ChaLearn LAP 2016: First Round Challenge on First Impressions - Dataset and Results

Víctor Ponce-López[1,2,6], Baiyu Chen[4], Marc Oliu[6], Ciprian Cornearu[1], Albert Clapés[2], Isabelle Guyon[3,5], Xavier Baró[1,6], Hugo Jair Escalante[3,7], Sergio Escalera[1,2,3]

[1] Computer Vision Center, Campus UAB, Barcelona, Spain
[2] Dept. Mathematics, University of Barcelona, Spain
[3] ChaLearn, California, USA
[4] UC Berkeley, California, USA
[5] U. Paris-Saclay, Paris, France
[6] EIMT/IN3 at the Open University of Catalonia, Barcelona, Spain
[7] INAOE, Puebla, Mexico

Abstract. This paper summarizes the ChaLearn Looking at People 2016 First Impressions challenge data and results obtained by the teams in the first round of the competition. The goal of the competition was to automatically evaluate five "apparent" personality traits (the so-called "Big Five") from videos of subjects speaking in front of a camera, by using human judgment. In this edition of the ChaLearn challenge, a novel data set consisting of 10.000 shorts clips from YouTube videos has been made publicly available. The ground truth for personality traits was obtained from workers of Amazon Mechanical Turk (AMT). To alleviate calibration problems between workers, we used pairwise comparisons between videos, and variable levels were reconstructed by fitting a Bradley-Terry-Luce model with maximum likelihood. The CodaLab open source platform was used for submission of predictions and scoring. The competition attracted, over a period of 2 months, 84 participants who are grouped in several teams. Nine teams entered the final phase. Despite the difficulty of the task, the teams made great advances in this round of the challenge.

Keywords: Behavior Analysis, Personality Traits, First Impressions.

1 Introduction

"You don't get a second chance to make a first impression", a saying famously goes. First impressions are rapid judgments of personality traits and complex social characteristics like dominance, hierarchy, warmth, and threat [1,2,3]. Accurate first impressions of personality traits have been shown to be possible when observers were exposed to relatively short intervals (4 to 10 min) of ongoing streams of individuals behavior [1,4], and even to static photographs present for 10s [2]. Most extraordinarily, trait assignment among human observers has been shown to be as fast as 100ms [5].

Personality is a strong predictor of important life outcomes like happiness and longevity, quality of relationships with peers, family, occupational choice, satisfaction, and performance, community involvement, criminal activity, and political ideology [6,7]. Personality plays an important role in the way people manage the images they convey in self-presentations and employment interviews, trying to affect the audience first impressions and increase effectiveness. Among the many other factors influencing employment interview outcomes like social factors, interviewer-applicant similarity, application fit, information exchange, preinterview impressions, applicant characteristics (appearance, age, gender), disabilities and training [8], personality traits are one of the most influential [9].

The key-assumption of personality psychology is that stable individual characteristics result into stable behavioral patterns that people tend to display independently of the situation [10]. The Five Factor Model (or the Big Five) is currently the dominant paradigm in personality research. It models the human personality along five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness. Many studies have confirmed consistency and universality of this model.

In the field of Computer Science, Personality Computing studies how machines could automatically recognize or synthesize human personality [10]. The literature in Personality Computing is considerable. Methods were proposed for recognizing personality from nonverbal aspects of verbal communication [11,12], multimodal combinations of speaking style (prosody, intonation, etc.) and body movements [13,14,15,16,17,18], facial expressions [19,20], combining acoustic with visual cues or physiological with visual cues [19,21,22,23]. Visual cues can refer to eye gaze [14], frowning, head orientation [22,23], mouth fidgeting [14], primary facial expressions [19,20] or characteristics of primary facial expressions like presence, frequency or duration [19].

As far as we know, there is no consistent data corpus in personality computing and no bench-marking effort has yet been organized. It is a great impediment in the further advancement of this line of research and the main motivator of this challenge. This challenge is part of a larger project which studies outcomes of job interviews. We have designed a dataset collected from publicly available YouTube videos where people talk to the camera in a self-presentation context. The setting is similar to video-conference interviews. Consistent to research in psychology and the related literature in automatic personality computing we have labeled the data based on the Big Five model using the Amazon Mechanical Turk (see Section 3). We are running a second round for the ICPR 2016 conference. It will take the form of a coopetition in which participants both compete and collaborate by sharing their code.

This challenge belongs to a series of events organized by ChaLearn since 2011[1]: the 2011-2012, user dependent One-shot-learning Gesture Recognition challenge [24,25], the 2013-2014 user independent Multi-modal Gesture Recognition challenge, the 2014-2015 human pose recovery and action recognition [26,27], and the 2015-2016 cultural event recognition [28] and apparent age es-

[1] `http://gesture.chalearn.org/`

timation [29,30]. In this 2016 edition, it is the first time we organize the First Impression challenge on automatic personality recognition.

The rest of this paper is organized as follows: in Section 2 we present the schedule of the competition and the evaluation procedures, in Section 3 we describe the data we have collected, Section 4 is dedicated to presenting, comparing and discussing the methods submitted in the competition. Section 5 concludes the paper with an extended discussion and suggestions about future work.

2 Challenge protocol, evaluation procedure, and schedule

The ECCV ChaLearn LAP 2016 challenge consisted in a single track competition to quantitatively evaluate the recognition of the apparent Big Five personality traits on multi-modal audio+RGB data from YouTube videos. The challenge was managed using the CodaLab open source platform of Microsoft[2]. The participants had to submit prediction results during the challenge. The winners had to publicly release their source code.

