Preprint submitted to International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Explaining First Impressions: 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 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**

Received: date / Accepted: date

* Means equal contribution by the authors.

Hugo Jair Escalante
INAOE, Mexico and ChaLearn, USA E-mail: hugojair@inaoep.mx

Heysem Kaya
Namık Kemal University, Department of Computer Engineering, Turkey
E-mail: hkaya@nku.edu.tr

Albert Ali Salah
Boğaziçi University, Dept. of Computer Engineering, Turkey and Nagoya University, FCVRC, Japan
E-mail: salah@boun.edu.tr

Furkan Gürpınar
Boğaziçi University, Computational Science and Engineering, Turkey
E-mail: furkan.gurpinar@boun.edu.tr

Sergio Escalera
University of Barcelona and Computer Vision Center, Spain
E-mail: sergio@maia.ub.es

Meysam Madadi
Computer Vision Center, Spain
E-mail: mmadadi@cvc.uab.es

Yağmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven and Rob van Lier
Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, the Netherlands E-mail: {y.gucluturk,u.guclu,m.vangerven,r.vanlier}@donders.ru.nl

Xavier Baró and Julio Jacques Junior
Universitat Oberta de Catalunya and Computer Vision Center, Spain
E-mail: {xbaro,jsilveira}@uoc.edu

Isabelle Guyon
UPSud/INRIA, Université Paris-Saclay, France and ChaLearn, USA

Abstract Explainability and interpretability are two critical aspects of decision support systems. Within computer vision, they are critical in certain tasks related to human behavior analysis such as in health care applications. 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 computer vision with an emphasis on looking at people tasks. Specifically, we review and study those mechanisms in the context of first impressions analysis. To the best of our knowledge, this is the first effort in this direction. Additionally, 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, and 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.

Keywords Explainable computer vision · First impressions · Personality analysis · Multimodal information · Algorithmic accountability

1 Introduction

Looking at People (LaP) – the field of research focused on the visual analysis of human behavior – has been a very active research field within computer vision in the last decade [23,24,53]. Initially, LaP focused on tasks associated with basic human behaviors that were *obviously* visual (e.g., basic gesture recognition [61,60] or face recognition in restricted scenarios [6,72]). Research progress in LaP has now led to models that can solve those initial tasks relatively easily [57,71]. Instead, attention on human behavior analysis has now turned to problems that are not *visually evident* to model / recognize [73,42,62]. For instance, consider the task of assessing personality traits from visual information [62]. Although there are methods that can estimate *apparent* personality traits with (relatively) acceptable performance, model recommendations by themselves are useless if the end user is not confident on the model's *reasoning*, as the primary use for such estimation is to understand bias in human assessors.

Explainability and interpretability are thus critical features of decision support systems in some LaP tasks. The former focuses on mechanisms that can tell what is the rationale behind the decision or recommendation made by the

E-mail: guyon@chalearn.org

Stephane Ayache
Aix Marseille Univ, CNRS, LIF, Marseille, France E-mail: Stephane.Ayache@lif.univ-mrs.fr

Evelyne Viegas
Microsoft Research, USA E-mail: evelynev@microsoft.com

Achmadnoer Sukma Wicaksana, Cynthia C. S. Liem
Multimedia Computing Group, Delft University of Technology, Delft, The Netherlands
E-mail: sukmawicaksana@gmail.com, c.c.s.liem@tudelft.nl

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. The model representation, the way its parameters are learned or the manner in which inference is performed gives these models a somewhat *white box* appearance. In fact, explainable and interpretable models have been available for a while for some applications within Artificial Intelligence (AI) and machine learning. However in computer vision, 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 computer vision tasks. 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[1]. This perception is starting to change, with more research focused on visualizing and understanding the structure of models and with cutting edge research focused on explaining the recommendations of LaP models.

This paper comprises a comprehensive study on explainability and interpretability in the context of computer vision, with emphasis on LaP tasks. In particular, we focus on those mechanisms in the context of first impressions analysis. We review concepts and the state of the art on the subject. Additionally, 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, and 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.

The main contributions of this paper are as follows:

- We provide a review of the state of the art on explainability and interpretability in the context of computer vision.
- We provide a comprehensive analysis of the job candidate screening competition, which is an academic challenge that aimed to push research on explainability for first impressions analysis. We present the challenge along with the participants' solutions described in detail.
- We analyze in depth the novel data set we have collected and annotated for the challenge, discussing critical aspects like annotation agreements and annotator biases.
- We outline the main findings of our study and identify promising venues for research on explainability in the context of LaP.

The remainder of this paper is organized as follows. Section 2 reviews related work on explainability and interpretability in the context of computer

[1] Please note that there have been efforts since a while trying to demystify the structures of deep networks, see e.g., [81].

vision. Section 3 describes the LAP First Impression Challenge, summarizing its results and main findings. Section 4 describes in more detail some methodologies proposed for explaining first impressions. Section 5 presents an in-depth study on the data set associated with the challenge. Finally, Section 6 presents a discussion on the lessons learned and outlines further research ideas in this area.

2 Explainability and interpretability in computer vision

Looking at People is the field of computer vision dealing with the visual analysis of all kinds of human behaviors from images and video [53]. Traditional LaP tasks include action and gesture recognition, facial image analysis, pose estimation, and human computer interaction, among others. Because of the large number of applications and of the difficulty in modeling human behavior in all of its varieties, LaP is a complex problem in computer vision. One should note that LaP inherits problems from other tasks of scene understanding, including illumination changes, partial occlusions, viewpoint changes, rigid and elastic deformations, and high inter- and intra-class variability, just to mention a few. In spite of these difficulties, LaP has made impressive progress in several classical tasks, see e.g., [24,53,67] and is starting to focus on much more complex problems, see e.g. [25,62,21,22].

