of the distribution of the choices of the participants averaged per each personality trait is displayed, i.e. low values mean low consistency and high values mean high consistency. Note that the consistency of the evaluations for all traits were significantly above the chance level.

by the machine learning and computer vision communities.

3.3 Evaluation protocol

The job candidate screening challenge was divided into two tracks/stages, comprising quantitative and qualitative variants of the challenge. The qualitative track being associated to the explainability capabilities of the models developed for the first track. The tracks were run in series as follows:

- **Quantitative competition (first stage).** Predicting whether the candidates are promising enough that the recruiter wants to invite him/her to an interview.
- **Qualitative coopetition (second stage).** Justifying/explaining with an appropriate user interface the recommendation made such that a human can understand it. Code sharing was expected.

Figure 3(a) depicts the information that was evaluated in each stage. In both cases, participants were free (and encouraged) to use information from apparent personality analysis. However, please note that the personality traits labels were provided *only with training data*. This challenge adopted a *coopetition* scheme; participants were expected to share their code and use other participants's code, mainly for the second stage of the challenge: e.g., a team could participate only in the qualitative competition using the solution of another participant in the quantitative competition.

As in other challenges organized by ChaLearn[5], the job candidate screening coopetition ran in CodaLab[6]; a platform developed by Microsoft Research and Stanford University in close collaboration with the organizers of the challenge.

3.3.1 Data partitioning

For the evaluation, the data set was split as follows:

- **Development (training)** data with ground truth for all of the considered variables (including personality traits) was made available at the beginning of the competition.
- **Validation** data **without labels** (neither for personality traits nor for the "job-interview variable") was also provided to participants at the beginning of the competition. Participants could submit their predictions on validation data to the CodaLab platform and received immediate feedback on their performance.
- **Final evaluation (test)** unlabeled data was made available to participants one week before the end of the quantitative challenge. Participants had to submit their predictions in these data to be considered for the final evaluation (no ground truth was released at this point). Only five test set submissions were allowed per team.

In addition to submitting predictions for test data, participants desiring to compete for prizes submitted their code for verification, and a description of their solutions.

3.3.2 Evaluation measures

For explainability, qualitative assessment is crucial. Consequently, we provide some detail about our approach in this section. The competition stages were independently evaluated, as follows:

- **Quantitative evaluation (interview recommendation).** The performance of solutions was evaluated according to their ability for predicting the interview variable in the test data. Specifically, similar in spirit to a regression task, the evaluation consists in computing the accuracy over the invite-for-interview variable, defined as:

$$A = 1 - \frac{1}{N_t} \sum_{i=1}^{Nt} |t_i - p_i| / \sum_{i=1}^{Nt} |t_i - \bar{t}| \qquad (1)$$

  where p_i is the predicted score for sample i, t_i is the corresponding ground truth value, with the sum running over N_t test videos, and $\bar{t}$ is the average ground truth score over all videos.
- **Qualitative evaluation (explanatory mechanisms).** Participants had to provide a textual description that explains the decision made for the interview variable. Optionally, participants could also submit a visual description to enrich and improve clarity and explainability. Performance was evaluated in terms of the creativity of participants and the explanatory effectiveness of the descriptions. For this evaluation, we invited a set of experts in the fields of psychological behavior analysis, recruitment, machine learning and computer vision.
  Since the explainability component of the challenge requires qualitative evaluations and hence human effort, the scoring of participants was made based on a small subset of the videos. Specifically, subsets of videos from the validation and test sets were systematically selected to better represent the variability of the personality traits and invite-for-interview values in the entire dataset. The jury only evaluated a single validation and a single test phase submission per participant. A separate jury member served as a tiebreaker. At the end, the creativity criterion was judged globally, according to the evaluated clips, as

5. http://chalearn.org
6. http://codalab.org/

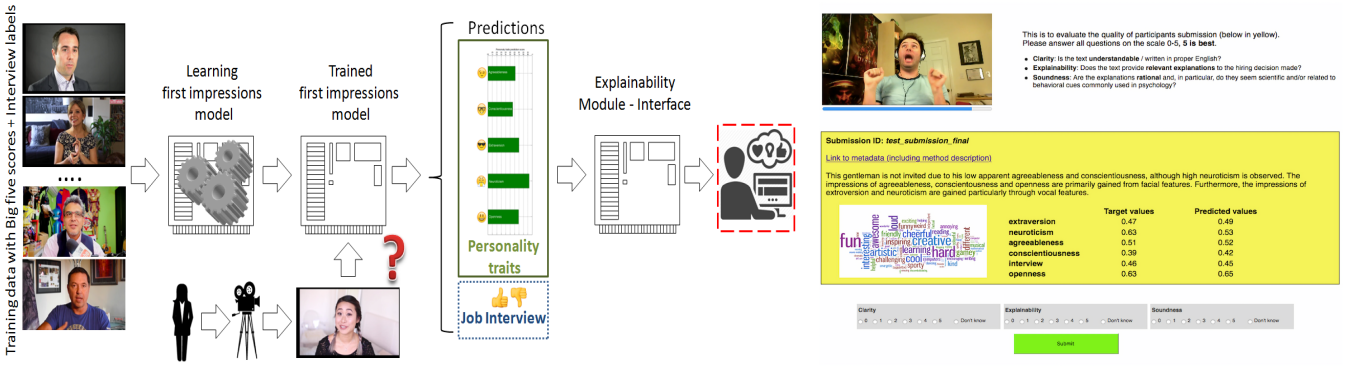


Fig. 3. a) Diagram of the considered scenario in the job candidate screening coopetition. The solid (green) top square indicates the variables evaluated in past editions of the challenge [15], [56]. The dotted (blue) bottom square indicates the variable evaluated in the quantitative track. The dashed (red) square indicates what is evaluated in the qualitative track. b) Qualitative evaluation interface. The explainable interface of a submission is shown to the judge who had to evaluate it along the considered dimensions.

well as an optional video that participants could submit to describe their method. Figure 3(b) shows an illustration of the interface used by the jury for the qualitative evaluation phase.

For each evaluated clip, the evaluation criteria for the jury were:

- *Clarity:* Is the text understandable / written in proper English?
- *Explainability:* Does the text provide relevant explanations on 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?

The following two criteria were evaluated globally, based on the evaluated clips and the optional submitted video.

- *Model interpretability:* Are the explanations useful to understand the functioning of the predictive model?
- *Creativity:* How original / creative are the explanations?

- **Coopetition evaluation (code sharing).** Participants were evaluated by the usefulness of their shared code in the collaborative competition scheme. The coopetition scheme was implemented in the second stage of the challenge.

3.4 Baselines

We considered several baselines for solving the aforementioned tasks in different input modalities. Please note that the factors underlying the predictions of the (baseline) models for the quantitative phase were investigated in these two earlier publications [25], [30]. Briefly, these two studies show that many factors including face features, gender, audio features and context have varying contributions to the predictions of different traits. Also note that we only describe the baseline adopted for the first stage for the quantitative stage of the challenge. For the qualitative evaluation we built a demo[7] successfully presented at the demo session of NIPS2016[8]. The purpose of the demo was to give an idea to participants on possibilities for designing their systems.

3.4.1 *Language models: audio transcripts*

We evaluated two different language models, each on the same modality (transcriptions). Both of the models were a variation of the following (linearized) ridge regression model: $y = \text{embedding}(\mathbf{x})\boldsymbol{\beta} + \varepsilon$, where y is the annotation, $\mathbf{x}$ is the transcription, $\boldsymbol{\beta}$ represents the parameters and ε is the error term. This formulation describes a (nonlinear) embedding, followed by a (linear) fully-connected computation. Both models were trained by analytically minimizing the L2 penalized least squares loss function on the training set, and model selection was performed on the validation set.

3.4.1.1 Bag-of-words model.: This model uses an embedding that represents transcripts as 5000-dimensional vectors, i.e. the counts of the 5000 most frequent non-stopwords in the transcriptions.