The competition had two phases:

- A development phase during which the participants had access to 6,000 manually labeled continuous video sequences of 15 seconds each. Thus, 60% of the videos used for training are randomly grouped in 75 training batches. They could get immediate feedback on their prediction performance by submitting results on an unlabeled validation set of 2000 videos. These 2,000 videos used in validation represent 20% over the total set of videos and are also randomly grouped in 25 validation batches.
- A final phase during which the competitors could submit their predictions on 2,000 new test videos (the remainder 20% over the total set of videos, also grouped in 25 test batches). The prediction scores on test data were not revealed until the end of the challenge.

2.1 Evaluation metrics

The participants of the different teams trained their models to imitate human judgments consisting in continuous target values in the range [0, 1] for each trait. Thus, their goal was to produce for each video in the validation set or the test set, 5 continuous prediction values in the range [0, 1], one for each trait.

For this task (similar in spirit to a regression) the evaluation consisted in computing the **mean accuracy** over all traits and videos. Accuracy for each trait is 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 are the predicted scores, t_i are the ground truth scores, with the sum running over the N_t test videos, and $\bar{t}$ is the average ground truth score over

[2] https://competitions.codalab.org/

all videos[3]. Additionally, we also computed (but did not use to rank the participants) the coefficient of determination:

$$R^2 = 1 - \sum_{i=1}^{Nt}(t_i - p_i)^2 / \sum_{i=1}^{Nt}(t_i - \bar{t})^2 \ . \tag{2}$$

We also turned the problems into classification problems by thresholding the target values at 0.5. This way we obtained 5 binary classification problems (one for each trait). We used the Area under the ROC curve (AUC) to estimate the classification accuracy[4].

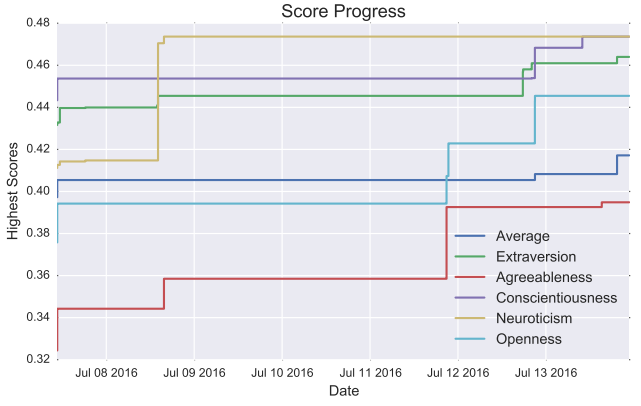

Fig. 1. Progress of validation set leaderboard highest scores of all teams for each trait and progress of the highest ranking score (mean accuracy over all traits). The score used is the accuracy, normalized as in Equation 1.

2.2 Schedule

The competition lasted two months and attracted 84 participants, who were grouped into several teams. The schedule was the following:

May 15, 2016: Beginning of the quantitative competition, release of the development data (with labels) and validation data (without labels).

June 30, 2016: Release of encrypted final evaluation data (without labels). Participants can start training their methods with the whole data set.

July 2, 2016: Deadline for code submission.

July 3, 2016: Release of final evaluation data decryption key. Participants start predicting the results on the final evaluation data.

[3] This definition is slightly different from what we used on the leaderboard. The leaderboard accuracy is not normalized $A = 1 - \frac{1}{Nt}\sum_{i=1}^{Nt}|t_i - p_i|$. This change does not affect the ranking.

[4] See e.g. `https://en.wikipedia.org/wiki/Receiver_operating_characteristic`.

July 13, 2016: End of the quantitative competition. Deadline for submitting the predictions over the final evaluation data. The organizers started the code verification by running it on the final evaluation data.

July 15, 2016: Deadline for submitting the fact sheets. Release of the verification results to the participants for review. Participants of the top ranked teams are invited to follow the workshop submission guide for inclusion at ECCV 2016 ChaLearn LAP 2016 Workshop on Apparent Personality Analysis.

As can be seen in Figure 1 progresses were made throughout the challenge and improvements were made until the very end. At the date the challenge ended, there was still a noticeable difference between the average of the best accuracies on the individual traits and the best accuracy of the teams, due to the fact that some of the team's methods performed better on some traits than others. This shows that there is still room for improvement and that the methods of the teams are complementary. We expect further improvements from the ongoing coopetition (second round of the challenge).

3 Competition data

The data set consists of 10,000 clips extracted from more than 3,000 different YouTube high-definition (HD) videos of people facing and speaking in English to a camera. The people appearing are of different gender, age, nationality, and ethnicity, which makes the task of inferring apparent personality traits more challenging. In this section, we provide the details about the data collection, preparation, and the final data set[5].

3.1 Video data

We collected a large pool of HD (720p) videos from YouTube. After visioning a large number of videos, we found Q&A videos to be particularly suitable and abundant talking-to-the-camera videos. These are generally videos with fewer people appearing, little moving background, and clear voice. Since YouTube videos are organized in channels, which can contain a variable number of videos, we limited the number of videos per YouTube channel (author) to 3 in order to keep a balance of unique subjects.