2.1 Audio-visual analysis of First impressions

As previously mentioned, LaP has pushed the state of the art in classical problems that have strong visual aspects, such as face recognition, body pose estimation, and gesture recognition. However, there are several problems for which research is still in its infancy. In this paper, we tackle estimating the apparent personality of people and related variables, see e.g. [62]. Personality and conduct variables in general are rather difficult to infer precisely from visual inspection, this holds even for humans. Accordingly, the LaP field is starting to pay attention to a less complex problem, that of estimating *apparent* personality from visual data [62]. Related topics receiving increasing attention from the LaP community are first impressions analysis, depression recognition and hiring recommendation systems [62,73,21,14], all of them starting from visual information.

The impact that LaP methods can have is huge. According to Vinciarelli and Mohammadi [75], any technology involving understanding, prediction and synthesis of human behavior is likely to benefit from personality computing approaches. One of such application is *job candidate screening*. According to Nguyen and Gatica-Perez [58], 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.This is precisely the aim of the speed interviews project[2]. 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.

Efforts on automating the interview process by analyzing videos are scarce [3]. In [58] 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. [28] aimed to determine whether first impressions of stress (from an annotated database of job interviews) 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. Then, two regression techniques, ridge regression and random forest are evaluated. Stress impressions were found to be significantly negatively correlated with "hirability" ratings (i.e., individuals who were perceived to be more stressed were more likely to obtain lower "hirability" scores). Regression results show evidence that visual features are better predictors of stress impressions than audio features. In the same line, Naim et al. [55] 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.

2.2 Explainability and interpretability 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 (including related tasks such as job recommendation) [32,33,74]. Although very competitive results have been reported with such methods (see e.g., [22]), 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 [74]. Because of this, explainability and interpretability have received special attention

[2] `http://gesture.chalearn.org/speed-interviews`

[3] However, one should note that the analysis of non-verbal behavior to predict the outcome of a social interaction is a topic that has been studied for a while in different domains [14].

in different fields, see e.g., [15]. In fact, the interest from the community on this topic is evidenced by the organization of dedicated events, such as thematic workshops [43,44,54,77,78] and challenges [22]. This is particularly important to ensure fairness and to verify that the models are not plagued with various kinds of biases, 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 [46,82,50,84,76]. 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 [16,65,66,49]. Selvaraju et al. presented a technique for making convolutional neural network (CNN) based models more transparent [65,66]. A novel class-discriminative localization technique is proposed - Gradient-weighted Class Activation Mapping - and combined with existing high-resolution visualizations to produce visual explanations for CNN-based models. When image classification is approached, this form of visualization lend insights into failure modes of the model (showing that seemingly unreasonable predictions have reasonable explanations). Das et al. conducted large-scale studies on "human attention" in Visual Question Answering (VQA) in order to understand where humans look when answering questions about images [16]. Attention maps generated by deep VQA models were evaluated against human attention. Their experiments showed that current attention models in VQA do not seem to be looking at the same regions as humans. This may not be a handicap, as a computer system potentially has access to more detailed information (e.g. it can look at the acceleration profile of a mouth corner movement to assess whether a smile is genuine or fake, which is very difficult for a human to compute [17]). Koh and Liang used influence functions - a classic technique from robust statistics - to trace the prediction of a black-box model through the learning algorithm and back to its training data [49], thereby identifying training points most responsible for a given prediction. They demonstrated that influence functions are useful for multiple purposes: understanding model behavior, debugging models, detecting dataset errors, and even creating visually-indistinguishable training-set attacks.

2.3 Explainability and interpretability of 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[4], sought to address this question.

[4] `http://openaccess.thecvf.com/CVPR2017_workshops/CVPR2017_W26.py`

In the remainder of this section, we review these first efforts on explainability and interpretability for first impressions and "hirability" analyses.

Güçlütürk et al. [32] 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 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 [33], 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 [32], 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. [74] 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. [40] 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 [70] 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. These two

approaches are discussed in detail in Section 4, as their proposed methods obtained the best performance in the Job Candidate Screening Competition we organized [40, 70].

2.4 A word of caution

The previous sections have reviewed research progress on LaP focusing on the explainability and interpretability of models. Researchers in LaP have made a great progress in different areas of LaP, 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. 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. This section describes the challenge and in the next sections we elaborate on the associated data set and main findings. The design of the challenge is further detailed in [22].

3.1 Job candidate screening: perspectives from organizational psychology

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 first impression judgments on people in video blogs, and whether they will be considered to be invited to a job interview. Accordingly, in this section we briefly review relevant aspects of organizational psychology on job candidate screening.

Traditionally, job candidate screening and application sifting would be conducted based on information in CVs and application forms, supported by references. The sifting procedure can be further improved by assessing behavioral

competences, weighted application blanks and bio-data, training and experience ratings, minimum qualifications, background investigations or positive vetting, structured questioning, Internet tests, and application scanning software [13].

For personnel selection, various types of assessment may further be performed to assess the candidate's suitability. Seven main aspects are identified in [13]:

- Mental ability (GMA) (intelligence, problem solving, practical judgement, clerical ability, mechanical comprehension, sensory abilities)
- Personality traits
- Physical characteristics (strength, endurance, dexterity)
- Interests, values and fit
- Knowledge (declarative: facts, procedural: knowing how to act, tacit: 'knowing how things really happen') (note: mastery of higher-level knowledge may require higher levels of mental ability)
- Work skills (dedicated skills required for the job, e.g. bricklaying or diagnosing an illness)
- Social skills (e.g. communication, persuasion, negotiation, influence, leadership, teamwork).