3.4.1.2 Skip-thought vectors model.: This model uses an embedding that represents transcripts as 4800-dimensional mean skip-thought vectors [43] of the sentences in the transcriptions. A recurrent encoder-decoder neural network pretrained on the BookCorpus dataset [74] was used for extracting the skip-thought vectors from the transcriptions.

3.4.2 *Sensory models: audio visual information processing*

We evaluated three different sensory models, each on a different modality (audio, visual, and audio visual, respectively). All models were a variation of the 18-layer deep residual neural network (ResNet18) in [31]. As such, they comprised several convolutional layers followed by rectified linear units and batch normalization, and connected to one another with (convolutional or identity) shortcuts, as well as a final (linear) fully-connected layer preceded by global average pooling. The models were trained by minimizing the

7. http://sergioescalera.com/wp-content/uploads/2016/12/TraitVideo.mp4
8. https://nips.cc/Conferences/2016/Schedule?showEvent=6314

mean absolute error loss function iteratively with stochastic gradient descent (Adam [42]) on the training set, and model selection was performed on the validation set.

3.4.2.1 Audio model.: This model is a variant of the original ResNet18 model, in which $n \times n$ inputs, kernels, and strides are changed to $n^2 \times 1$ inputs, kernels, and strides [24], as well as changing the size of the last layer to account for the different number of outputs. Prior to entering the model, the audio data were temporally preprocessed to 16kHz. The model was trained on random 3s crops of the audio data and tested on the entire audio data.

3.4.2.2 Visual model.: This model is a variant of the original ResNet18 model, in which the size of the last layer is changed to account for the different number of outputs. Prior to entering the model, the visual data are spatiotemporally preprocessed to 456×256 pixels and 25 frames per second. The model was trained on random 224×224 pixel single frame crops of the visual data and tested on the entire visual data.

3.4.2.3 Audiovisual model.: This model is obtained by a late fusion of the audio and visual models. The late fusion took place after the global average pooling layers of the models via concatenation of their latent features. The entire model was jointly trained from scratch.

3.4.3 Language and sensory model

3.4.3.1 Skip-thought vectors and audiovisual model.: This model is obtained by a late fusion of the pretrained skip-thought vectors and audiovisual models. The late fusion took place after the embedding layer of skip-thought vectors model and the global average pooling layer of the audiovisual model via concatenation of their latent features. Only the last layer was trained from scratch and the rest of the layers were fixed.

3.4.4 Results

The baseline models were used to predict the trait annotations as a function of the language and/or sensory data. Table 1 shows the baseline results. The language models had the lowest overall performance with skip-thought vectors model performing better than the bag-of-word model. The performance of the sensory models were better than those of the language models with the audiovisual fusion model having the highest performance and the audio model having the lowest performance. Among all models, the language and sensory fusion model (skip-thought vectors and audiovisual fusion model) achieved the best performance. All prediction accuracies were significantly above the chance-level ($p < 0.05$, permutation test), and were consistently improved by fusing more modalities.

TABLE 1
Baseline results. Results are reported in terms of 1 - relative mean absolute error on the test set. AGR: Agreeableness; CON: Conscientiousness; EXT: Extroversion; $\overline{NEU}$: (non-)Neuroticism; OPE: Openness; AVE: average over trait results; INT: interview.

Model	AGR	CON	EXT	$\overline{NEU}$:	OPE	AVE	INT
language							
bag-of-words	0.8952	0.8786	0.8815	0.8794	0.8875	0.8844	0.8845
Skip-Thought Vec.	0.8971	0.8819	0.8839	0.8827	0.8881	0.8867	0.8865
sensory							
audio	0.9034	0.8966	0.8994	0.9000	0.9024	0.9004	0.9032
visual	0.9059	0.9073	0.9019	0.8997	0.9045	0.9039	0.9076
Audio-Visual	0.9102	0.9138	0.9107	0.9089	0.9111	0.9109	0.9159
language+sensory							
STV + AV	*0.9112*	*0.9152*	*0.9112*	*0.9104*	*0.9111*	*0.9118*	*0.9162*

4 TWO EXPLAINABLE SYSTEMS

This section provides a detailed description of two systems that completed the second stage of the job candidate screening challenge.

4.1 BU-NKU: Explainabilty with decision trees

The BU-NKU system is based on audio, video, and scene features, similar to the system that won the ChaLearn First Impression Challenge at ICPR 2016 [29]. Here, the face, scene, and audio modalities are first combined at feature level, followed by stacking the predictions of sub-systems to an ensemble of decision trees [35]. The flow of this system is illustrated in Figure 4.

For the qualitative stage, the idea is to produce intermediate interpretable variables (i.e. apparent personality traits), and base the interview decision on these. To generate the explanations, the proposed system takes the apparent personality trait predictions, discretizes them into low/high, and maps them to the binarized (invite/do not invite) interview variable via a decision tree (DT). DT is employed to allow visualization and ease of interpretation for the model. This tree is traced to generate a verbal explanation.

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. Faces are detected, aligned, and resized to 64×64 pixels. A deep neural network pre-trained with VGG-Face [55] and finetuned with FER-2013 database [23] is used.

After extracting frame-level features from each aligned face using the deep neural network, videos are summarized by computing functional statistics of each dimension over time, including mean, standard deviation, offset, slope, and curvature. Facial features are combined with the Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) video descriptor, shown to be effective in emotion recognition [1], [36].

In order to use ambient information in the images, a set of features is extracted using the VGG-VD-19 network [62], which is trained for an object recognition task on the ILSVRC 2012 dataset. Similar to face features, a 4 096-dimensional representation from the 39^{th} layer of the 43-layer architecture is used. This gives 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 [28], [29].

The open-source openSMILE tool [20] is popularly used to extract acoustic features in a number of international paralinguistic and multi-modal challenges. The BU-NKU approach uses the toolbox with a standard feature configuration that served as the challenge baseline sets in INTERSPEECH 2013 Computational Paralinguistics Challenge [59]. This configuration was found to be the most

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.2973984, IEEE Transactions on Affective Computing

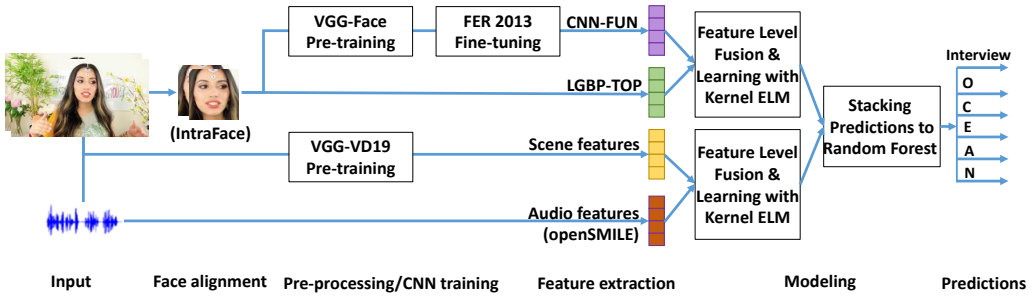


Fig. 4. Flowchart of the BU-NKU system.

effective acoustic feature set among others for personality trait recognition [29].

In order to model personality traits from audio-visual features, kernel extreme learning machines (ELM) were used [33]. 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 [6]. 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.

The validation set performances of individual features, as well as their feature-, score- and multi-level fusion alternatives are shown in Table 2. Here, System 0 corresponds to the top entry in the ICPR 2016 Challenge [29], 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. Systems 1 to 6 are sub-components of the proposed system, namely System 8, whereas System 7 is a score fusion alternative that uses linear weights instead of a Random Forest. Systems 1 and 2 are trained with facial features as explained before: VGGFER33 is 33^{rd} layer output of FER fine-tuned VGG CNN and LGBPTOP is also extracted from face. These two facial features are combined in the proposed framework, and their feature-level fusion performance is shown as System 5. Similarly, Systems 3 (scene sub-system) and 4 (audio sub-system) are combined at feature level as System 6.

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 2) 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.

