Design of an Explainable Machine Learning Challenge for Video Interviews

Hugo Jair Escalante[1,2], Isabelle Guyon[1,6], Sergio Escalera[3,4], Julio Jacques Jr.[3,4],
Meysam Madadi[3], Xavier Baró[3,5], Stephane Ayache[9], Evelyne Viegas[7],
Yağmur Güçlütürk[8], Umut Güçlü[8], Marcel A. J. van Gerven[8], Rob van Lier[8]
[1] *ChaLearn*, California, USA, [2] *Instituto Nacional de Astrofísica, Óptica y Electrónica*, Mexico,
[3] *Computer Vision Center, UAB*, Barcelona, Spain, [4] *Dept. Mathematics and Computer Science, UB*, Spain,
[5] *EIMT, Open University of Catalonia*, Barcelona, Spain, [6] *Université Paris-Saclay*, France, [7] *Microsoft Research*, USA
[8] *Radboud University, Donders Institute for Brain, Cognition and Behaviour*, Nijmegen, the Netherlands
[9] *Aix Marseille Univ, CNRS, LIF, Marseille, France*
hugojair@inaoep.mx
http://chalearnlap.cvc.uab.es/

***Abstract*—This paper reviews and discusses research advances on "explainable machine learning" in computer vision. We focus on a particular area of the "Looking at People" (LAP) thematic domain: *first impressions and personality analysis*. Our aim is to make the computational intelligence and computer vision communities aware of the importance of developing explanatory mechanisms for computer-assisted decision making applications, such as automating recruitment. Judgments based on personality traits are being made routinely by human resource departments to evaluate the candidates' capacity of social insertion and their potential of career growth. However, inferring personality traits and, in general, the process by which we humans form a first impression of people, is highly subjective and may be biased. Previous studies have demonstrated that learning machines can learn to mimic human decisions. In this paper, we go one step further and formulate the problem of *explaining the decisions of the models* as a means of identifying what visual aspects are important, understanding how they relate to decisions suggested, and possibly gaining insight into undesirable negative biases. We design a new challenge on explainability of learning machines for first impressions analysis. We describe the setting, scenario, evaluation metrics and preliminary outcomes of the competition. To the best of our knowledge this is the first effort in terms of challenges for explainability in computer vision. In addition our challenge design comprises several other quantitative and qualitative elements of novelty, including a "coopetition" setting, which combines competition and collaboration.**

I. Introduction

Research progress in computer vision and pattern recognition has lead to a variety of modeling techniques with (almost) human-like performance in a variety of tasks. A clear example of this type of models are neural networks, whose deep variants dominate the arenas of computer vision and natural language processing among other fields. Although this type of models have obtained astonishing results in a variety of tasks (e.g., face recognition with facenet [1]), they are limited in their explainability and interpretability. That is, in general, users cannot say too much about:

- What is the rationale behind the decision made? (*explainability*)
- What in the model structure explains its functioning? (*interpretability*)

This in turns raises multiple questions about decisions – why a decision is preferred over others and how confident is the learning machine in its decision, what steps lead the learning machine's decision – and model structure – why a determined parameter configuration was chosen, what the parameters mean, how a user could interpret the learned model, what additional knowledge would be required from the user/world to improve the model. Hence, while reaching good prediction performance is critical, explainability/interpretability is a much desirable feature to include learning machines as part of *decision support systems*, for instance in medicine or security.

In this paper we focus our attention on explainability of learning machines in the area of computer vision. We briefly review on-going efforts in this direction, with emphasis on a very specific application within the so-called Looking at People (LAP) field: first impressions and personality analysis. We elaborate on the importance that explainability can have in this particular domain and review efforts from related fields. In addition, we describe a challenge we are organizing with the aim of advancing the state of the art on explainability of learning machines in first impressions and personality analysis tasks. The data, evaluation protocol, expected outcomes and preliminary results of this challenge are discussed.

The remainder of this paper is organized as follows. Section II briefly reviews related work on explainability of learning machines. Section III elaborates on the importance of explainability for first impressions and personality analysis tasks. Section IV describes in detail the proposed challenge. Section V presents a discussion on explainability for first impressions and personality trait analysis and outlines ongoing and future research directions.