All of these aspects were considered for the design of the challenge. However, in order to limit the scope of the challenge and to facilitate the objective evaluation of methods for automatically job candidate screening, some aspects were simplified. Obviously, different requirements are required from, say, an HR specialist and a GPU coder. We focused on the aspects that are independent from the job type to obtain general results.

3.2 Overview

The challenge relied on a novel data set that we made publicly available recently[5] [21,62]. 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 (HD) 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 shows snapshots of sample videos from the data set.

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. 3.3). 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 screening interview steps e.g. based

[5] Data set is available at `http://chalearnlap.cvc.uab.es/dataset/24/description/`

Fig. 1 Snapshots of sample videos from the First Impressions data set [62].

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: `http://gesture.chalearn.org/speed-interviews`, 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 [62] and ICPR2016 [21]. 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 (a task not previously explored). 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 (coopetition) by rewarding participants who share their code during the challenge, weighting rewards with the usefulness/popularity of their code.

3.3 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) [51], 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. Figure 2 illustrates the interface that Turkers had access to.

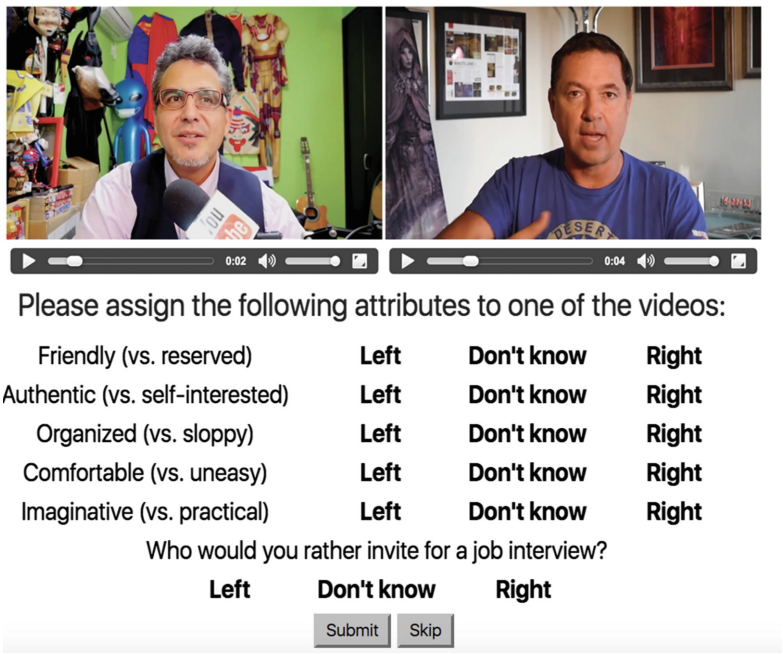


Fig. 2 Snapshots of the interface for labeling videos [62]. The "big five" traits are characterized by adjectives: Extroversion = Friendly (vs. Reserved); Agreeableness = Authentic (vs. Self-interested); Conscientiousness = Organized (vs. Sloppy); (non-)Neuroticism = Comfortable (vs. Uneasy); Openness = Imaginative (vs. Practical).

We post-processed the rankings provided by the annotators to generate scores for each annotated variable. By doing so, cardinal scores were obtained by pairwise fitting a Bradley-Terry-Luce (BTL) model. Additionally, we used

a small-world algorithm for sampling the pairs that Turkers had to annotate. Details on the labeling procedure can be found in [11].

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[6] 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 was greater than 0.65. Furthermore, the apparent trait annotations were highly predictive of invite-for-interview annotations, with a significantly above-chance coefficient of determination of 0.91.

3.4 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 at this stage.

Figure 3 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.

3.4.1 Platform

As in other challenges organized by ChaLearn[7], the job candidate screening coopetition ran in CodaLab[8]; a platform developed by Microsoft Research and Stanford University in close collaboration with the organizers of the challenge.

[6] `http://www.rev.com`
[7] `http://chalearn.org`
[8] `http://codalab.org/`

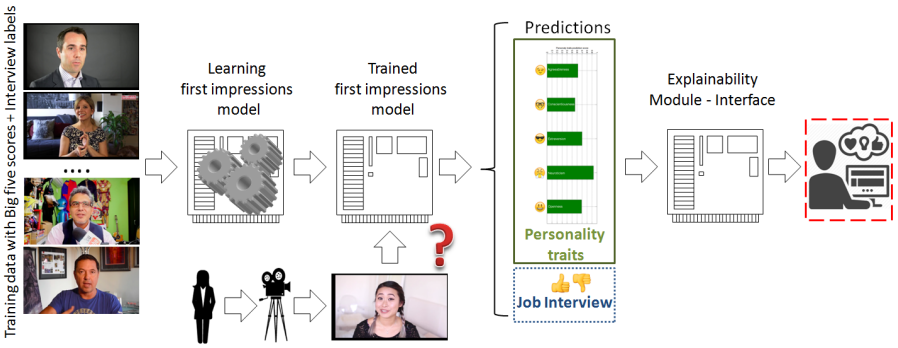


Fig. 3 Diagram of the considered scenario in the job candidate screening coopetition. The solid (green) top square indicates the variables evaluated in past editions of the challenge [21, 62]. The dotted (blue) bottom square indicates the variable evaluated in the quantitative track. The dashed (red) square indicates what is evaluated in the qualitative track.

3.4.2 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, together with fact sheets summarizing their solutions.

3.4.3 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:

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

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 well as an optional video that participants could submit to describe their method. Figure 4 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.