TABLE 2
Validation set performance of the BU-NKU system and its sub-systems, using the performance measure of the challenge (1-relative mean abs error). FF: Feature-level fusion, WF: Weighted score-level fusion, RF: Random Forest based score-level fusion. INTER: Interview invite variable. $AGRE$: Agreeableness. $CONS$: Conscientiousness. $EXTR$: Extroversion. $\overline{NEUR}$: (non-)Neuroticism. $OPEN$: Openness to experience.

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

4.1.1 Qualitative System

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 decision tree classifier. The proposed approach for decision explanation uses the trace of each decision from the root of the tree to the leaf. 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.

The DT trained on the predicted OCEAN dimensions gives a classification accuracy of 94.2% for binarized interview variable. The illustration of the trained DT is given in Figure 5.

The model is intuitive as 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 Extroversion. Figure 6 illustrates automatically generated verbal and visual explanations for this stage.

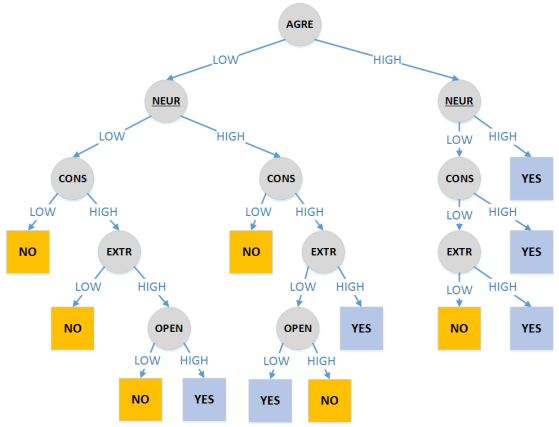


Fig. 5. Illustration of the decision tree for job interview invitation. NEUR denotes (non-) Neuroticism. Leaves denote a positive or a negative invitation response.

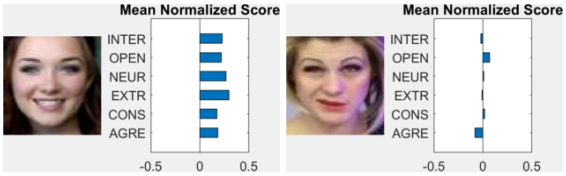


Fig. 6. Sample verbal and visual explanations from qualitative stage for the BU-NKU entry. **Left image:** This lady is invited for an interview due to her high apparent agreeableness and non-neuroticism impression. The impressions of agreeableness, conscientiousness, extroversion, non-neuroticism and openness are primarily gained from facial features. **Right image:** This lady is not invited for an interview due to her low apparent agreeableness and extraversion impressions, although predicted scores for non-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, non-neuroticism and openness are primarily gained from facial features. Furthermore, the impression of conscientousness is predominantly modulated by voice.

4.2 TUD: Explainability via Linear models

This section describes the TUD approach for the second stage of the job candidate screening challenge. This system was particularly designed to give assistance to a human assessor. The proposed model employs features that can easily be described in natural language, with a linear (PCA) transformation to reduce dimensionality, and simple linear regression models for predicting scores, such that scores can be traced back to and justified with the underlying features. While state-of-the-art automatic solutions rarely use hand-crafted features and models of such simplicity, there are clear gains in explainability. As demonstrated within the ChaLearn benchmarking campaign, this model did not obtain the strongest quantitative results, but the human-readable descriptions it generated were well appreciated by human judges.

The model considers two modalities, visual and textual, for extracting features. In the visual modality, it considers features capturing facial movement and expression, as they are one of the best indicators for personality [5], [52]. Specifically, action units (AUs) extracted from segmented frames depicting the face of the user were considered.

For each of the considered AUs, three features were generated: (1) the percentage of time frames is computed, during which the AU was visible in a video; (2)the maximum intensity of the AU in the video was stored; (3) the mean intensity of the AU over the video was also recorded. A total of 18 features were extracted using the OpenFace tool [2], resulting in 54 OpenFace features.

In addition to AUs based features, a Weighted Motion Energy Image (wMEI) is constructed [4]s. The base face image of each video is obtained and the overall movement for each pixel is computed over video frames. For each wMEI, three statistical features (mean, median, and entropy) are extracted to constitute a MEI representation.

Textual features are generated by using transcripts that were provided as the extension of the ChaLearn dataset. As reported in the literature [9], [58] and confirmed in private discussions with organizational psychologists, assessment of GMA (intelligence, cognitive ability) is important for many hiring decisions. While this information is not reflected in personality traits, language usage of the subjects may possibly reveal some related information. To assess that, several Readability indices were used with the transcripts. This was done by using open source implementations of various readability measures in the NLTK toolkit. Eight measures were selected as features for the Readability representation: ARI [63], Flesch Reading Ease [22], Flesch-Kincaid Grade Level [40], Gunning Fog Index [27], SMOG Index [49], Coleman Liau Index [8], LIX, and RIX [3]. While these measures are originally developed for written text (and ordinarily may need longer textual input than a few sentences in a transcript), they do reflect complexity in language usage. In addition, two simple statistical features were used for an overall Text representation: total word count in the transcript, and the amount of unique words within the transcript, respectively.

The feature spaces considered by the TUD predictive model are: OpenFace, wMEW, Readability, and Text based features. A separate model per representation was trained to predict personality traits and interview scores. For a final prediction score, late fusion is used and the predictions made by the four different models were averaged. A diagram of the proposed system can be seen in Figure 7.

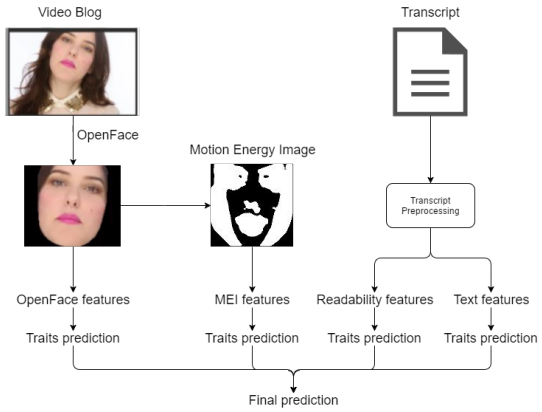


Fig. 7. Overall system diagram for the TUD system.

As one of the goals of the system is to trace back the prediction scores to each underlying feature, linear models were selected. First, principal Component Analysis (PCA) is

used to reduce dimensionality, retaining 90% variance. The resulting transformed features are used as input for a simple linear regression model to predict the scores.

TABLE 3
Performance comparison between the TUD system and lowest / highest performance for each prediction category in the ChaLearn CVPR 2017 Quantitative Challenge.

Categories	TUD System	Lowest	Highest
Interview	0.8877	0.8721	0.9209
Agreeableness	0.8968	0.8910	0.9137
Conscientiousness	0.8800	0.8659	0.9197
Extroversion	0.8870	0.8788	0.9212
Neuroticism	0.8848	0.8632	0.9146
Openness	0.8903	0.8747	0.9170

Table 3 shows the overall test accuracy of the TUD system, for each of the Big Five personality traits and the interview invitation assessment. For each predicted class, scores are compared to the lowest and highest scores of participants in the challenge While the system did not achieve the top scores, it was parsimonious in its use of computational resources, and the linear models allowed easier explainability, see neext section. This is clearly a trade-off in such systems, as more parameters in the model and increased complexity makes interpretation more difficult.

4.2.1 Qualitative System

The TUD team implemented a text descriptor generator based on information derived from features and the model aiming to explain the model recommendations. Please note that the raw values of features or coefficients of linear models were not used directly, this is because it has not been proved that indeed the considered features are indicators of intervewability. However, features and coefficients were used indirectly as follows. For each sample, it was first indicated whether the person scores were "unusually" high-low with respect to the population of "representative subjects", formed by the vloggers represented in the 6000-video training set. Therefore, for each feature measurement, the system reports in a fragment of text the typical range of the features and the percentile of the subject, compared to scores in the training set.