II. Related work

Explainability is a fundamental topic within artificial intelligence (AI) [2]. In fact, one of the main motivations for the fields of knowledge-based [3] and probabilistic reasoning [4]

was developing explainable and interpretable models. So far, there are models for which gaining insights into their decisions and recommendations is possible, e.g., decision trees, causal models, Bayesian networks (even when an expert may be required to provide an explanation). However, for many modeling techniques (including those recently exhibiting best performance, such as "deep learning" techniques), the process behind a decision generated by the model remains largely unexplained despite recent efforts[1].

A lack of interpretability is particularly pressing for neural networks and has been an argument that is frequently employed by detractors of these models. The recent success of deep learning in several fields, like computer vision [1] and natural language processing [2][5], has motivated renewed efforts on methods that can help users gaining insights into the "behavior" of deep learning models. This ranges from the visual analysis of intermediate layers of models for image classification [6], which *visually* explain the low-level components (weights-level) of the model, to the explanation on decisions of the model based on mid-level predictions [7], [8] (e.g., explaining an event detected in a video by generating sentences using concepts recognized in frames), as well as models that in addition incorporate external knowledge [9].

Explanatory mechanisms not related to neural networks have been proposed for robotics [10], assistant and training systems [11], [12], health consultation systems [13] and computer vision developments not relying on deep learning [14]. There are also few attempts to generate explanation of predictions for generic models, seeing them as black boxes [15] (related to this direction there are the attempts to *justify* model predictions as well [16]).

Concerning computer vision and its applications, although there is a growing number of efforts on developing explanatory models and mechanisms, model interpretation is still in its infancy. We anticipate this field will become one of the hot topics in the next few years within AI in general (e.g., see the DARPA call on explainable AI [17]). In this direction, there are two research fields which will have a broad impact: natural language processing (for generating natural language explanations) and machine learning (the most promising modeling tool for developing explainable mechanisms). We anticipate that many efforts will be oriented to explaining deep learning models since these models presently dominate in challenges in computer vision (see e.g. the winning solutions for some recent challenges, [18], [19], [20], [21], [22], [23], [24], [25], [26]).

This paper lays the foundations for building explainable systems in computer vision. Concretely, we outline the design of the first academic challenge that aims at evaluating the explainability of models used to tackle a difficult computer vision problem. We focus on a problem for which explanations are critical for end users to take informed decisions, namely predicting first impressions. Deep learning/neural network models having achieved best results in past editions of our challenges[3] [18], [19], we expect that solutions developed in the context of this challenge will have an impact in related deep learning models and computer vision challenges.

[1]See for instance the contributions to the 2016 NIPS workshop https://sites.google.com/site/nips2016interpretml/

[2]http://googleblog.blogspot.com/2015/07/neon-prescription-or-rather-new.html

III. Explaining first impressions

It is well known that the first impression one makes is highly important in many contexts, such as job interviews, teaching/learning environments, presentations/talks, networking, and of course in the daily social context (e.g., meeting new people, dating, etc.). First impressions can be defined as rapid judgments of personality traits and complex social characteristics like dominance, hierarchy, warmth, and threat [27], [28], [29]. It is known that humans can form first impressions on *stereotypical* personality traits from faces as fast as 100ms after stimulus onset [30]. Other studies suggest that, with more time, observers can form very accurate first impressions of traits when exposed to streams (video, audio, text, etc.) of individuals' behavior [27], [31]. These findings in social psychology have motivated computer scientist to explore the capabilities of natural language processing, computer vision and pattern recognition methods for recognizing personality traits and forming first impressions.

Automatic methods for personality trait recognition have been studied for a while in natural language processing [32], [33], [34], [35], [36], [37]. However, first impression recognition techniques in computer vision is an emerging topic. Some efforts for recognizing personality traits from still images [38] and videos [39], have been proposed. Very recently we organized two rounds of a challenge in which we aimed to automatically infer apparent personality traits of people from very short clips [18], [19]. Contrary to existing work, our focus was to recognize traits with limited information (15s video clips). We found that top ranked participants were able to predict personality traits with an area under ROC curve above 0.8 for most traits, by processing 15s video clips. It is important to mention that the top performing methods were based on deep learning approaches, see [40], [41], [42], [43], [19], [18] for details.