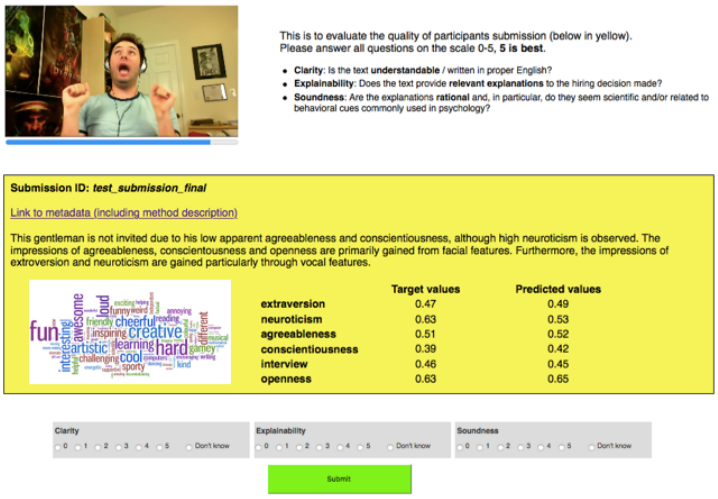


Fig. 4 Qualitative evaluation interface. The explainable interface of a submission is shown to the judge who had to evaluate it along the considered dimensions.

3.5 Baselines

We considered several baselines for solving the aforementioned tasks in different input modalities. Here, we briefly describe the baseline models and results (see [32,37] for more details).

3.5.1 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.

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.

Skip-thought vectors model. This model uses an embedding that represents transcripts as 4800-dimensional mean skip-thought vectors [48] of the sentences in the transcriptions. A recurrent encoder-decoder neural network pretrained on the BookCorpus dataset [83] was used for extracting the skip-thought vectors from the transcriptions.

3.5.2 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 [38]. 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 mean absolute error loss function iteratively with stochastic gradient descent (Adam [47]) on the training set, and model selection was performed on the validation set.

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 [31], 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.

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.

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.

3.5.3 Language and sensory model

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.

3.5.4 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

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*

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.

4 Two 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: Decision Trees for Modality Fusion and Explainable Machine Learning

The BU-NKU system is based on audio, video, and scene features. A similar pipeline was used in the system that won the ChaLearn First Impression Challenge at ICPR 2016 [36]. The main difference is that 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 [40]. The flow of this system is illustrated in Figure 5. The qualitative stage inputs the final predictions from the system proposed for the quantitative stage, discretizes them using the training set mean scores of each target dimension, then maps the binarized (low/high) personality traits 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, hence allow explainability for the decision. Finally the DT is traced to generate a verbal explanation. A wider but brief summary of the components used in this system is provided in the subsequent subsections.

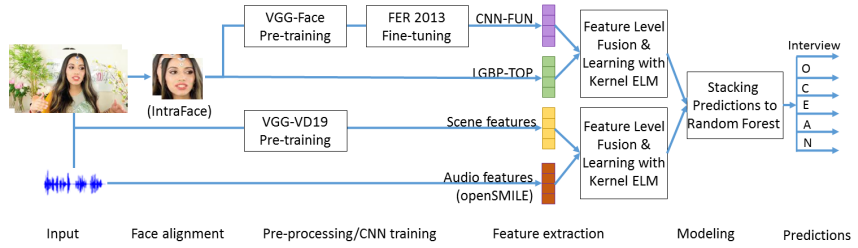


Fig. 5 Flowchart of the BU-NKU system.

4.1.1 Quantitative System

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.

Face Features Faces are detected on all frames of the video input. The Supervised Descent Method (SDM) is used for face registration, which gives 49 landmarks on each detected face [80]. The roll angle is estimated from the eye corners to rotate the image accordingly. Then a margin of 20% of the interocular distance around the outer landmarks is added to crop the facial image. Each image is resized to 64×64 pixels.

After aligning the faces, image-level deep features are extracted from a convolutional neural network trained for facial emotion recognition. To prepare this feature extractor, the system starts with the pre-trained VGG-Face network [59], which is optimized for the face recognition task on a very large set of faces. Then this network is fine-tuned for emotion [41], using more than 30K training images of the FER-2013 dataset [30]. The final trained network has a 37-layer architecture (involving 16 convolution layers and 5 pooling layers). The response of the 33^{rd} layer is used, which is the lowest-level 4 096-dimensional descriptor.

After extracting frame-level features from each aligned face, videos are summarized by computing functional statistics of each dimension over time. The functionals include mean, standard deviation, offset, slope, and curvature. Offset and slope are calculated from the first order polynomial fit to each feature contour, while curvature is the leading coefficient of the second order polynomial.

In the BU-NKU approach, deep 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, 41]. It is extracted by applying 18 Gabor filters on aligned facial images with varying orientation and scale parameters. The resulting feature dimensionality is 50 112.

Scene Features In order to use ambient information in the images, a set of features is extracted using the VGG-VD-19 network [68], 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 [35,36]. For Job Candidate Screening task, these features contribute to the final decision both directly and indirectly over the personality trait predictions.

Acoustic Features The open-source openSMILE tool [27] is popularly used to extract acoustic features in a number of international paralinguistic and multi-modal challenges. The idea is to obtain a large pool of potentially relevant features by passing an extensive set of summarizing functionals on the low level descriptor contours (e. g. Mel Frequency Cepstral Coefficients, pitch, energy and their first/second order temporal derivatives). 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 [64]. This configuration was found to be the most effective acoustic feature set among others for personality trait recognition [36].