In addition, to derive indicators from the linear model and reflect them in the description, for each representation (OpenFace, MEI, Readability, Text) the two largest linear regression coefficients were identified. Then, for PCA dimensions corresponding to these coefficients, the features contributing most strongly to these dimensions were traced back, and their sign is checked. For these features, a short text description is added, expressing how the feature commonly affects the final scoring (e.g. 'In the model, a higher score on this feature typically leads to a higher overall assessment score') for a positive linear contribution.

As a result, for each video in the validation and test sets, a fairly long, but consistent textual description was generated. An example fragment of the description is given in Figure 8.

```
******************
* USE OF LANGUAGE *
******************

Here is the report on the person's language use:

** FEATURES OBTAINED FROM SIMPLE TEXT ANALYSIS **
Cognitive capability may be important for the job. I looked at a few very simple
text statistics first.

*** Amount of spoken words ***
This feature typically ranges between 0.000000 and 90.000000. The score for this
video is 47.000000 (percentile: 62).
In our model, a higher score on this feature typically leads to a higher overall
assessment score.
```

Fig. 8. Example description fragment for the TUD system.

4.3 Challenge results

4.3.1 Stage 1 results: Recognizing first impressions

For the first stage, 72 participants registered for the challenge. Four valid submissions were considered for the prizes as summarized in Table 4. The leading evaluation measure is the Recall of the Invite-for-interview variable (see Equation 1), although we also show results for personality traits.

TABLE 4
Results of the first stage of the job screening coopetition. * Leading evaluation measure.

Rank	**Team**	INTER	*AGRE*	*CONS*	*EXTR*	*NEUR*	*OPEN*
1	BU-NKU [35]	0.9209 (1)	0.9137 (1)	0.9197 (1)	0.9212 (1)	0.9146 (1)	0.9170 (1)
-	Baseline [25]	0.9162 (2)	0.9112 (2)	0.9152 (2)	0.9112 (3)	0.9103 (2)	0.9111 (2)
2	PML [14]	0.9157 (3)	0.9103 (3)	0.9137 (3)	0.9155 (2)	0.9082 (3)	0.9100 (3)
3	ROCHCI	0.9018 (4)	0.9032 (4)	0.8949 (4)	0.9026 (4)	0.9011 (4)	0.9047 (4)
4	FDMB	0.8721 (5)	0.8910 (5)	0.8659 (5)	0.8788 (5)	0.8632 (5)	0.8747 (5)

The performance of the top three methodologies was quite similar, however the methodologies were not. In the following, we provide a short description of the different methods.

- **BU-NKU.** See Section 4.1, and [35].
- **PML. [14]** Adopted a purely visual approach based on multi-level appearance. After face detection and normalization, Local Phase Quantization (LPQ) and Binarized Statistical Image Features (BSIF) descriptors were extracted at different scales of each frame using a grid. Feature vectors from each region and each resolution were concatenated, the representation for a video was obtained by averaging the per-frame descriptors. For prediction, the authors resorted in a stacking formulation: personality traits are predicted with Support Vector regression (SVR), the outputs of these models are used as inputs for the final decision model, which, using Gaussian processes, estimates the invite for interview variable.
- **ROCHCI.** Extracted a set of predefined multi-modal features and used gradient boosting for predicting the interview variable. Facial features and meta attributes extracted with SHORE[9] were used as visual

9. https://www.iis.fraunhofer.de/en/ff/bsy/tech/bildanalyse/shore-gesichtsdetektion.html

descriptors. Pitch and intensity attributes were extracted from the audio signal. Finally, hand picked terms were used from the ASR transcriptions. The three type of features were concatenated and gradient boosting regression was applied for predicting traits and interview variable.

- **FDMB.** Used frame differences and appearance descriptors at multiple fixed image regions with a SVR method for predicting the interview variable and the five personality traits. After face detection and normalization, differences between consecutive frames was extracted. LPQ descriptors were extracted from each region in each frame and were concatenated. The video representation was obtained by adding image-level descriptor. SVR was used to estimate traits and the interview variable.

It was encouraging that the teams that completed the final phase of the first stage proposed methods that relied on diverse and complementary features and learning procedures. In fact, it is quite interesting that solutions based on deep learning were not that popular for this stage. This is in contrast with previous challenges in most aspects of computer vision (see e.g. [18]), including the first impressions challenge [15], [56]. In terms of the information/modalities used, all participants considered visual information, through features derived from faces and even context. Audio was also considered by two out of the four teams. Whereas ASR transcripts were used only by a single team. Finally, information fusion was performed at a feature level.

4.3.2 Stage 2 results: Explaining recommendations

The two teams completing the final phase of the qualitative stage were BU-NKU and TUD, and their approaches were detailed in previous subsections. Other teams also developed solutions to the explainability track, but did not succeed in submitting predictions for the test videos. BU-NKU and TUD were tied for the first place in the second stage. Table 5 shows the results of participants in the explainability stage of the challenge. Recall that a committee of experts evaluated a sample of videos labeled with each methodology, using the measures described in Section 3.3.2, and a [0,5] scale was adopted. It can be seen from this table that both methods obtained comparable performance.

TABLE 5
Results of the second stage of the job screening coopetition.

Rank	Team	Clarity	Explainability	Soundness	Interpretability	Creativity	Mean score
1	BU-NKU	4.31	3.58	3.4	3.83	2.67	3.56
1	TUD	3.33	3.23	3.43	2.4	3.4	3.16

The performances in Table 5 illustrate that there is room for improvement for developing proper explanations. In particular, evaluation measures for explainability deserve further attention.

5 Analysis of the First Impressions Dataset

The collected personality traits dataset is rich in terms of the number of videos and annotations, and hence suitable for training models with high generalization power. The ground truth annotations used in training models are those given by individuals and may reflect their bias/preconception towards the person in the video, even though it may be unintentional and subconscious. Thus, the classifiers trained can inherently contain this subjective bias.

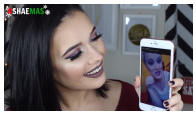
In the next subsection we analyze different aspects of the First Impression database, such as the video split procedure used to build the dataset (Sec. 5.1), the intra/inter-video labeling variation (Sec. 5.2), and subjective bias with respect to gender, ethnicity (Sec. 5.3) and age (Sec. 5.4). We finally handle the problem as a binary classification task, which provides more room for improvement using multimodal approaches (Sec. 5.5).

5.1 Video Split

As briefly mentioned at the beginning of Section 3.1, the 10k clips of the First Impression dataset were obtained by partitioning 3k videos obtained from YouTube into small clips of 15 seconds each. Some of these small clips, generated from the same video, can be found in the train, validation and test set. However, it should be noted that even though different clips generated fom the same video can be found in different sets, they were captured at different time intervals. Thus, both the scene as well as the person appearing on each clip can have high variation in appearance due to different views/poses, camera motion and behavior.

We analyzed the percentage of clips generated from the same video that are contained in different sets and results show that 83.7% of the clips in the validation set have at least one clip in the train set which was generate from the same video. Similarly, 84% of the clips in the test set have at least one clip in the train set which was generated from the same video. From the exact $3,060$ original videos, 721 have not been split and $2,339$ have been split at least once. The average split per video for the whole dataset is 3.27 (± 1.8) and the maximum number of splits found is 6. Figure 9 shows two frames obtained from the same video but at different splits and respective labels associated to each clip, to illustrate the intra-video labeling variation.

(a) {0.67, 0.81, 0.74, 0.53, 0.55, 0.63}



(b) {0.46, 0.62, 0.63, 0.44, 0.42, 0.47}

Fig. 9. Two frames extracted from different clips (generated from the same video) and respective labels {Interview, O, C, E, A, N}, which illustrate the intra-video variation in appearance/pose and labeling.