However, even though remarkable progress has been reported recently, what determines a first impression is still debatable and it is not even clear to trained humans. Understanding which aspects/features/variables may trigger a decision in favor of a particular trait or favor a particular positive/negative first impression is decisive in at least the following scenarios:

- **Analysis:** Measurable/quantitative aspects important for automatic systems can be corroborated by findings in psychology and social sciences to justify hiring decisions [44].
- **Training:** Explainable first impression recognition systems may be instrumental in developing training curricula

[3]Past editions exclusively dealt with the prediction problem rather than explainability, see Section IV and [18], [19]

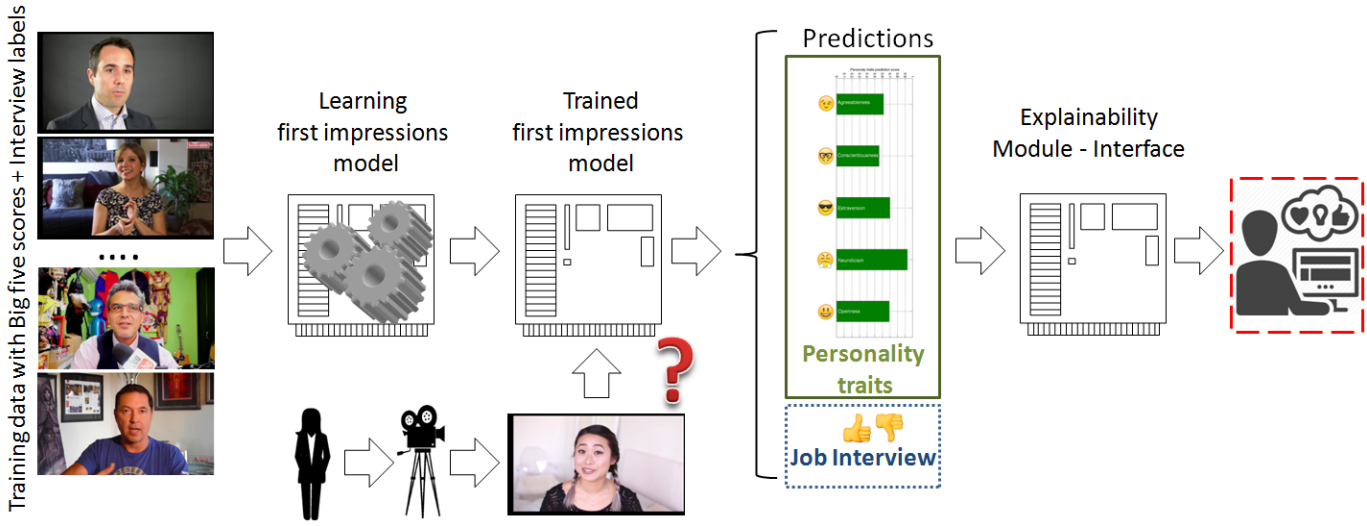

Fig. 1. Diagram of the considered scenario in the job candidate screening coopetition. The solid (green) square indicates the variables evaluated in past editions of the challenge [19], [18]. The dotted (blue) square indicates the variable to be evaluated in the quantitative track of this challenge. The dashed (red) square indicates what will be evaluated in the qualitative track of the challenge.

for job seekers and recruiters, or more generally speakers, negotiators, etc.

Hence, fully explainable models for first impression recognition would have a broad impact in the fields of affective computing, social signal processing, social psychology, and social sciences in general. Achieving a level of explainability capable of supporting the above fields is a daunting task that requires coordinated efforts from multiple disciplines. This motivated us to organize an academic competition of explainable models for first impressions recognition.

Figure 1 depicts the general scenario of the challenge. In a nutshell, participants of the so called *Job Candidate Screening Challenge* will develop methods for deciding whether a job candidate should be invited or not for an interview, using a short video of that candidate. Target values for hiring preferences have been obtained from human annotators. To facilitate the task, the videos will also be annotated (using human expertise) with personality traits. Hence, as a subtask, the challenge participants will be invited to also predict personality traits and eventually use them to predict hiring preferences (the main goal). The principal focus of the challenge is to induce participants into developing user interfaces that generate explanations and recommendations to users, supporting, and explaining the predictions. To that end, the challenge implements a coopetition scheme favoring collaborations between competitors towards advancing the state-of-the-art.

Because we aim to exploit creativity of participants for approaching this novel problem, we place no restrictions on the type of information they can use (e.g., audio, video, text information derived from the clips, and even external knowledge) as long its use does not interfere with the rules of the challenge. In addition, the participants can use any type of methodology (rule based, inductive model based, etc.) and any type of explanation recommendation (textual, audiovisual, etc.). The winners of the challenge will be determined by a panel of experts in the fields of psychological behavior analysis, recruitment, machine learning and computer vision.