Model Learning In order to model personality traits from audio-visual features, kernel extreme learning machines (ELM) were used, due to the learning speed and accuracy of the algorithm. Initially, ELM is proposed as a fast learning method for Single Hidden Layer Feedforward Networks (SLFN): an alternative to back-propagation [39]. To increase the robustness and the generalization capability of ELM, a regularization coefficient is included in the optimization procedure.

Score Fusion 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 [9]. Sampling with replacement leaves approximately one third of the training set instances *out-of-bag*, which are used to cross-validate the models and optimize the hyper-parameters at the training stage. This is an important aspect of the method regarding the challenge conditions, as cross validation gives an unbiased estimate of the expected value of prediction error [10].

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 [36], 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.

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**

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.

4.1.2 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 (DT) classifier. The use of a DT is motivated by the fact that the resulting model is self-explanatory and can be converted into an explicit recommender algorithm using "if-then" rules. The proposed approach for decision explanation uses the trace of each decision from the root of the tree to the leaf. The verbal explanations are finally accompanied with the aligned image from the first face-detected frame and the bar graphs of corresponding mean normalized scores.

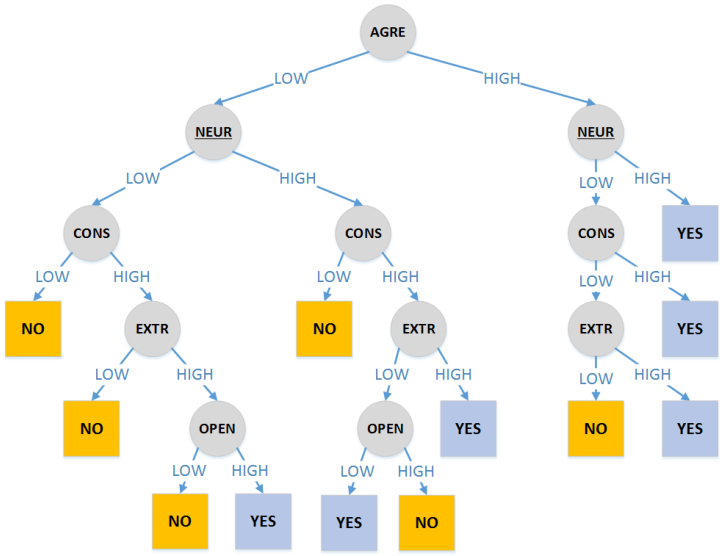
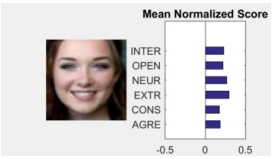


Fig. 6 Illustration of the decision tree for job interview invitation. NEUR denotes (non-) Neuroticism. Leaves denote a positive or a negative invitation response.

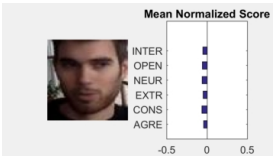
The DT trained on the predicted OCEAN dimensions gives a classification accuracy of 94.2% for binarized interview variable. The illustration of the trained DT is given in Figure 6. The model is intuitive as higher scores of traits generally increase the chance of interview invitation. As can be seen 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. The high/low scores of these top four traits are correlated with target variable and are observed to be consistent throughout the DT. If the Openness score is high, then having a high score in any of the variables (non-)Neuroticism, Conscientiousness or Extroversion suffices for invitation. Chances of invitation decrease if Agreeableness is low: only three out of eight leaf nodes are "YES" in this branch. In two of these cases, one has to have high scores in three out of four remaining traits. Figure 7 illustrates automatically generated verbal and visual explanations for this stage.

4.2 TUD: Layered Linear Regression and Weight Analysis

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



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.



This gentleman is not invited due to his low apparent agreeableness, non-neuroticism, conscientiousness, extroversion and openness scores. The impression of conscientousness is modulated by facial expressions. Furthermore, the impressions of agreeableness, extroversion, non-neuroticism and openness are gained particularly through vocal features.

Fig. 7 Sample verbal and visual explanations from qualitative stage for the BU-NKU entry.

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 [56,8]. However, considering findings in organizational psychology, personality traits are not the only (and neither the strongest) predictors for job suitability. In fact, GMA (General Mental Ability) tests, such as intelligence tests, have the highest validity at the lowest application cost [63,13]. While there is no formal GMA assessments for subjects in the dataset, language use of the vlogger may indirectly reveal GMA characteristics, such as the use of difficult words. Consequently, textual features, including speaking density, as well as linguistic sophistication, were considered for this approach.

4.2.1 Visual Features

For the visual representation, the system was not built to focus on the video in general, but particularly on facial expression and movement. OpenFace tools were used to segment only the face from each video [2], standardizing the segmented facial video to 112x112 pixels. OpenFace is an open source toolkit which does not only segment faces, but offers a feature extraction library that can extract and characterize facial movements and gaze [4]. OpenFace is able to recognize a subset of individual Action Units (AU) that construct facial expressions encoded in Facial Action Code System (FACS) as shown in

Table 3 [19,20]. These AUs then can be described in two ways: in terms of presence (indicating whether a certain AU is detected in a given time frame) and intensity (indicating how intense an AU is at a given time frame).

Table 3 Action Units that are recognized by OpenFace.

Action Unit	Description	Action Unit	Description
AU1	Inner Brow Raiser	AU14	Dimpler
AU2	Outer Brow Raiser	AU15	Lip Corner Depressor
AU4	Brow Lowerer	AU17	Chin Raiser
AU5	Upper Lid Raiser	AU20	Lip stretcher
AU6	Cheek Raiser	AU23	Lip Tightener
AU7	Lid Tightener	AU25	Lips part
AU9	Nose Wrinkler	AU26	Jaw Drop
AU10	Upper Lip Raiser	AU28	Lip Suck
AU12	Lip Corner Puller	AU45	Blink

For each of these AUs, three features were constructed for input to the system. First, the percentage of time frames is computed, during which the AU was visible in a video. Second, the maximum intensity of the AU in the video was stored. Lastly, the mean intensity of the AU over the video was also recorded. These three features per AU add up to 52 features in total for the OpenFace representation.