After having downloaded an initial pool of 13,951 YouTube videos using pytube Python's API[6], we manually filtered out unsuitable footage (too short sequences or non English speakers). From the remaining 8,581 videos, we automatically generated a set of 32,139 clips of 15 seconds each. The clip generation was automatically done by searching continuous 15-second video segments in which one and only one face appeared. Faces were detected using Viola-Jones from OpenCV [31]. We retained only faces with at least one visible eye – with

[5] Data set is available at `http://gesture.chalearn.org/2016-looking-at-people-eccv-workshop-challenge/data-and-description`

[6] PyTube API: `https://github.com/nficano/pytube`.

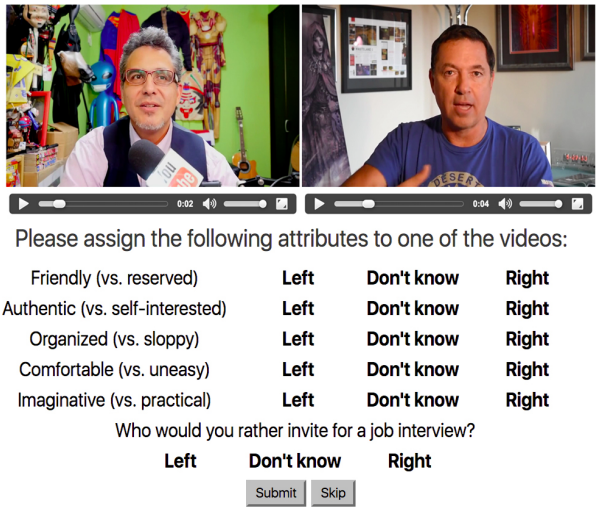

Fig. 2. Data collection web page. Comparing pairs of videos, the AMT workers had to indicate their preference for five attributes representing the "Big Five" personality traits, following these instructions: *"You have been hired as a Human Resource (HR) specialist in a company, which is rapidly growing. Your job is to help screening potential candidates for interviews. The company is using two criteria: (A) competence, and (B) personality traits. The candidates have already been pre-selected for their competence for diverse positions in the company. Now you need to evaluate their personality traits from video clips found on the Internet and decide to invite them or not for an interview. Your tasks are the following. (1) First, you will compare pairs of people with respect to five traits: Extraversion = Friendly (vs. reserved); Agreeableness = Authentic (vs. self-interested); Conscientiousness = Organized (vs. sloppy); Neuroticism = Comfortable (vs. uneasy); Openness = Imaginative (vs. practical). (2) Then, you will decide who of the 2 people you would rather interview for the job posted."* In this challenge we did not use the answers to the last question.

eyes being also detected using Viola-Jones. To increase robustness, we kept only those clips meeting both criteria ("one and only one face bounding box containing at least one eye") in 75% of the frames. Videos were of various duration, hence we limited the number of clips per video to at most 6.

We then performed a second fine-grained manual filtering – this time considering clips, instead of whole videos – using a custom web interface to filter out those clips not meeting the following criteria:

- One unique person as foreground at a safe distance from the camera.
- Good quality of audio and images.
- Only English speaking.
- People above 13-15 years old. Non-identified babies appearing with the parents might be allowed.
- Not too much camera movement (changing background allowed, but avoid foreground constantly blurred).

- No adult or violent contents (except people casually talking about sex or answering Q&A in an acceptable manner). Discard any libelous, doubtful or problematic contents.
- No nude (except if only parts above shoulders and neck are visible).
- Might have people in the background (crowd, audience, without talking, with low resolution of faces to avoid any confusion with the speaker).
- No advertisement (visual or audio information about products or company names).
- Avoid visual or audio cuts (abrupt changes).

From this second manual filter, we obtained the final set of 10,000 clips. These correspond to 3,060 unique originating videos. From those, we were able to generate a mean of 3.27 clips per video. In terms of time duration, the clips correspond to 41.6 hours of footage pooled from 608.7 hours of originating videos.

On the other hand, the originating videos were provided by 2,764 unique YouTube channels. Note, however, that the number of channels do not correspond to number of people (a youtuber can have different channels or participate in other youtubers' channels), but it provides an estimation of the diversity of people appearing in the data set. The originating videos are also quite diverse in both their number of views and their 5-star ratings, which also helped to alleviate bias towards any particular kind of videos. This information is summarized in Table 1 together with other statistics computed from videos' meta-data. The table is completed with the 20 most common keywords (or tags) associated to the originating videos. As we stated before, we focused on Q&A videos, often related to other video content such as vlogging, HOW TOs, and beauty tips (mostly makeup).

preparation	Downloaded videos	13,951* (HD 720p @ 30 FPS)
	Remaining videos (supervised from *)	8,581**
	Sampled videos per channel	3 (at most)
	Sampled clips per video	6 (at most)
	Clip length	15 seconds
	Candidate clips (sampled from **)	32,139†
final data set	**Final set of clips** (supervised from †)	**10,000**‡
	Total duration of clips	41.6 hours (4.5M frames)
	Unique channels (originating ‡)	2,764; {1 : 2,584, 2 : 161, 3 : 19}§
	Unique videos (originating ‡)	3,060; {1 : 721, 2 : 533, 3 : 464, 4 : 398, 5 : 435, 6 : 509}¶
	Mean no. clips per video	3.27
	Duration of originating videos	608.7 hours
	Total no. views of originating videos	More than 115M; {0-100 : 27.64%, 100-1K : 34.15%, 1K-10K : 22.68%, 10K-100K : 11.44%, >100K : 4.08%}‖
	Originating videos' avg. rating	4.6/5.0; {1 : 8, 2 : 11, 3 : 43, 4 : 1340, 5 : 1,395}
	Originating videos' keywords (top 20)	'Q&A', 'q&a', 'vlog', 'questions', 'makeup', 'beauty', 'answers', 'funny', 'Video Blog (Website Category)', 'question and answer', 'answer', 'question', 'fashion', 'Vlog', 'Questions', 'vlogger', 'how to', 'tutorial', 'q and a', 'Answers'

§ is a frequency count, i.e. how many channels contribute to the final set of 10,000 clips with 1, 2, or 3 clips respectively;
¶ analogously to (§), that is how many videos contribute to the 10,000 clips with 1, 2, . . ., 6 clips;
‖ is a relative frequency count of videos with a number of views ranging in different intervals (0 to 100, 100 to 1K, etc).