The challenge has the potential to advance the state of the art in a number of directions related to explanatory modeling. We foresee this first challenge will motivate further research on explainability in computer vision systems and will have an impact in a number of novel applications.

IV. Job Candidate Screening Coopetition

This section describes in some details the setting of the academic challenge we propose, aiming at motivating research on explainability for first impressions and apparent personality analysis. A general diagram of the considered scenario is shown in Figure 1.

A. Overview

With the aim of assessing the importance that explainability has in first impressions and apparent personality analysis, we are organizing the first academic challenge on explainable computer vision and pattern recognition to assess "first impressions" on personality traits. The specific goal of the challenge is to devise automated methods for deciding whether a job candidate has to be interviewed or not, using short video clips (see data description below). It is implicitly assumed that the candidate has already sucessfully passed technical screening interview steps e.g. based on CV review. We address the part of the interview process related only to **human factors**, complementing aptitudes and competence, supposed to have been separately evaluated. Although this setting is simplified, the challenge is 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. This challenge is 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.

This challenge is related to two previous 2016 competitions on first impressions that were part of the contest programs of ECCV2016 [18] and ICPR2016 [19]. Both past challenges focused on predicting the apparent personality of candidates in video. In this new round of the challenge, we aim at predicting **hiring recommendations** in a candidate screening process, i.e. whether a job candidate is worth interviewing (a task not previously explored). In addition, we focus 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.

The job candidate screening challenge has been divided into two tracks, comprising quantitative and qualitative variants of the competition, tracks being 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 is expected at this stage.

Each competition adopts a different evaluation protocol. Figure 1 graphically indicates what information will be evaluated in each variant. In both cases, participants are free (and encouraged) to use information from apparent personality analysis. Likewise, since this challenge is a *coopetition*, participants are expected to share their code and use other participants' code, mainly for the second stage of the challenge: e.g., a team can participate only in the qualitative competition using the solution of another participant in the quantitative competition.

B. Data

For the challenge we use the data set used in previous competitions [19], [18], but extended with a predictive variable that has not been used previously: "Invite for interview" (referred to as "job-interview variable"). The first impressions data set, comprises 10,000 clips (average duration 15s) extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing and speaking in English to a camera. People in videos show different gender, age, nationality, and ethnicity. Figure 2 shows snapshots of sample videos from the data set.

Videos are labeled both with personality traits and the "job-interview variable". Amazon Mechanical Turk (AMT) was used for generating the labels. A principled procedure was adopted to guarantee the reliability of labels, converting rankings provided by labelers into normalized real valued scores (see [45] for details). The considered personality traits

Fig. 2. Snapshots of sample videos from the First Impressions data set [18].

were those from the Five Factor Model (also known as the Big Five), which is the dominant paradigm in personality research. It models human personality along five dimensions: *Extraversion, Agreeableness, Conscientiousness, Neuroticism* and *Openness*. Thus each clip has ground truth labels for these five traits. In addition to labeling the *apparent* personality traits, AMT workers labeled each video with a variable indicating whether the person should be invited or not to a job interview (the "job-interview variable"). This variable was also subject to the post processing reported in [45], so the variable to be predicted is a real number. The reader is referred to [18] where the data set is described in more details. In the previous editions of the first impressions challenge, participants had to predict only the personality traits of people (see detailed results in [18], [19]).

The data set used for the job candidate screening competition has been also extended in terms of the information that participants can use as input for their models. Every video was annotated to contain the transcriptions of audio. Each 15s YouTube video in the data set was transcribed independently. 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 (http://www.rev.com) to ensure maximum quality of the ground truth annotations. This newly added data dimension will make it possible for competitors to use higher level, contextual information in their models. Likewise, we expect participants use transcriptions to generate explanations of their methods.

For the quantitative track of the job candidate screening coopetition, the participants will have to predict the "job-interview variable". For the qualitative track, the participants will have to "explain", why a predictive model makes a recommendation. For both tracks, participants are encouraged to use information from personality traits, which could be used as extra features helping to predict the "job-interview variable" and/or serve as *arguments* for the explanatory mechanisms. Note that the personality traits labels will be provided *only with training data*.