5.2 Intra-Video and Inter-Video labeling variation

In this section we analyze the labeling variation with respect to different clips generated from the same video (intra-video), as well as different videos from the same YouTube user (inter-video). For the former, we compute the standard deviation per trait for each video which has at least one split (i.e., number of clips > 1). Then, we computed the average deviation per video, taking into account the deviation for all traits. Finally, given the average deviation per video, we

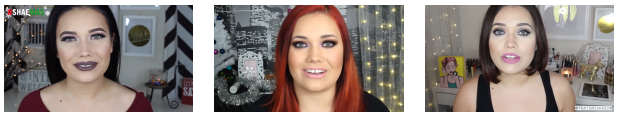
compute the global average deviation with respect to the whole dataset. Results are shown in Table 6.

TABLE 6
Intra-Video and Inter-Video variation analysis. Second and third columns show the average standard deviation per video (intra-video) and per user (inter-video) for the 6 variable under analysis, i.e., {Interview, O, C, E, A, N}. "Global avg deviation" represents the mean standard deviation over all traits with respect to the whole dataset (following the intra/inter-video procedures described in this section).

	Intra-video	Inter-video
Interview	0.064 ($\pm$0.033)	0.054 ($\pm$0.040)
O	0.070 ($\pm$0.034)	0.059 ($\pm$0.044)
C	0.063 ($\pm$0.031)	0.055 ($\pm$0.041)
E	0.068 ($\pm$0.035)	0.056 ($\pm$0.044)
A	0.071 ($\pm$0.035)	0.048 ($\pm$0.037)
N	0.071 ($\pm$0.034)	0.052 ($\pm$0.040)
Avg.	0.068 ($\pm$0.024)	0.054 ($\pm$0.032)

For the inter-video analysis, we first grouped all videos from each user (although some users might have videos from different people, we assume most videos of the same user are from the same individual, i.e., having the same individual appearing on them). First, consider the 10k videos in our dataset have been obtained from around 2762 YouTube users (i.e., 313 videos have no user associated with and are not included in this analysis). The average number of videos per user is 1.07 ($\pm$0.29) and the maximum number of videos per user is 4. The average number of clips (after split) per user is 3.51 ($\pm$2.12) and the maximum number of clips per user is 16. The procedure described next is performed for all users with number of videos > 1 (i.e., 181 users).

For each user, within these 181 users, we first computed the average label per trait for each video (assuming each video can be split into different clips). Then, we computed the standard deviation per trait taking into account the precomputed "average labels" with respect to all videos of a specific user. Then, we computed the average deviation over all traits with respect to each user, and finally, given the average deviation of all traits/users, we computed the global average deviation, see Table 6.



(a) {0.67, 0.81, 0.74, 0.53, 0.55, 0.63}

(b) {0.58, 0.79, 0.58, 0.66, 0.66, 0.62}

(c) {0.74, 0.83, 0.78, 0.62, 0.68, 0.64}

Fig. 10. Three frames extracted from different videos of the same user (i.e., with the same person appearing on them) and respective labels {Interview, O, C, E, A, N}, which illustrate the inter-video variation in appearance and labeling.

It can be observed from Table 6 that intra-video variation is higher than inter-video variation. This may happened because each set of pairs was annotated by a unique annotator. Thus, different clips of the same video may be annotated by different annotators. It can be also observed that there is not a global trend with respect to all analyzed traits and scenarios, i.e., the trait with lower/higher variation in the inter-video scenario may not have the lower/higher variation in the intra-video scheme. To illustrate the inter-video variation, Figure 10 shows diferent frames, obtained from different videos of the same user, and respective labels.

5.3 Gender and Ethnicity analysis

In this section, existence of this latent bias towards gender and apparent ethnicity is analyzed. For this purpose, the videos used in the challenge are further manually annotated for gender and ethnicity, to complement the challenge meta data. Then a linear (Pearson) correlation analysis is carried out between these traits and apparent personality annotations. The results are summarized in Table 7. Although the correlations range from weak to moderate, the statistical strength of the relationships are very high.

We first observe that there is an overall positive attitude/preconception towards females in both personality traits (except Agreeableness) and job interview invitation. The second observation is that the gender bias is stronger compared to ethnicity bias. Concerning the ethnicities, the results indicate an overall positive bias towards Caucasians, and a negative bias towards African-Americans. There is no discernible bias towards Asians in either way.

TABLE 7
Pearson correlations between annotations of gender-ethnicity versus personality traits and interview invitation. * and ** indicate significance of correlation with $p < 0.001$ and $p < 10^{-6}$, respectively.

Correlation	Gender	Ethnicity		
Dimension	Female	Asian	Caucasian	Afro-American
Agreeableness	-0.023	-0.002	**0.061****	**-0.068****
Conscientiousness	**0.081****	0.018	**0.056****	**-0.074****
Extroversion	**0.207****	**0.039***	**0.039***	**-0.068****
Neuroticism	**0.054***	-0.002	**0.047***	**-0.053****
Openness	**0.169****	0.010	**0.083****	**-0.100****
Interview	**0.069****	0.015	**0.052***	**-0.068****

When correlations are analyzed closely, we see that women are perceived as more "open" and "extroverted" compared to men, noting that the same but negated correlations apply for men. It is also seen that women have higher prior chances to be invited for a job interview. We observe a similar, but negative correlation with the apparent Afro-American ethnicity. To quantify these, we first measure the trait-wise means from the development set, comprised of 8000 videos. We then binarize the interview variable using the global mean score, and compute prior probability of job invitation conditioned on gender and ethnicity traits. The results summarized in Table 8 clearly indicate a difference in the chances for males and females to be invited for a job interview. Furthermore, the conditional prior probabilities show that Asians have an even higher chance to be called for a job interview compared to Caucasian ethnicity, while Afro-Americans are disfavored. Since these biases are present in the annotations, supervised learning will result in systems with similar biases. Such algorithmic biases should be made explicit for preventing the misuse of automatic systems.

TABLE 8
Gender and ethnicity based mean scores and conditional prior probabilities for job interview invitation.

	Male	Female	Asian	Caucasian	Afro-American
mean scores	0.539	0.589	0.515	0.507	0.475
p(invite | trait)	0.495	0.560	0.562	0.539	0.444

5.4 Age analysis

We annotated the subjects into eight disjoint age groups using the first image of each video. The people on the videos are classified into one of the following groups: 0-6, 7-13, 14-18, 19-24, 25-32, 33-45, 46-60 and 61+ years old. We excluded the 13 subjects under 14 years old from the analysis. We subsequently analyzed the prior probability of job interview invitation for each age group, with and without gender breakdown. Apart from the ground truth annotations, we also analyzed the held out test set predictions from BU-NKU (winner) system explained in Section 4.1. The results are summarized in Figure 11, where"(Anno)" refers to statistics from ground truth annotations and "(Sys Pred)" are those from the test set predictions.

Regarding the ground truth annotations, on the overall, the prior probability is lower than 0.5 (chance level) for people under 19 or over 60. This is understandable, as very young or old people may not be seen (legally and/or physically) fit to work. For people whose age range from 19 to 60 (i.e. working-age groups), the invitation chance is slightly (but not significantly) higher than the chance level. Within the working-age groups, the female prior probability peaks at 19-24 age group and decreases with increasing age, while for the male gender the prior probability of job invitation steadily increases with age. The analysis shows that annotators preferred to invite women when they are younger and men when they are older to a job interview. This is also verified with correlation analysis: the Pearson correlation between the ordinal age group labels and ground truth interview scores are 0.126 ($p < 10^{-13}$) and 0.074 ($p < 10^{-5}$) for male and female gender, respectively. The results indicate that likability/fitness may be an underlying factor in female job invitation preference. For males, the preference may be attributed to the perceived experience and authority of the subject.

While the analysis of the ground truth reflects a latent bias towards age and gender combination, we see that the classifiers trained on these annotations implicitly model the bias patterns. For male candidates, the classifier exhibits the same age-gender bias pattern for job interview invitation compared to the ground truth and the statistics for males older than 25 years are very similar in both. A similar pattern exists for the female candidates however peaking at the age of 25-32, instead of 19-24. The largest difference between the ground truth and the classifier statistics is observed in non-working-age groups (14-18 and 61+) and particularly among females. This may be attributed to the fact that these groups form a small proportion (3.5%) of the data.