Table 1. Video data preparation and final data set statistics.

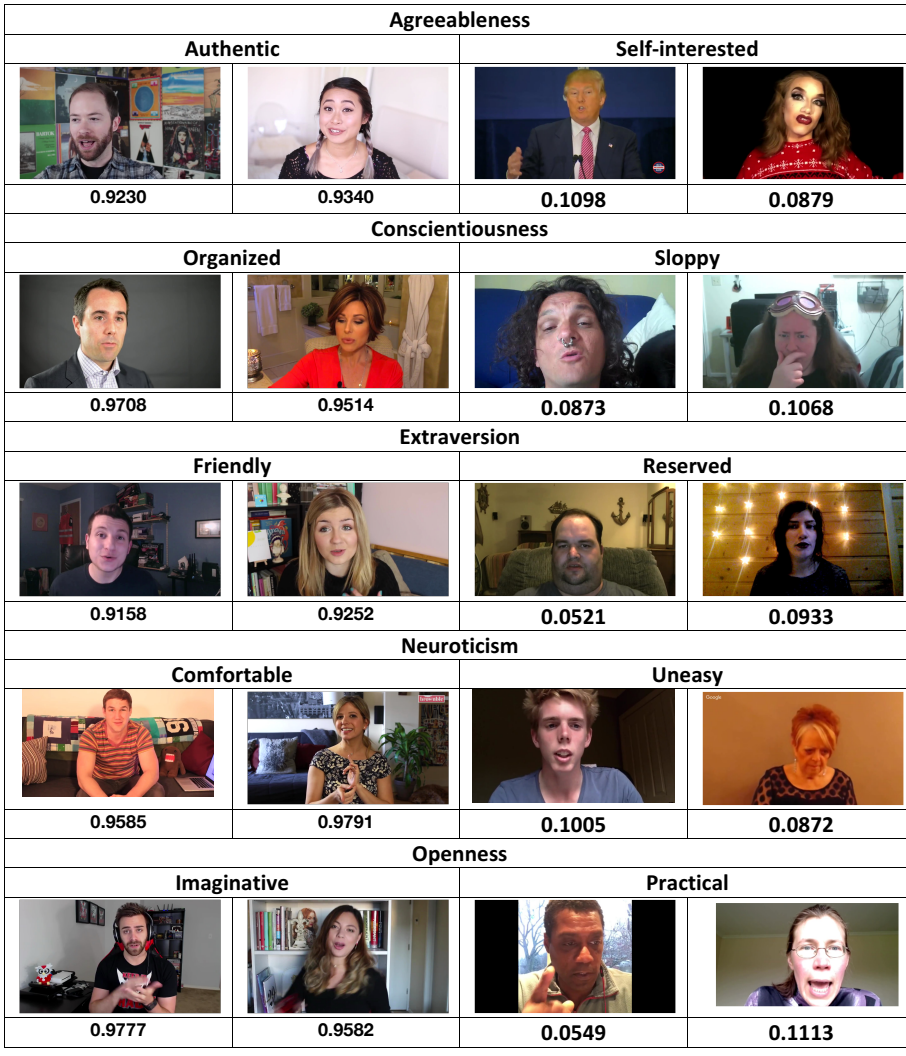

Fig. 3. Screenshot of sample videos voted to clearly perceive the traits, on either end of the spectrum.

3.2 Ground-truth estimation

Obtaining ground truth for Personality Traits can be challenging. Before deciding to use human labeling of videos, we considered using self-administered personality tests on subjects we interviewed. We concluded that such test results are biased and variable. Additionally, performing our own interviews did not allow us to collect massive amounts of data. Therefore, for this dataset, we resorted to use the perception of human subjects visioning the videos. This is a

different task than evaluating real personality traits, but equally useful in the context of human interaction (e.g. job interviews, dating, etc.).

To rapidly obtain a large number of labels, we used Amazon Mechanical Turk (AMT), as is now common in computer vision [32]. Our budget allowed us to get multiple votes per video, in an effort to reduce variance. However, because each worker (aka voter) contributes only a few labels in a large dataset, this raises the problem of bias and the need for calibrating the labels. Biases, which can be traced for example to harshness, prejudices for race, age, or gender, and cultural prejudices, are very hard to measure.

We addressed this problem by using pairwise comparisons. We designed a custom interface (see Figure 2).

Each AMT worker labeled small batches of pairs of videos. To ensure a good coverage and some overlap in the labeling of pairs of videos across workers, we generated pairs with a small-world algorithm [33]. Small-world graphs provide high connectivity, avoid disconnected regions in the graph, have a well distributed edges, and minimum distance between nodes [34].