The feasibility of the challenge annotations has already been successfully evaluated. The reconstruction accuracy of all annotations is greater than 0.65. Furthermore, the apparent trait annotations are highly predictive of invite-for-interview annotations with a significantly above-chance coefficient of determination of 0.91.

C. Evaluation protocol

As in previous challenges organized by ChaLearn[4] the job candidate screening coopetition will run in CodaLab[5]; a platform developed by Microsoft Research and Stanford University in close collaboration with the organizers of the challenge.

For the evaluation, the data set will be divided into the following partitions:

- **Development (training)** data with ground truth for all of the considered variables (including personality traits) will be made available at the beginning of the competition.
- **Validation** data **without labels** (neither for personality traits nor for the "job-interview variable") will be also provided to participants at the beginning of the competition. Participants can submit their predictions on validation data to the CodaLab platform and receive immediate feedback on their performance (there will be a validation leader board in the platform).
- **Final evaluation (test)** unlabeled data will be made available to participants one week before the end of the quantitative challenge. Participants will have to submit their predictions in these data to be considered for the final evaluation (no ground truth will be released at this point).

In addition to submitting predictions for test data, participants desiring to compete for prizes will submit their code for verification together with fact sheets summarizing their solutions.

The winners of the challenge in the different tracks will be determined as follows (where both competition stages will be independently evaluated, and top 3 ranked participants at each stage will be awarded):

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

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

  where p_i is the predicted score for sample i, t_i is the corresponding ground truth value, with the sum running over the N_t test videos, and $\bar{t}$ is the average ground truth score over all videos.
- **Qualitative evaluation (explanatory mechanisms).** Participants should provide a textual description that explains the decision made for the interview variable in test data. Optionally, participants can also submit a visual description to enrich and improve clarity and explainability. Performance will be evaluated in terms of the creativity of participants and the explanatory effectiveness of the mechanisms-interface. For this evaluation we will invite 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 will be made based on a small subset of the videos. Specifically, a small subset of videos from the validation data and a small subset of videos from the test data will be systematically selected to best represent the variability of the personality traits and invite-for-interview values in the entire dataset. The jury will only evaluate a single validation and a single test phase submission per participant. A separate jury member will serve as a tiebreaker. At the end, the creativity criteria will be judge globally according to the evaluated clips, as well as an optional video that participant can submit to describe their method. Figure 3 shows an illustration of the jury interface for the qualitative evaluation phase.

For each evaluated clip, the evaluation criteria for the jury will be:

- *Clarity:* Is the text understandable / written in proper English?
- *Explainability:* Does the text provide relevant explanations to the hiring decision made?
- *Soundness:* Are the explanations rational and, in particular, do they seem scientific and/or related to behavioral cues commonly used in psychology.

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

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

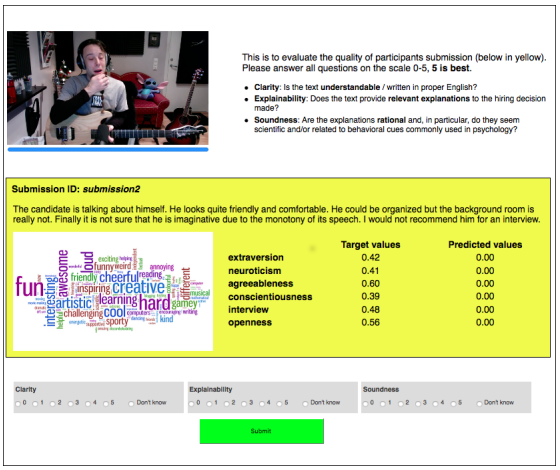

Fig. 3. Qualitative evaluation interface

[4]http://chalearn.org
[5]http://codalab.org/

It is expected that at this stage, participants of the first stage share their code, which can be used by any user participant of the second stage, see below.

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

The timeline for the challenge is as follows:

- *10th January, 2017:* Beginning of the quantitative competition, release of development and validation data.
- *10th February, 2017:* Test data release and deadline for code submission for the first (quantitative) track. Code and fact sheets submission deadline.
- *15th February, 2017:* End of quantitative competition. Start of qualitative coopetition. Code is shared among participants.
- *8th April, 2017:* End of qualitative coopetition. Qualitative coopetition code and fact sheets submission deadline.
- *14-19th May, 2017:* Presentation of results in IJCNN 2017.
- *26th July, 2017:* Presentation of results at the ChaLearn Workshop on Explainable Computer Vision Multimedia and Job Candidate Screening Coopetition at CVPR2017.