The analysis provided in Sections 5.3 and 5.4 evidences potential biases for systems trained on the first impressions data set we released. Therefore, even when the annotation procedure aimed to be objective, biases are difficult to avoid. Explainability could be an effective way to overcome data biases, or at least to point out these potential biases so that decision takers can take them into account. Also, please note that explainable mechanisms could use data-bias information to provide explanations on their recommendations.

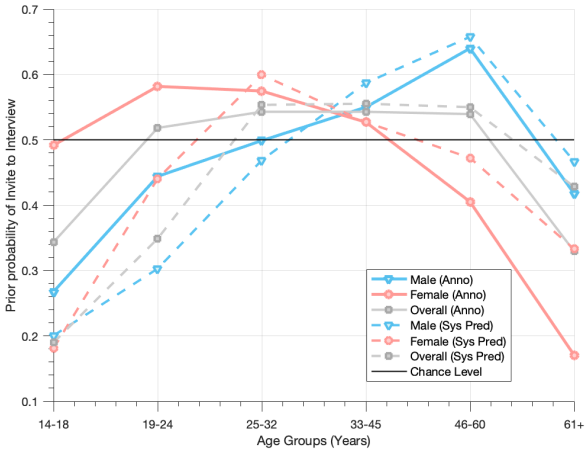


Fig. 11. The prior probability of job interview invitation over age groups, and over gender and age groups jointly.

5.5 Handling the Problem as a Classification Task

With the goal of bringing more light on the difficulty of the problem, in this section we report preliminary results when approaching the problem as one of classification instead of regression. For this experiment we considered the BU-NKU system described in Section 4. 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) [54], Histogram of Oriented Gradients (HOG) [11] and Scale Invariant Feature Transform [46]. The test set classification performances of the top systems for single- and multi-modal approaches are shown in Table 9.

First of all, it can be noticed that the performance of the classification model lies between 69% and 77% of test set accuracy. As the values of traits follow a nearly symmetric distribution, random guessing would approximate a 50% of accuracy. This illustrates the difficulty of the classification task and suggests that the apparently low gap to reach perfect performance in the regression problem is far from being reached.

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 highly 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.

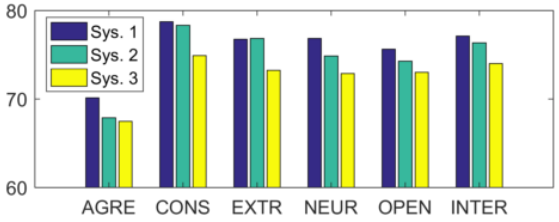


Fig. 12. Test set classification performance of top three fusion systems over personality traits and the interview variable. Sys. 1: Audio-Video system, Sys. 2: Video only system, Sys. 3: Audio plus a single image based system. NEUR refers to non-Neuroticism as it is used throughout the paper.

The dimension that is the hardest to classify is agreeableness, whereas accuracy for conscientiousness was consistently the highest (see Figure 12). Among the conventional image descriptors, HOG was the most successful, 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 outperforms acoustic features on both the development and test sets by 3%.

TABLE 9
Test set classification accuracies for the top single and multimodal systems. The scene feature is extracted from the first video frame.

Sys.	Modality	Interview	Trait Avg.
1	Audio + Video	77.10	75.63
2	Video (Face Seq.)	76.35	74.45
3	Audio + Scene + First Face	74.00	72.31
4	Audio + Scene	71.95	70.47
5	First Face + Scene	71.15	69.97
6	Audio Only	69.25	67.93

6 Lessons Learned and Open Issues

Explainable decision making is particularly important for algorithmic discrimination, where people are affected by the decisions given by an algorithm [72]. Such algorithms may be used for prioritization (e.g. multimedia search engines), classification (e.g. credit scoring), association (e.g. predictive policing), or filtering (e.g. recommender systems). We have described the first comprehensive challenge on apparent personality estimation, proposed two end-to-end solutions, and investigated issues of algorithmic accountability.

The first thing we would like to stress is that explainability requires user studies for its evaluation. Testbeds and protocols, such as the one we contribute in this paper, will be useful for advancing research in explainability. Additionally, it is essential that algorithmic accountability is broken down into multiple dimensions, along which systems are evaluated. In [19], these dimensions are proposed as responsibility, explainability, accuracy, auditability, and fairness, respectively. We have proposed five dimensions for explainability in this work.

Several applications and domains within computer vision will benefit from having explainable and interpretable models. Such mechanisms are essential in scenarios in which the outcome of the model can have serious implications. We foresee that the following application domains will significantly benefit from research in this direction: health applications (e.g., model-assisted diagnosis, remote patient monitoring, etc.); *non-visually-obvious* human behavior analysis (e.g., personality analysis, job screening); recognition tasks involving people (e.g., gender, ethnicity, age recognition); cultural-dependent tasks (e.g., adult content classification, cultural event recognition); security applications (e.g., biometrics of potential offenders, detection/verification/scanning systems, smart surveillance, etc.). The availability of new explicable/interpretable models will increase the scope of research for computer vision problems, and allow the creation of human-computer mixed decision systems in more sensitive application areas.

There is a marked distinction between visual question answering (VQA) and explainable decision making. VQA produces a narrative of the multimedia input, whereas explainability requires making a narrative of the decision process itself. This can be seen as a meta-cognitive property of the system. At the moment, the focus is on natural language based explanations, as well as strong visualizations suitable for human interpretation. Once such systems are sufficiently advanced, we could expect machine-interpretable explanations (such as through micro ontologies) to be produced as by-products, and compartmentalized systems taking advantage of such explanations to improve their decision making.

Black box models, such as deep neural network approaches, require external mechanisms (such as systematic examination of internal responses for ranges of input conditions) for the interpretation of their workings, which increase the annotation and training burden of these systems. On the other hand, transparent (white) models trade off accuracy. Balanced systems, such as the solutions we proposed in this work, combining early black box modeling with transparent decision-level modeling, could be the ideal solution.

Considering the current scenario of first-impression analysis in the context of job candidate screening, the quest for algorithmic accountability will inevitably also bring ethical questions. Will automatically trained pipelines inadvertently pick up on data properties that actually are irrelevant to the problem at hand [64]? May certain individuals get disadvantaged because of this, or due to inherent data biases? Generally, can judgments of external assessors be trusted? These are challenging questions, that will need comprehensive, interdisciplinary effort in order to be answered. In [45], an extensive discussion of this topic is given, considering the algorithmic job candidate screening problem from a machine learning and organizational psychology perspective.

Through the explicit focus of the ChaLearn challenge on *apparent* personality analysis, the question on whether prediction labels reflect 'true' personality is not the central question. Instead, the focus lies on inferring potential patterns in first-impression judgments made by outsiders. While these may be biased, the goal at this stage is not to mitigate this bias, but rather to gain deeper insight into what data characteristics may be underlying observed biases.

Contrasting typical approaches in organizational psychology with the approach taken in the ChaLearn challenge, traditional psychometrically validated personality measurement instruments would involve many more item questions than the ones that were currently offered to MTurk workers.

In addition, vlog data in the challenge may not originally have been intended as a video resume for a professional job candidacy. At the same time, the current data acquisition setup allowed for considerable scaling in terms of the amount of videos that could be annotated and analyzed; while several dozens of video resumes may already be a considerably sized corpus in the psychology domain, in the current challenge, several thousands of clips were studied.

These very different approaches therefore can offer different viewpoints and insights on the problem. A next step for future work will be to more explicitly compare them. Generally, in case inherent data and judgment bias would lead to unfair or unethical system predictions, the responsibility for avoiding this is shared. Machine learning researchers and engineers should be aware of this, and can conceive algorithmic solutions that explicitly mitigate unethical outcomes. At the same time, the identification of the most critical ethical challenges also needs explicit input and insight from data providers and domain specialists.

REFERENCES

[1] Timur R Almaev and Michel 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.

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU CS 16 118, CMU School of Computer Science, 2016.