Cardinal scores were obtained by fitting a BTL model [35]. This is a probabilistic model such that the probability that an object j is judged to have more of an attribute than object i is a sigmoid function of the difference between cardinal scores. Maximum Likelihood estimation was used to fit the model. Deeper details and explanations of the procedure to convert from pairwise scores to cardinal scores are provided in a companion paper [36], where a study is conducted to evaluate how many videos we could label with the constraints of our financial budget. We ended up affording $321,684$ pairs to label $10,000$ videos.

4 Challenge results and methods

In this section we summarize the methods proposed by the teams and provide a detailed description of the winning methods. The teams submitted their code and predictions for the test sets; the source code is available from the challenge website.[7]. Then, we provide a statistical analysis of the results and highlight overall aspects of the competition.

4.1 Summary of methods used

In Table 2 we summarize the various approaches of the teams who participated in the final phase, uploaded their models, and returned a survey about methods we asked them to answer (so-called "fact sheets").

The vast majority of approaches, including the best performing methods, used both audio and video modalities. Most of the teams represented the audio with handcrafted spectral features, a notable exception being the method proposed by team *DCC*, where a residual network [37] was used instead. For the

[7] http://gesture.chalearn.org/2016-looking-at-people-eccv-workshop-challenge/winner_code

	Pretraining	Preprocessing	Modality				Fusion
			Audio		Video		
			R[1]	L[2]	R[1]	L[2]	
NJU-LAMDA	VGG-face	-	logfbank[3]	NN	CNN	CNN	late
evolgen	-	face alignment	spectral	RCNN[10]	RCNN[10]	RCNN[10]	early
DCC	-	-	ResNet	ResNet+FC	ResNet	ResNet+FC	late
ucas	VGG, AlexNet, ResNet	face alignment	spectral	PSLR[4],SVR[5]	CNN(face/scene)	PSLR[4],SVR[5]	late
BU-NKU	VGG-face, FER2013	face alignment	-	-	CNN(face/scene)	KELM[6]	early
pandora	-	face alignment	LLD[8]	Bagged Regressor	CNN(face/scene)	CNN	early
Pilab	-	-	spectral	RF regressor	-	-	-
Kaizoku	-	-	MFCC[9]/CNN	CNN	CNN	CNN	late
ITU-SiMiT	VGG-face, VGG-16	face detection	-	-	CNN(face/scene)	SVR[5]	late

[1] R = Representation [2] L = Learning Strategy [3] logfbank = Logarithm Filterbank Energies [4] PSLR = Partial Least Square Regressor [5] SVR = Support Vector Regression [6] KELM = Kernel Extreme Learning Machine [7] FER = Facial Expression Recognition Dataset [8] LLD = Low Level Descriptor [9] MFCC = Mel Frequency Cepstral Coefficient [10] RCNN = Recurrent Convolutional Neural Networks.

Table 2. Overview of the team methods comparing pretraining (topology and data), preprocessing if performed, representation, learning strategy per modality and fusion.

video modality the dominant approach was to learn the representations through convolutional neural networks [38]. The modalities were late-fused in most methods before being fed to different regression methods like fully connected neural networks or Support Vector Regressors. A notable exception is the method proposed by team *evolgen*, which includes temporal structure by partitioning the video sequences and sequentially feeding the learned audio-video representation to a recurrent Long Short Term Memory layer [39].

Most teams made semantic assumptions about the data by separating face from background. Usually, this was achieved by preprocessing such as face frontalisation. However, it is important to notice that the winning method of team *NJU-LAMDA* does not make any kind of semantic separation of the content.

Finally, a common approach was to use pre-trained deep models fine-tuned on the dataset provided for this challenge. The readers are referred to Table 2 for a synthesis of the main characteristics of the methods that have been submitted to this challenge and to Table 3 for the achieved results. Next, we provide a more detailed description of the three winning methods.

First place: The *NJU-LAMDA* team proposed two separate models for still images and audio, processing multiple frames from the video and employing a two-step late fusion of the frame and audio predictions. For the video modality, it proposed DAN+, an extension to Descriptor Aggregation Networks [40] which applies max and average pooling at two different layers of the CNN, normalizing and concatenating the outputs before feeding them to a fully connected layers. A pretrained VGG-face model [41] is used, replacing the fully-connected layers and fine-tuning the model with the First Impressions dataset. For the audio modality it employs log filter bank (logfbank) features and a single fully-connected layer with sigmoid activations. At test time, a predefined number of frames are fed to the visual network and the predictions averaged. The final visual predictions are averaged again with the output of the audio predictor.

Second place: The *evolgen* team proposed a multimodal LSTM architecture for predicting the personality traits. In order to maintain the temporal structure, the input video sequences are split in six non-overlapping partitions. From each of the partitions the audio representation is extracted using classical spectral features and statistical measurements, forming a 68-dimensional feature vector. The video representation is extracted by randomly selecting a frame from the partition, extracting the face and centering it through face alignment. The preprocessed data is passed to a Recurrent CNN, trained end-to-end, which uses a separate pipeline for audio and video. Each partition frame is processed with convolutional layers, afterwards applying a linear transform to reduce the dimensionality. The audio features of a given partition go through a linear transform and are concatenated with the frame features. The Recurrent layer is sequentially fed with the features extracted from each partition. In this way, the recurrent network captures variations in audio and facial expressions for personality trait prediction.

Third place: The *DCC* team proposed a multimodal personality trait recognition model comprising of two separate auditory and visual streams (deep residual networks, 17 layers each), followed by an audiovisual stream (one fully-connected layer with hyperbolic tangent activation) that is trained end-to-end to predict the big five personality traits. There is no pretraining, but a simple preprocessing is performed where a random frame and crop of the audio are selected as inputs. During test, the whole audio and video sequences are fed into the auditory and visual streams, applying average pooling before being fed to the fully-connected layer.