D. Preliminary results

At the moment of writing this paper, the first round of the competition has been finished. At this stage, 53 participants were registered for the challenge. We expect more participants to join for the second stage, as code of the top ranked participants will be shared. At this stage, four valid submissions were considered for the prizes, the performance of these submissions are shown in Table I. Recall, the leading evaluation measure is the classification performance (see Equation 1) in the Invite-for-interview variable, although we also show results for personality traits, as participants submitted predictions for these variables as well.

Interestingly, only one out of the 4 valid submissions outperformed the baseline method described in detail in [46]. This is not surprising as this strong baseline was built on top of the solution of a top ranked entry of the first impressions challenge [43], [18]. Nevertheless, the performance of the top 3 methodologies is quite similar. In fact, the differences in performance can only be appreciated at the third decimal of the accuracy (this applies for personality traits as well). All four participants agreed to share their code for the coopetition stage, therefore, participants of the second stage can choose on the different methods and be sure that performance is quite similar. Despite performance is similar, methodologies were not, in the following we provide a succinct description of the different methods.

- **BU-NKU.** The top ranked team used four types of features. Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) features were extracted from facial images, also, features extracted with a (pretrained) deep convolutional network (DCNN) were considered. As audio descriptors, this team extracted the baseline set of features as specified in [47] using the openSMILE library [48]. Finally, the authors also used a DCNN to extract scene information, in an effort to extract useful contextual information. This team also performed preliminary experiments including ASR transcripts, however text descriptors were not included in the winning entry. The four feature types were combined in two stages for generating its predictions. In a first stage, LGBP-TOP with facial DCNN and audio with scene features were combined via linear kernel extreme learning machines (ELM). Next a random forest model combined the outputs of the two ELMs. This last method was trained to predict the interview variable and the 5 personality traits.
- **PML.** This team 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 are extracted at different scales of each frame using a grid. Feature vectors from each region and each resolution are concatenated, the representation for a video is 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.** This team extracted a set of predefined multimodal features and used gradient boosting for predicting the interview variable. Facial features and meta attributes extracted with the SHORE[6] library 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.** This team used frame differences and appearance descriptors at multiple fixed image regions with a Support Vector regression (SVR) method for predicting the interview variable and the five personality traits. After face detection and normalization, differences between consecutive frames is extracted. Local Phase Quantization (LPQ) descriptors are extracted from each region in each frame and are concatenated. The video representation is obtained by adding the image-level descriptor. SVR is used to estimate traits and the interview variable. This method only relied on visual information.

Despite that only four teams provided qualified to the final phase of the first stage, it was encouraging to have 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

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

Rank	Team	Invite-Interview $*$	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
1	BU-NKU	0.920916 (1)	0.913731 (1)	0.919769 (1)	0.921289 (1)	0.914613 (1)	0.917014 (1)
-	baseline[46]	0.916202 (2)	0.911230 (2)	0.915228 (2)	0.911220 (3)	0.910378 (2)	0.911123 (2)
2	PML	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)

TABLE I
RESULTS OF THE FIRST STAGE OF THE JOB SCREENING COOPETITION.

this stage. This is in contrast with previous challenges in most aspects of computer vision (see e.g. [25]), including the first impressions challenge [18], [19]. 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.

We are certain that the code from all four submissions will be quite helpful for participants in the second stage of the challenge; not only in terms of performance, but in terms of features and learning methodologies, which can be extended to generate explanations.

V. DISCUSSION

Explaining model decisions and recommendations has been a topic of interest for AI since its early days. In fact, explainalibity has been considered as a core problem in numerous subfields and entire research areas in AI (e.g., expert systems). Because of that, research progress on explainability for different tasks and applications has been impressive. Yet, explainability of computer vision models is a field still in its infancy. Therefore, there are many open problems and research opportunities. This paper represents a first effort in this direction by describing the design of a challenge on explainability of models for first impressions and personality analysis. The approached task is interesting from different perspectives (e.g., computer vision, pattern recognition, machine learning, affective computing, social signal processing, social psychology, etc.) and can have impact into practical applications (e.g., training systems for job candidates, recruiters and actors).