The resulting segmented video is also used for another video representation. In order to capture overall movement of the vlogger's face, a Weighted Motion Energy Image (wMEI) is constructed from the resulting face segmented video. MEI is a grayscale image that shows how much movement happens on each pixel throughout video, with white indicating a lot of movement and black indicating less movement [7]. wMEI was proposed in the work of Biel et al. [5] as a normalized version of MEI, by dividing each pixel values with the maximum pixel value. The method proposed by TUD is inspired by the aforementioned work with improvement on background noise reduction. In [5], the whole video frame is used as an input to compute wMEI, which makes background movement contribute to the overall wMEI measurements. Thus, there are cases in which the resulting wMEI is all white due to background or camera movements, rather than movement of a human subject. For example, this happens when the vlogger recorded the video in a public space or while on the road. Using the face segmented video instead of a whole video frame, the involvement of background is minimized to get a better representation of the subject's true movement, as can be seen in Figure 8. In order to create wMEI, 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.

The ChaLearn dataset has been carefully prepared so that only one unique foreground person faces the camera in the video. However, the current OpenFace implementation has limitations when the video still contains other visual

Fig. 8 wMEI for face segmented video

sources with faces, such as posters or music covers in the background. While the situation is rare, it was occasionally noticed that a poster was detected and segmented as 'main face' rather than the subject's actual face. For such misdetections, no movement will be detected at all, so these outliers are easily captured by the system.

4.2.2 Textual Features

Textual features are generated by using transcripts that were provided as the extension of the ChaLearn dataset. For a handful of videos, transcript data was missing; those videos were manually annotated, such that all videos have a transcript, with exception of one video in which the person speaks in sign language.

As reported in the literature [63,13] 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-contrib package of the Natural Language Toolkit (NLTK). More specifically, eight measures were selected as features for the Readability representation: ARI [69], Flesch Reading Ease [29], Flesch-Kincaid Grade Level [45], Gunning Fog Index [34], SMOG Index [52], Coleman Liau Index [12], 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.

4.2.3 Quantitative System

The building blocks of the TUD predictive model encompass four feature representations; OpenFace, MEI, Readability, and Text. Employing the 6000 training set videos, for each representation, a separate model 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 9.

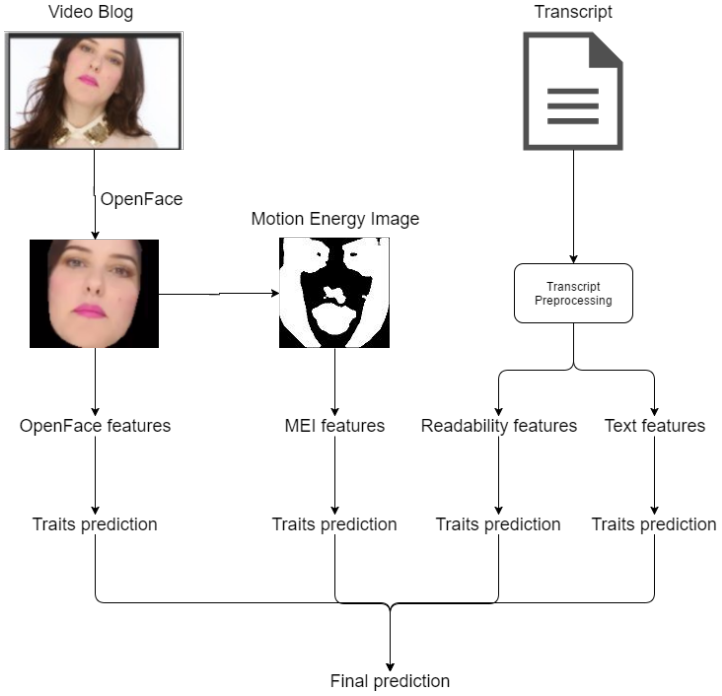


Fig. 9 Overall system diagram for the TUD system.

As the goal of the system is to trace back the prediction scores to each underlying feature, linear models were selected. Linear regression is a commonly used model in the social sciences literature. 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.

The linear regression coefficients indicate for each PCA dimension whether it contributes positively or negatively to the prediction. Furthermore, considering the PCA transformation matrix, the strength of teh contribution for original features can be determined.

Table 4 shows the overall quantitative accuracy of the TUD system on the 2000 videos in the benchmark training set, 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 (from all of the participants) in the ChaLearn CVPR 2017 Quantitative 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 explain-

Table 4 Accuracy (1 - Relative Mean Absolute Error) comparison between the TUD system and the lowest and highest accuracy for each prediction category in the ChaLearn CVPR 2017 Quantitative Challenge.

Categories	TUD System	Lowest	Highest
Interview	0.887744	0.872129	0.920916
Agreeableness	0.896825	0.891004	0.913731
Conscientiousness	0.880077	0.865975	0.919769
Extroversion	0.887040	0.878842	0.921289
Neuroticism	0.884847	0.863237	0.914613
Openness	0.890314	0.874761	0.917014

ability. This is clearly a trade-off in such systems, as more parameters in the model and increased complexity makes interpretation more difficult.