The approaches of all three winning methods use separate streams for audio and video, applying neural networks for both streams. The first and second places both use some kind of data preprocessing, with the NJU-LAMDA team using logfbank features for the audio and the evolgen team using face cropping and spectral audio features. The second and third methods both use end-to-end training, fusing the audio and video streams with fully-connected layers.

4.2 Statistical analysis of the results

Table 3 lists the results on test data using different metrics. One can also observe very close and competitive results among the top five teams. The results of the top ranking teams are within the error bar.

For comparison, we indicated the results obtained by using the median predictions of all ranked teams. No improvement is gained by using this voting scheme. We also show "random guess", which corresponds to randomly permuting these random predictions.

Accuracy score (normalized)							
Rank	Team Name	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Average
	Median Pred.	0.4188 ± 0.0132	0.3179 ± 0.0148	0.4193 ± 0.0097	0.3892 ± 0.0121	0.3749 ± 0.0116	0.3840 ± 0.0123
1	NJU-LAMDA	0.4215 ± 0.0146	**0.3450 ± 0.0210**	**0.4497 ± 0.0145**	0.4087 ± 0.0171	0.3876 ± 0.0171	**0.4025 ± 0.0169**
2	evolgen	**0.4358 ± 0.0164**	0.3318 ± 0.0178	0.4295 ± 0.0126	0.4069 ± 0.0238	**0.3920 ± 0.0181**	0.3992 ± 0.0178
3	DCC	0.3987 ± 0.0217	0.3236 ± 0.0157	0.4310 ± 0.0153	**0.4091 ± 0.0116**	0.3740 ± 0.0184	0.3873 ± 0.0165
4	ucas	0.4180 ± 0.0129	0.3123 ± 0.0111	0.4128 ± 0.0168	0.3891 ± 0.0134	0.3811 ± 0.0118	0.3827 ± 0.0132
5	BU-NKU	0.4416 ± 0.0188	0.2990 ± 0.0175	0.4324 ± 0.0217	0.3586 ± 0.0156	0.3651 ± 0.0162	0.3794 ± 0.0180
6	pandora	0.3771 ± 0.0150	0.3008 ± 0.0187	0.3770 ± 0.0156	0.3767 ± 0.0211	0.3670 ± 0.0200	0.3597 ± 0.0181
7	Pilab	0.2825 ± 0.0142	0.2464 ± 0.0214	0.2581 ± 0.0124	0.2897 ± 0.0142	0.2977 ± 0.0166	0.2749 ± 0.0158
8	Kaizoku	0.1620 ± 0.0314	0.1848 ± 0.0242	0.2183 ± 0.0299	0.1885 ± 0.0313	0.2353 ± 0.0179	0.1978 ± 0.0270
9	ITU-SiMiT	0.1847 ± 0.0067	0.1953 ± 0.0106	0.1750 ± 0.0082	0.1990 ± 0.0099	0.1915 ± 0.0091	0.1891 ± 0.0089
	Random Guess	0.0697 ± 0.0423	0.1253 ± 0.0456	0.0865 ± 0.0512	0.1039 ± 0.0383	0.0799 ± 0.0490	0.0931 ± 0.0453

R^2 **score**							
Rank	Team Name	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Average
	Median Pred.	0.5048 ± 0.0307	0.2972 ± 0.0361	0.5239 ± 0.0301	0.4565 ± 0.0284	0.4144 ± 0.0353	0.4394 ± 0.0321
1	NJU-LAMDA	0.4808 ± 0.0367	**0.3381 ± 0.0247**	**0.5435 ± 0.0293**	**0.4745 ± 0.0318**	**0.4370 ± 0.0276**	**0.4548 ± 0.0300**
2	evolgen	**0.5151 ± 0.0312**	0.3289 ± 0.0366	0.4883 ± 0.0298	0.4554 ± 0.0287	0.4141 ± 0.0391	0.4404 ± 0.0331
3	DCC	0.4312 ± 0.0405	0.2961 ± 0.0293	0.4781 ± 0.0360	0.4484 ± 0.0322	0.4026 ± 0.0249	0.4113 ± 0.0326
4	ucas	0.4890 ± 0.0394	0.2921 ± 0.0242	0.5195 ± 0.0330	0.4573 ± 0.0369	0.4391 ± 0.0295	0.4394 ± 0.0326
5	BU-NKU	0.5143 ± 0.0318	0.2339 ± 0.0313	0.4866 ± 0.0318	0.3634 ± 0.0325	0.3721 ± 0.0279	0.3941 ± 0.0311
6	pandora	0.4141 ± 0.0316	0.2440 ± 0.0290	0.4020 ± 0.0375	0.3772 ± 0.0406	0.3675 ± 0.0255	0.3610 ± 0.0328
7	Pilab	0.2204 ± 0.0343	0.1208 ± 0.0349	0.1554 ± 0.0342	0.2292 ± 0.0237	0.2266 ± 0.0288	0.1905 ± 0.0312
8	Kaizoku	0.2260 ± 0.0324	0.1098 ± 0.0210	0.2248 ± 0.0326	0.2246 ± 0.0248	0.2269 ± 0.0346	0.2024 ± 0.0291
9	ITU-SiMiT	0.0074 ± 0.0082	0.0020 ± 0.0035	0.0061 ± 0.0059	0.0015 ± 0.0015	0.0033 ± 0.0047	0.0040 ± 0.0048
	Random Guess	0.0024 ± 0.0025	0.0020 ± 0.0023	0.0017 ± 0.0016	0.0019 ± 0.0036	0.0015 ± 0.0026	0.0019 ± 0.0025