The considered scenario, data, evaluation protocol, and timeline are described. Preliminary results, covering the first stage of the challenge, were also reported. More than 5o participants registered for the challenge. Performance was very similar across entries that qualified to the final phase and all participants agreed to share their code. Interestingly, most of these entries did not relied on deep learning methodologies and they are quite diverse to each other. In general terms, we believe this diversity is enriching and quite beneficial for the upcoming second stage of the challenge, as participants will have available plenty of varied resources to develop explanatory mechanisms.

ACKNOWLEDGMENTS

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. The first author was supported by *Red Temática CONACyT en Tecnologías del Lenguaje.*

REFERENCES

[1] F. Schroff, D. Kalenichenko, and P. J., "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[2] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Pearson Education, 2003.

[3] R. Brachman and H. Levesque, *Readings in Knowledge Representation*, ser. Morgan Kaufmann readings series. M. Kaufmann Publishers, 1985.

[4] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann;, 1998.

[5] R. Collobert, j. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.

[6] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[7] Q. Yu, J. Liu, H. Cheng, A. Divakaran, and H. Sawhney, "Multimedia event recounting with concept based representation," in *Proceedings of the 20th ACM International Conference on Multimedia*, ser. MM '12. New York, NY, USA: ACM, 2012, pp. 1073–1076. [Online]. Available: http://doi.acm.org/10.1145/2393347.2396386

[8] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann, "Devnet: A deep event network for multimedia event detection and evidence recounting," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2568–2577.

[9] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *arXiv:1603.08507*, 2016.

[10] M. Lomas, R. Chevalier, E. V. Cross, II, R. C. Garrett, J. Hoare, and M. Kopack, "Explaining robot actions," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '12. New York, NY, USA: ACM, 2012, pp. 187–188. [Online]. Available: http://doi.acm.org/10.1145/2157689.2157748

[11] F. Elizalde, E. Sucar, J. Noguez, and A. Reyes., "Integrating probabilistic and knowledge-based systems for explanation generation," in *Proceedings of the Third International Workshop on Explanation-aware Computing*, ser. Exact '08, 2008.

[12] H. C. Lane, M. Core, M. van Lent, S. Solomon, and D. Gomboc, "Explainable Artificial Intelligence for Training and Tutoring," in *12th International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands, Jul. 2005. [Online]. Available: http://ict.usc.edu/pubs/Explainable%20Artificial%20Intelligence%20for%20Training%20and%20Tutoring.pdf

[13] R. L. Teach and E. H. Shortliffe, "An analysis of physician attitudes regarding computer-based clinical consultation systems," *Computers and Biomedical Research*, vol. 14, no. 6, pp. 542 – 558, 1981. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0010480981900124

[14] T. Berg and P. N. Belhumeur, "How do you tell a blackbird from a crow?" in *Proceedings of ICCV*, 2013.

[15] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *arXiv:1602.04938*, 2016.

[16] O. B. orb and K. McKeown, "Justification narratives for individual classifications," in *Proceedings of ICML AutoML Workshop 2014*, 2015.

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

[18] V. P. Lopez, B. Chen, A. Clapes, M. Oliu, C. Corneanu, X. Baro, H. J. Escalante, I. Guyon, and S. Escalera., "Chalearn lap 2016: First round challenge on first impressions - dataset and results," in *ChaLearn Looking at People Workshop on Apparent Personality Analysis. ECCV Workshop proceedings, LNCS, Springer, 2016, in press.*, 2016.

[19] H. J. Escalante, V. Ponce, J. Wan., M. Riegler, C. 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.

[20] S. Escalera, M. Torres-Torres, B. Martinez, X. Baro, H. J. Escalante, I. Guyon, G. Tzimiropoulos, C. Corneanu, M. O. Simón, M. A. Bagheri, and M. Valstar., "Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016," in *Proceedings of CVPRW*, 2016.

[21] S. Escalera, J. Fabian, P. Pardo, X. Baro, J. Gonzalez, H. J. Escalante, D. Misevic, U. Steiner, and I. Guyon, "Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results," in *Proceedings of ICCVW*, 2015.

[22] X. Baro, J. Gonzalez, J. Fabian, M. A. Bautista, M. Oliu, H. J. Escalante, I. Guyon, and S. Escalera, "Chalearn looking at people 2015 challenges: action spotting and cultural event recognition," in *Proceedings of CVPRW*, 2015.

[23] S. Escalera, X. Baro, J. Gonzalez, M. A. Bautista, M. Madadi, M. Reyes, V. Ponce, H. J. Escalante, J. Shotton, and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results," in *Proceedings of ECCVW*, 2014.