4.2.4 Qualitative System

In the Qualitative phase of the ChaLearn CVPR 2017 Challenge, the goal was to explain predictions with a human-understandable text. The TUD system implements a simple text description generator, with the following justifications:

- Each of the visual and textual features were picked to be explainable in natural language to non-technical users. However, no formal proof was given that the features are fully valid predictors of personality traits or interviewability. While the model gives indicators on the strongest linear coefficients, the assessments it was trained on are made by external observers (crowdsourcing workers), which poses a very different situation from the assessment settings in the formal psychology studies, as reported in [63]. Therefore, these features do not constitute a comprehensive panel of "good" features, despite their good predictive power.
- It may be possible to aggregate feature observations to higher-level descriptions (in particular, regarding AU detections, as combinations of AUs may indicate higher-level emotional expressions), but as this would increase the complexity of the model, only a basic explanation using individual low-level features was kept.
- As the feature measurements did not formally get tested (yet) in terms of psychometric validity, it is debatable to consider feature measurements and predicted scores as absolute indicators of interviewability. However, for each person, it was indicated whether the person scores "unusually" with respect to a larger population of "representative subjects" (formed by the vloggers represented in the 6000-video training set). Therefore, for each feature measurement, the system reports the typical range of the features and the percentile of the subject, compared to scores in the training set.
- Finally, to reflect major indicators from the linear model in the description, for each representation (OpenFace, MEI, Readability, Text) the two largest linear regression coefficients that were picked. 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 notice is added to the description, expressing how the feature commonly affects the final scoring (e.g. 'In our 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 set, a fairly long, but consistent textual description was generated. An example fragment of the description is given in Figure 10.

```
******************
* 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. 10 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 5. 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 5 Results of the first stage of the job screening coopetition. * Leading evaluation measure.

Rank	**Team**	**Invite-Interview** *	*Agreeableness*	*Conscientiousness*	*Extroversion*	*Neuroticism*	*Openness*
1	BU-NKU [40]	0.920916 (1)	0.913731 (1)	0.919769 (1)	0.921289 (1)	0.914613 (1)	0.917014 (1)
-	Baseline [32]	0.916202 (2)	0.911230 (2)	0.915228 (2)	0.911220 (3)	0.910378 (2)	0.911123 (2)
2	PML [18]	0.915746 (3)	0.910312 (3)	0.913775 (3)	0.915510 (2)	0.908297 (3)	0.910078 (3)
3	ROCHCI	0.901859 (4)	0.903216 (4)	0.894914 (4)	0.902660 (4)	0.901147 (4)	0.904709 (4)
4	FDMB	0.872129 (5)	0.891004 (5)	0.865975 (5)	0.878842 (5)	0.863237 (5)	0.874761 (5)

Interestingly, only one out of the four valid submissions outperformed the baseline method described in the previous section. The performance of the top three methodologies was quite similar, but the methodologies were not. In the following, we provide a short description of the different methods.

- **BU-NKU.** This approach was detailed in Section 4.1, see [40] for further details.
- **PML.** [**18**] 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 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. [23]), including the first impressions challenge [62,21]. 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.

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

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 6 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 3.4.3, and a [0,5] scale was adopted. It can be seen from this table that both methods obtained comparable performance. BU-NKU outperformed clearly the TUD team in terms of perceived clarity and interpretability, whereas the opposite happened in terms of creativity.

Table 6 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 6 illustrate that there is room for improvement for developing proper explanations. In particular, evaluation measures for explainability deserve further attention.

4.4 Discussion

This section described the design of the job candidate screening challenge, as well as its top performing submissions. The challenge comprised two stages, one of which focused entirely on generating explanations for the recommendations made by models. Out of 72 participating teams, only two teams successfully completed both stages, illustrating the difficulty in generating explanations for complex machine learning pipelines.

The two solutions that were described in detail comprise different methodologies and focus on different aspects. The BU-NKU system focused mostly on visual information for explaining recommendations. Audio was used as a complementary feature. The TDU system, on the other hand, gave more priority to textual information. Both methodologies are quite interesting and surely will be the basis for further research in this growing topic.

5 Analysis of the First Impressions data set

The collected personality traits dataset is rich in terms of the number of videos and annotations, and hence suitable for training models with high generaliza-

tion 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 this section, existence of this latent bias towards gender[10] and apparent ethnicity is analyzed. For this purpose, the videos used in the challenge are further manually annotated for gender and ethnicity, to complememnt 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

[10] We follow the computer science literature here, and use "gender estimation," but distinguishing male vs. female is more appropriately termed as "sex estimation". Gender is a more complex and subjective construct.

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

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. The results are summarized in Figure 11.

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 although non-consciously, people prefer 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.

The analysis provided in this section 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.

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 [79].

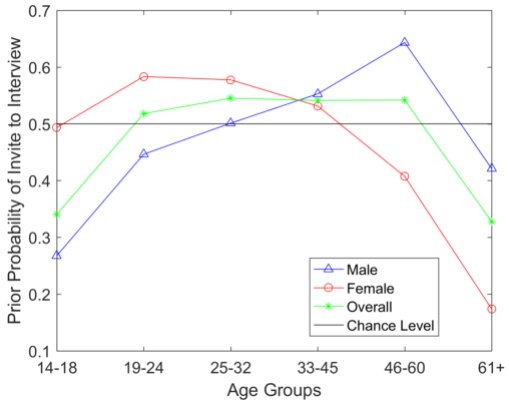


Fig. 11 The prior probability of job interview invitation over age groups, and over gender and age groups jointly.

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). In this paper, 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 [26], 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.

Acknowledgements The challenge organizers gratefully acknowledge a grant from Azure for Research, which allowed running the challenge on the Codalab platform and the technical support of Université Paris-Saclay. ChaLearn provided prizes and travel awards to the winners. This work was partially supported by CONACyT under grant 241306, Spanish Ministry projects TIN2016-74946-P and TIN2015-66951-C2-2-R (MINECO/FEDER, UE), and CERCA Programme / Generalitat de Catalunya. H.J. Escalante was supported by *Red Temáticas CONACyTs en Tecnologías del Lenguaje (RedTTL) e Inteligencia Computacional Aplicada (RedICA)*. A.A. Salah was supported by the BAGEP Award of the Science Academy. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