[3] Jonathan Anderson. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26(6):490–496, 1983.

[4] Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube. In *5th Int. AAAI Conference on Weblogs and Social Media*, pages 446–449, 2011.

[5] Peter Borkenau, Steffi Brecke, Christine Möttig, and Marko Paelecke. Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, 43(4):703–706, 2009.

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

[7] Baiyu Chen, Sergio Escalera, Isabelle Guyon, Víctor Ponce-López, Nihar Shah, and Marc Oliu Simón. *Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings: Application to Personality Traits*, pages 419–432. Springer International Publishing, 2016.

[8] Meri Coleman and T. L. Liau. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975.

[9] Mark Cook. *Personnel Selection: Adding Value Through People*. Wiley-Blackwell, fifth edition, 2009.

[10] A. Vinciarelli D. Gatica-Perez and J. M. Odobez. *Interactive Multimodal Information Management*, chapter Nonverbal Behavior Analysis. EPFL Press, 2013.

[11] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893. IEEE, 2005.

[12] DARPA. Broad agency announcement explainable artificial intelligence (XAI). In *DARPA-BAA-16-53, August 10, 2016*, 2016.

[13] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[14] Salah Eddine Bekhouche, Fadi Dornaika, Abdelkrim Ouafi, and Abdelmalik Taleb-Ahmed. Personality traits and job candidate screening via analyzing facial videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[15] H. J. Escalante, V. Ponce, J. Wan., M. Riegler, Chen. B., A. Clapes, S. Escalera, I. Guyon, X. Baro, P. Halvorsen, H. Müller, and M. Larson. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *Proc. ICPRW*, 2016.

[16] Hugo Jair Escalante, Sergio Escalera, Isabelle Guyon, Xavier Barí, Yagmur Guçluturk, Umut Guclu, and Marcel A. J. van Gerven, editors. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018.

[17] Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Cezar Silveira Jacques, Meysam Madadi, Xavier Baró, Stéphane Ayache, Evelyne Viegas, Yagmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven, and Rob van Lier. Design of an explainable machine learning challenge for video interviews. In *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, pages 3688–3695, 2017.

[18] Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. ChaLearn Looking at People: A review of events and resources. In *Proc. IJCNN*, 2017.

[19] Diakopoulos N. et al. Principles for accountable algorithms and a social impact statement for algorithms. *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017.

[20] Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: the Munich Versatile and Fast open-source Audio Feature Extractor. In *Proc. of the Intl. Conf. on Multimedia*, pages 1459–1462. ACM, 2010.

[21] Ailbhe N. Finnerty, Skanda Muralidhar, Laurent Son Nguyen, Fabio Pianesi, and Daniel Gatica-Perez. Stressful first impressions in job interviews. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI, pages 325–332, New York, NY, USA, 2016. ACM.

[22] Rudolf Flesch. A New Readability Yardstick. *The Journal of Applied Psychology*, 32(3):221–233, 1948.

[23] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun 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.

[24] Umut Güçlü, Jordy Thielen, Michael Hanke, and Marcel A. J. van Gerven. Brains on beats. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2101–2109. Curran Associates, Inc., 2016.

[25] Yagmur Gucluturk, Umut Guclu, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel A. J. van Gerven, and Rob van Lier. Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, 2017.

[26] Yagmur Gucluturk, Umut Guclu, Marc Perez, Hugo Jair Escalante, Xavier Baro, Isabelle Guyon, Carlos Andujar, Julio Jacques Junior, Meysam Madadi, Sergio Escalera, Marcel A. J. van Gerven, and Rob van Lier. Visualizing apparent personality analysis with deep residual networks. In *The IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[27] R Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[28] Furkan Gürpınar, Heysem Kaya, and Albert Ali 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.

[29] Furkan Gürpınar, Heysem Kaya, and Albert Ali Salah. Multimodal Fusion of Audio, Scene, and Face Features for First Impression Estimation. In *23rd International Conference on Pattern Recognition*, pages 43–48, Cancun, Mexico, December 2016.

[30] Y. Güçlütürk, U. Güçlü, M. Pérez, H. Escalante, X. Baró, I. Guyon, C. Andujar, J. Jacques Junior, M. Madadi, S. Escalera, M. van Gerven, and R. van Lier. Visualizing apparent personality analysis with deep residual networks. In *ICCV Workshops*, 2017.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[32] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 3–19. Springer International Publishing, 2016.

[33] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong 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.

[34] Julio C. S. Jacques Junior, Yagmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andújar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. van Gerven, Rob van Lier, and Sergio Escalera.

First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 2019.

[35] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Multimodal Score Fusion and Decision Trees for Explainable Automatic Job Candidate Screening from Video CVs. In *CVPR Workshops*, pages 1651–1659, Honolulu, Hawaii, USA, December 2017.

[36] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[37] B. Kim, D. M. Malioutov, and K. R. Varshney. Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). *ArXiv e-prints*, July 2016.

[38] B. Kim, D. M. Malioutov, K. R. Varshney, and A. Weller. Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017). *ArXiv e-prints*, August 2017.

[39] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 2942–2950, 2017.

[40] J P Kincaid, R P Fishburne, R L Rogers, and B S Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training*, Research B(February):49, 1975.

[41] P.-J. Kindermans, K. Schütt, K.-R. Müller, and S. Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. *ArXiv e-prints*, November 2016.

[42] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[43] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015.

[44] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[45] Cynthia C. S. Liem, Markus Langer, Andrew Demetriou, Annemarie M. F. Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph. Born, and Cornelius J. König. *Explainable and Interpretable Models in Computer Vision and Machine Learning*, chapter Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. Springer, 2018.

[46] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[47] Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vision*, 120(3):233–255, December 2016.

[48] Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992.

[49] G.H. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[50] Klaus-Robert Müller, Andrea Vedaldi, Lars Kai Hansen, Wojciech Samek, and Grégoire Montavon. Interpreting, explaining and visualizing deep learning workshop. *NIPS*, Forthcoming, 2017.

[51] I. Naim, M. I. Tanveer, D. Gildea, and E. Hoque. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, PP(99), 2016.

[52] Laura P Naumann, Simine Vazire, Peter J Rentfrow, and Samuel D Gosling. Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12):1661–1671, 2009.

[53] Laurent Son Nguyen and Daniel Gatica-Perez. Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*, 18(7):1422–1437, 2016.

[54] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

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

[56] Victor Ponce Lopez, Baiyu Chen, Albert Places, Marc Oliu, Ciprian Corneanu, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, and Sergio 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.

[57] R. D. Pérez Principi, C. Palmero, J. C. Junior, and S. Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, pages 1–1, 2019.

[58] F L Schmidt and J E Hunter. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, 124(2):262–274, 1998.

[59] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim. The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism. pages 148–152, 2013.

[60] R. R Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. Grad-CAM: Why did you say that? *ArXiv e-prints*, November 2016.

[61] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016.

[62] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[63] E A Smith and R J Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, pages 1–14, 1967.

[64] Bob L. Sturm. A simple method to determine if a music information retrieval system is a 'horse'. *IEEE Transactions on Multimedia*, 2014.

[65] Achmadnoer Sukma Wicaksana and Cynthia C. S. Liem. Human-explainable features for job candidate screening prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[66] Michel F. Valstar, Björn W. Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge, AVEC@ACM Multimedia 2013, Barcelona, Spain, October 21, 2013*, pages 3–10, 2013.

[67] C. Ventura, D. Masip, and A. Lapedriza. Interpreting cnn models for apparent personality trait regression. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1705–1713, 2017.

[68] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291, 2014.

[69] Feng Wang, Haijun Liu, and Jian Cheng. Visualizing deep neural network by alternately image blurring and deblurring. *Neural Networks*, 2017.

[70] A. G. Wilson, B. Kim, and W. Herlands. Proceedings of NIPS 2016 Workshop on Interpretable Machine Learning for Complex Systems. *ArXiv e-prints*, November 2016.

[71] Andrew Gordon Wilson, Jason Yosinski, Patrice Simard, and Rich Caruana. Interpretable ml symposium. *NIPS*, Forthcoming, 2017.