[24] S. Escalera, J. Gonzalez, X. Baro, H. J. Escalante, and I. Guyon, "Chalearn looking at people events," *IAPR NewsLetter*, vol. 37, no. 4, pp. 13–15, 2015.

[25] S. Escalera, X. Baró, H. J. Escalante, and I. Guyon, "Chalearn looking at people: A review of events and resources," in *Proc. IJCNN*, 2017.

[26] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *arXiv:1606.01847*, 2016.

[27] N. Ambady, F. J. Bernieri, and J. A. Richeson, "Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream," *Adv. Exp. Soc. Psycho.*, vol. 32, pp. 201–271, 2000.

[28] D. S. Berry, "Taking people at face value: Evidence for the kernel of truth hypothesis," *Social Cognition*, vol. 8, no. 4, p. 343, 1990.

[29] R. Hassin and Y. Trope, "Facing faces: studies on the cognitive aspects of physiognomy." *JPSP*, vol. 78, no. 5, p. 837, 2000.

[30] J. Willis and A. Todorov, "First impressions making up your mind after a 100-ms exposure to a face," *PSS*, vol. 17, no. 7, pp. 592–598, 2006.

[31] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychol. Bull.*, vol. 111, no. 2, p. 256, 1992.

[32] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transaction on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[33] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *AAAI Technical Report WS-13-01 Computational Personality Recognition (Shared Task)*, 2013, pp. 2–5.

[34] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez, "Hi youtube!: Personality impressions and verbal content in social video," in *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ser. ICMI '13. New York, NY, USA: ACM, 2013, pp. 119–126. [Online]. Available: http://doi.acm.org/10.1145/2522848.2522877

[35] F. Alam, E. A. Stepanov, and G. Riccardi, "Personality traits recognition on social network - facebook," in *AAAI Technical Report WS-13-01*, 2013.

[36] F. Rangel, F. Celli, P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at PAN 2015," in *CLEF 2015 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings Vol. 1391*, 2015.

[37] M. A. A. Carmona, A. P. L. Monroy, M. M. Y. Gomez, L. V. Pineda, and H. J. Escalante, "INAOE's participation at PAN'15: Author profiling task," in *CLEF 2015 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings Vol. 1391*, 2015.

[38] F. Celli, E. Bruni, and B. Lepri, "Automatic personality and interaction style recognition from facebook profile pictures," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 1101–1104. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654977

[39] G. Chávez-Martínez, S. Ruiz-Correa, and D. Gatica-Perez, "Happy and agreeable?: Multi-label classification of impressions in social video," in *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '15. New York, NY, USA: ACM, 2015, pp. 109–120. [Online]. Available: http://doi.acm.org/10.1145/2836041.2836051

[40] C.-L. Zhang, H. Zhang, X.-S. Wei, and J. Wu, "Deep bimodal regression for apparent personality analysis." in *ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings*, 2016.

[41] F. Gürpınar, H. Kaya, and A. Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation," in *Proc. ICPRW*, 2016.

[42] A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," in *ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings*, 2016.

[43] Y. Güçlütürk, U. Güçlü, M. A. J. van Gerven, and R. van Lier, "Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition," in *ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings*, 2016.

[44] A. Todorov, C. Y. Olivola, R. Dotsch, and P. Mende-siedlecki, "Social Attributions from Faces : Determinants , Consequences , Accuracy , and Functional Significance," vol. 66, pp. 519–545, 2015.

[45] B. Chen, S. Escalera, I. Guyon, V. Ponce-Lopez, N. Shah, and M. O. Simon, "Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits," in *ChaLearn Looking at People Workshop on Apparent Personality Analysis. ECCV Workshop proceedings, LNCS, Springer, 2016, in press.*, 2016.

[46] Y. Güçlütürk, U. Güçlü, X. Baró, H. J. Escalante, I. Guyon, S. Escalera, M. A. J. van Gerven, and R. van Lier, "Multimodal first impression analysis with deep learning," *Submitted: IEEE Transaction on Affective Computing*, 2016.

[47] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. of INTERSPEECH*, 2013, pp. 148–152.

[48] F. E. amd M. Wöllmer and B. Schuller, "Opensmile: the munich versatile and fast opensource audio feature extractor," in *Proc. of the intl. conf. on Multimedia*. ACM, 2010, p. 1459–1462.