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. Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *FG*, volume 06, pages 1–6, 2015.
5. 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.
6. M. J. Black and Yaser Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *International Journal of Computer Vision*, 25(1):23–48, 1997.
7. Aaron F. Bobick and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

8. 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.
9. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
10. Leo Breiman et al. Heuristics of instability and stabilization in model selection. *The annals of statistics*, 24(6):2350–2383, 1996.
11. 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.
12. Meri Coleman and T. L. Liau. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975.
13. Mark Cook. *Personnel Selection: Adding Value Through People*. Wiley-Blackwell, fifth edition, 2009.
14. A. Vinciarelli D. Gatica-Perez and J. M. Odobez. *Interactive Multimodal Information Management*, chapter Nonverbal Behavior Analysis. EPFL Press, 2013.
15. DARPA. Broad agency announcement explainable artificial intelligence (XAI). In *DARPA-BAA-16-53, August 10, 2016*, 2016.
16. 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.
17. Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*, pages 525–538. Springer, 2012.
18. 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.
19. Paul Ekman and Wallace V Friesen. *Facial Action Coding System: Investigatoris Guide*. Consulting Psychologists Press, 1978.
20. Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.
21. 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.
22. 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.
23. Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. ChaLearn Looking at People: A review of events and resources. In *Proc. IJCNN*, 2017.
24. Sergio Escalera, Jordi Gonzàlez, Hugo Jair Escalante, Xavier Baró, and Isabelle Guyon. Editorial: Looking at people special issue. *International Journal of Computer Vision*, Forthcoming, 2017.
25. Sergio Escalera, Mercedes Torres Torres, Brais Martinez, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Georgios Tzimiropoulos, Ciprian Corneou, Marc Oliu, Mohammad Ali Bagheri, and Michel Valstar. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1–8, June 2016.
26. Diakopoulos N. et al. Principles for accountable algorithms and a social impact statement for algorithms. *Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2017.
27. 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.

28. 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.
29. Rudolf Flesch. A New Readability Yardstick. *The Journal of Applied Psychology*, 32(3):221–233, 1948.
30. 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.
31. 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.
32. 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.
33. 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.
34. R Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.
35. 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.
36. 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.
37. Y. Gucluturk, U. Guclu, M. Perez, H. Escalante, X. Baro, 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.
38. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
39. 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.
40. Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Multi-modal 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.
41. 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.
42. Heysem Kaya and Albert Ali Salah. Continuous mapping of personality traits: A novel challenge and failure conditions. In *Proceedings of the 2014 ICMI Workshop on Mapping Personality Traits Challenge*, pages 17–24. ACM, 2014.
43. 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.
44. 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.
45. 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.
46. 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.

47. Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
48. Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015.
49. 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.
50. Aravindh Mahendran and Andrea Vedaldi. Visualizing deep convolutional neural networks using natural pre-images. *Int. J. Comput. Vision*, 120(3):233–255, December 2016.
51. 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.
52. G.H. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.
53. Thomas B. Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal, editors. *Visual Analysis of Humans - Looking at People*. Springer, 2011.
54. Klaus-Robert Muller, Andrea Vedaldi, Lars Kai Hansen, Wojciech Samek, and Gregoire Montavon. Interpreting, explaining and visualizing deep learning workshop. *NIPS*, Forthcoming, 2017.
55. 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.
56. 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.
57. Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(8):1692–1706, 2016.
58. 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.
59. O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.
60. M. Pittore, M. Campani, and A. Verri. Learning to recognize visual dynamic events from examples. *International Journal of Computer Vision*, 38(1):35–44, 2000.
61. R. Polana and R.C. Nelson. Detection and recognition of periodic, nonrigid motion. *International Journal of Computer Vision*, 23(3):261–282, 1997.
62. 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.
63. 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.
64. 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.
65. 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.
66. 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.
67. Escalera Sergio, Isabelle Guyon, and Vassilis Athitsos, editors. *Gesture recognition*. Springer Series on Challenges in Machine Learning. Springer, 2017.

68. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
69. E A Smith and R J Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, pages 1–14, 1967.
70. 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.
71. Yaniv Tagiman, Ming Yang, Marc Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proc. of CVPR*, 2014.
72. Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
73. 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.
74. 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.
75. Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291, 2014.
76. Feng Wang, Haijun Liu, and Jian Cheng. Visualizing deep neural network by alternately image blurring and deblurring. *Neural Networks*, 2017.
77. 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.
78. Andrew Gordon Wilson, Jason Yosinski, Patrice Simard, and Rich Caruana. Interpretable ml symposium. *NIPS*, Forthcoming, 2017.
79. www foundation. Algorithmic accountability. *World Wide Web Foundation, Defense Advanced Research Projects Agency (DARPA)*, 2017.
80. X. Xiong and Fernando De la Torre. Supervised Descent Method and Its Application to Face Alignment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539, 2013.
81. Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *Proc. of ICML Deep Learning Workshop*, 2015.
82. Matthew D. Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks*, pages 818–833. Springer International Publishing, Cham, 2014.
83. 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.
84. Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. *CoRR*, abs/1702.04595, 2017.