AUC score							
Rank	Team Name	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness	Average
	Median Pred.	0.8333 ± 0.0138	0.7625 ± 0.0255	0.8504 ± 0.0196	0.8112 ± 0.0144	0.8179 ± 0.0170	0.8241 ± 0.0181
1	NJU-LAMDA	0.8391 ± 0.0247	0.7634 ± 0.0239	**0.8696 ± 0.0147**	0.8199 ± 0.0147	0.8217 ± 0.0173	0.8227 ± 0.0191
2	evolgen	0.8376 ± 0.0160	**0.7771 ± 0.0210**	0.8492 ± 0.0165	0.8260 ± 0.0184	0.8135 ± 0.0156	0.8207 ± 0.0175
3	DCC	0.8178 ± 0.0187	0.7528 ± 0.0227	0.8579 ± 0.0160	0.8131 ± 0.0239	0.8138 ± 0.0233	0.8111 ± 0.0209
4	ucas	0.8421 ± 0.0193	0.7767 ± 0.0253	0.8569 ± 0.0166	**0.8338 ± 0.0181**	**0.8290 ± 0.0159**	**0.8277 ± 0.0190**
5	BU-KNU	**0.8438 ± 0.0201**	0.7372 ± 0.0279	0.8586 ± 0.0154	0.7854 ± 0.0255	0.7991 ± 0.0172	0.8048 ± 0.0212
6	pandora	0.8097 ± 0.0173	0.7435 ± 0.0239	0.8074 ± 0.0161	0.7987 ± 0.0169	0.8026 ± 0.0150	0.7924 ± 0.0178
7	Pilab	0.7139 ± 0.0277	0.6608 ± 0.0229	0.6870 ± 0.0254	0.7321 ± 0.0199	0.7195 ± 0.0278	0.7026 ± 0.0247
8	Kaizoku	0.7286 ± 0.0198	0.6603 ± 0.0241	0.7393 ± 0.0263	0.7277 ± 0.0179	0.7051 ± 0.0218	0.7122 ± 0.0220
9	ITU-SiMiT	0.4410 ± 0.0255	0.4669 ± 0.0216	0.4778 ± 0.0212	0.4863 ± 0.0193	0.4706 ± 0.0154	0.4685 ± 0.0206
	Random Guess	0.4988 ± 0.0272	0.5129 ± 0.0214	0.5161 ± 0.0255	0.5010 ± 0.0264	0.5193 ± 0.0252	0.5096 ± 0.0252

Table 3. Results of the first round of the Personality Trait challenge. Top: the Accuracy score used to rank the teams (Equation 1). Middle: R^2 score (Equation 2). Bottom: Area under the ROC Curve (AUC) evaluating predictions by turning the problem into a classification problem. The error bars are the standard deviations computed with the bootstrap method. The best results are indicated in bold.

We treated the problem either as a regression problem or as a classification problem:

– **As a regression problem.** The metric that was used in the challenge to rank teams is the mean (normalized) accuracy (Equation 1). We normalized

it in such a way that making constant predictions of the average target values yields a score of 0. The best score is 1. During the challenge we did not normalize the accuracy; however this normalization does not affect the ranking. Normalizing makes the Accuracy more comparable to the R^2 and results are easier to interpret. The results obtained with the R^2 metric (Equation 2) are indeed similar, except that the third and fourth ranking teams are swapped. The advantage of using the Accuracy over the R^2 is that it is less sensitive to outliers.

– **As a classification problem.** The AUC metric (for which random guesses yield a score of 0.5, and exact predictions a score of 1) yields slightly different results. The fourth ranking team performs best according to that metric. Classification is generally an easier problem than regression. We see that classification results are quite good compared to regression results.

For the regression analysis, we graphically represented the Accuracy results (the official ranking score) as a box plot (Figure 4) showing the distribution of scores for each trait and the overall accuracy. For the classification analysis, we show ROC curves in Figure 5. In both cases Agreeableness seems significantly harder to predict than other traits, while Conscientiousness is the easiest (albeit with a large variance). We also see that all top ranking teams have similar ROC curves.

An analysis of the correlation between the five personality traits for both the ground truth and the median predictions (Figure 6) shows some correlation between labels, particularly the group Extraversion, Neurotism, and Openness. This remains true for the team' predictions; Agreeableness is also significantly correlated to that group. For the predictions the correlation between any given pair of traits is 25-35% higher for the team' predictions than for the ground truth. Nothing in the challenge setting encourages methods to "orthogonalize" decisions about traits, hence the predictors devised by the teams make joint predictions of all five personality traits and may easily learn correlations between traits.

In Figure 7, we also investigated the quality of the predictions by producing scatter plots of the predictions vs. the ground truth. We show an example for the trait Extraversion. On the x-axis coordinate is ground truth and on the y-axis the median prediction of all the teams. We linearly regressed the predictions to the ground truth. The first diagonal corresponds to ideal predictions. Similar plots are obtained for all traits and all teams. As can be seen, the points do not gather around the first diagonal and the two lines have different slopes. We interpret this as follows: there are two sources of error, a systematic error corresponding to a bias in prediction towards the average ground truth value, and a random error. Essentially the models are under-fitting (they are biased towards the constant prediction).