[72] www foundation. Algorithmic accountability. *World Wide Web Foundation, Defense Advanced Research Projects Agency (DARPA)*, 2017.

[73] Matthew D. Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*, pages 818–833. Springer International Publishing, Cham, 2014.

[74] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*, 2015.

[75] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *CoRR*, abs/1702.04595, 2017.

Hugo Jair Escalante is researcher scientist at Instituto Nacional de Astrofisica, Optica y Electronica, INAOE, Mexico. and a director of ChaLearn, a nonprofit dedicated to organizing challenges, since 2011. He has been involved in the organization of several challenges in computer vision and automatic machine learning. He is reviewer at JMLR, PAMI, and has served as coeditor of special issues in IJCV, PAMI, and TAC.

Heysem Kaya is an assistant professor at Utrecht University, Department of Information and Computing Sciences. He was awarded PhD degree from Computer Engineering Department, Boğaziçi University in 2015. He has co-authored over 40 publications on machine learning, computational paralinguistics and affective computing. His works won four Computational Paralinguistics Challenge (ComParE) awards and four awards in multimodal affective computing challenges, including three ChaLearn competitions.

Albert Ali Salah (M08, SM17) is professor and chair of Social and Affective Computing at the Information and Computing Sciences Department of Utrecht University. Albert has obtained his PhD from Boğaziçi University, Turkey. He has co-authored over 150 publications on pattern recognition, multimodal interfaces, and computer analysis of human behavior. He serves as a Steering Board member of ACM ICMI, as an associate editor of journals including IEEE Trans. on Cognitive and Developmental Systems, IEEE Trans. Affective Computing, and Int. Journal on Human-Computer Studies.

Sergio Escalera leads the Human Analysis Group (HuPBA) at UB and CVC. He is an associate professor at Universitat de Barcelona. He is member of the Computer Vision Center. He is vice-president of ChaLearn and chair of IAPR TC-12: Multimedia and visual information systems. He has been awarded with ICREA Academia. His research interests include affective and personality computing, with special interest in human pose recovery and behavior analysis from multimodal data.

Yağmur Güçlütürk is an assistant professor at the Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands, where she is working on neuroprosthetics and AI as a member of the Artificial Cognitive Systems Lab and the Neuronal Stimulation for Recovery of Function Consortium. Previously, she did her Ph.D. and M.Sc in Cognitive Neuroscience at the Radboud University, and B.IT in Artificial Intelligence at the Multimedia University.

Umut Güçlü is an assistant professor at Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands, and a member of the Artificial Cognitive Systems lab, where he works on combining deep learning and neural coding to systematically investigate the cognitive algorithms in vivo with neuroimaging and implement them in silico with artificial neural networks. He did his postdoctoral, doctoral and master's studies in cognitive neuroscience and bachelor's studies in artificial intelligence, as well as additional undergraduate studies in natural sciences and mathematics.

Xavier Baró received his B.S. degree in Computer Science from UAB in 2003. In 2005 he obtained his M.S. degree in Computer Science at UAB, and in 2009 the Ph.D degree in Computer Engineering. Currently, he is associate professor and researcher at the Computer Science, Multimedia and Telecommunications department at Universitat Oberta de Catalunya (UOC). From 2015, he is the director of the Computer Vision Master degree at UOC. He is co-founder of the SUNAI group, and collaborates within CVC at UAB as member of HUPBA group. His research interests are related to machine learning, evolutionary computation, and statistical pattern recognition, specially their applications to the field of Looking at People.

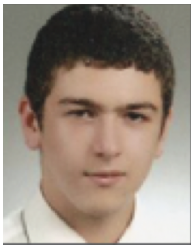
Isabelle Guyon is chaired professor in "big data" at the Université Paris-Saclay, specialized in statistical data analysis, pattern recognition and machine learning. Her areas of expertise include computer vision and bioinformatics. Her recent interest is in applications of machine learning to the discovery of causal relationships. Prior to joining Paris-Saclay she worked as an independent consultant and was a researcher at AT&T Bell Laboratories, where she pioneered applications of neural networks to pen computer interfaces (with collaborators including Yann LeCun and Yoshua Bengio) and co-invented with Bernhard Boser and Vladimir Vapnik Support Vector Machines (SVM), which became a textbook machine learning method. She organizes challenges in Machine Learning since 2003 supported by the EU network Pascal2, NSF, and DARPA, with prizes sponsored by Microsoft, Google, Facebook, Amazon, Disney Research, and Texas Instrument. She is action editor of the Journal of Machine Learning Research, and program chair of the NIPS 2016 conference.

Julio C. S. Jacques Junior is a postdoctoral researcher at the Computer Science, Multimedia and Telecommunications department at Universitat Oberta de Catalunya (UOC), within the Scene Understanding and Artificial Intelligence (SUNAI) group. He also collaborates within the Computer Vision Center (CVC) and Human Pose Recovery and Behavior Analysis (HUPBA) group at Universitat Autonoma de Barcelona (UAB) and University of Barcelona (UB), as well as within ChaLearn Looking at People.

Meysam Madadi obtained his MS degree and PhD in Computer Vision at the Universitat Autònoma de Barcelona (UAB) in 2013 and 2017, respectively. He is currently a post-doc researcher at Computer Vision Center (CVC), UAB. He has been a member of Human Pose Recovery and Behavior Analysis (HUPBA) group since 2012. He collaborated in ChalearnLAP ECCV2014, NIPS2016, CVPR2017, ICCV17, ECCV18 and CVPR19.

Stephane Ayache received his Ph.D. in Computer Science from Institut National Polytechnic de Grenoble (INPG), France in 2007, he is currently associate Professor at AMU/LIF since 2008. His research interests include Computer Vision, Image/Video content understanding and indexing, Machine Learning with a focus on multimodal analysis and multiview learning, transfer learning, deep learning.

Evelyne Viegas is Senior Director of Research Engagement at Microsoft Research. In her current role, she leads initiatives in the areas of Artificial Intelligence, Computing Systems, Experiences and AI and Society, working in partnership with the academic research community, and business groups, worldwide. She develops signature programs via research collaborations to drive open innovation.



Furkan Gürpınar has obtained his MS degree from Boğaziçi University, Computational Science and Engineering program, with a thesis on Video and Image Based Face Analysis with Extreme Learning Machines. His work received first place in ChaLearn LAP Job Candidate Screening Challenge (CVPR'17), ChaLearn LAP First Impressions Challenge (ICPR'16), and was first runner up in EMOTIW AFEW Challenge (ICMI'15).

Achmadnoer Sukma Wicaksana was a PhD student at Technical University of Delf.

Cynthia Liem is assistant Professor at the Multimedia Computing Group. In 2007 and 2009, obtained her BSc and MSc degree in Media and Knowledge Engineering (Computer Science) at the TU Delft, after which she continued pursuing a PhD at the same institution (defended in 2015). Besides, she obtained the BMus (2009) and MMus (2011) degree in classical piano performance at the Royal Conservatoire in The Hague.

Marcel A. J. van Gerven is professor of Artificial Cognitive Systems at Radboud University and principal investigator in the Donders Institute for Brain, Cognition and Behaviour. His research focuses on brain-inspired computing, understanding the neural mechanisms underlying natural intelligence and the development of new ways to improve the interaction between humans and machines. He is head of the Artificial Intelligence department at Radboud University.

Rob van Lier studied Experimental Psychology, and after that he took his PhD on the topic of human visual perception at Radboud University (Nijmegen, the Netherlands). Next, he spent a few postdoc years at the University of Leuven (Belgium) and returned to Nijmegen, based on a two-year grant from the Dutch National Science Foundation (NWO) and a 5-year grant from the Royal Netherlands Academy of Arts and Sciences (KNAW). Currently, he is an associate professor at the Radboud University and a principal investigator at the Donders Institute for Brain, Cognition and Behaviour (Nijmegen, The Netherlands), heading the research group 'Perception and Awareness'.