5 Discussion and Future Work

This paper has described the main characteristics of the ChaLearn Looking at People 2016 Challenge which included the first round competition on First Im-

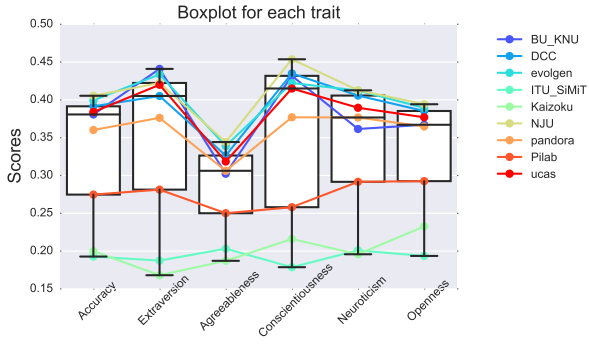

Fig. 4. Distribution of final scores for each trait and performance of the individual teams. We see that "Agreeableness" is consistently harder to predict by the top ranking teams.

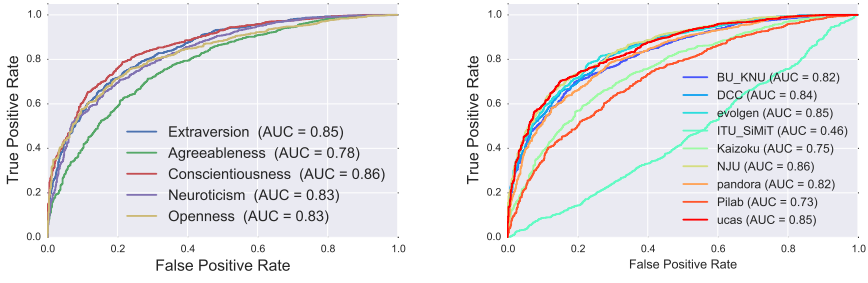

Fig. 5. Receiver Operating Characteristic curve of the median prediction of each trait, the median taken over all ranked teams predictions (left) and averaged over all traits for each team (right)

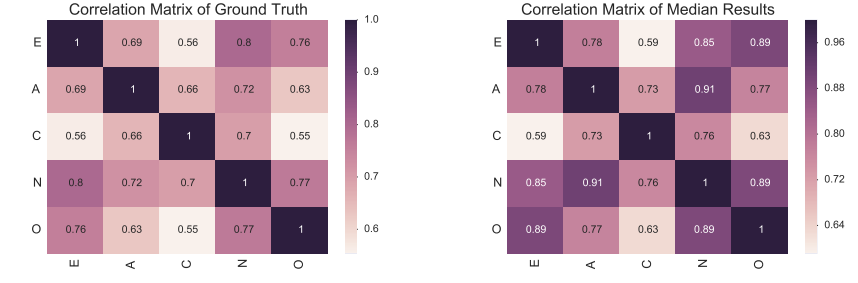

Fig. 6. Correlation matrices. Correlation for all videos between ground truth labels (left), and between the median predictions of the teams (right).

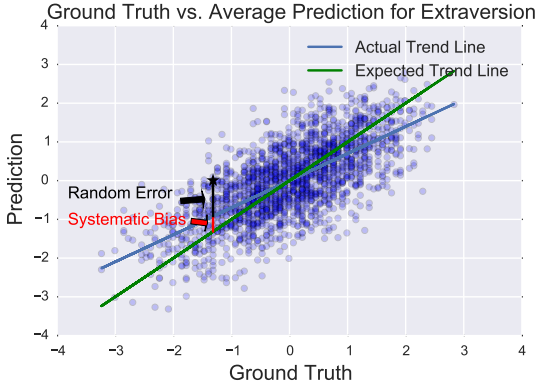

Fig. 7. Ground Truth vs. Average Prediction for Extraversion. Each dot represents a video. The average is taken over all final submissions.

pressions. A large dataset was designed with manual selection of videos, AMT pairwise video annotation to alleviate labeling bias, and reconstruction of cardinal ratings by fitting a BLT model. The data were made publicly available to the participants for a fair and reproducible comparison in the performance results. Analyzing the methods used by 9 teams that participated in the final evaluation and uploaded their models (out of a total of 84 participants), several conclusions can be drawn:

- There was a lot of emulation during the challenge and the final results are close to one another even though the methods are quite diverse.
- Feature learning (via deep learning methods) dominates the analysis, but pretrained models are widely used (perhaps due to the limited amount of available training data).
- Late fusion is generally applied, though additional layers fusing higher level representations from separate video and audio streams are often used.
- Video is usually analyzed at a per-frame basis, pooling the video features or fusing the predictions. The second place winner is an exception, using an LSTM to integrate the temporal information.
- Many teams used contextual cues and extracted faces, but some top ranking teams did not.

Even though performances are already quite good, from the above analysis it is still difficult to ensure the achievement of human level performance. Since there is a wide variety of complementary approaches, to push participants to improve their performances by joining forces, we are organizing a first coopetition (combination of competition and collaboration) for ICPR 2016. In this first edition of coopetition, we reward people for sharing their code by combining the traditional accuracy score with the number of downloads of their code. With this setting, the methods are not only evaluated by the organizers, but also by the other participants.

We are preparing a more sophisticated coopetition that will include more interactive characteristics, such as the possibility for teams to share modules of their overall system. To that end, we will exploit CodaLab worksheets (`http://worksheets.codalab.org`), a new feature resembling iPython notebooks, which allow user to share code (not limited to Python) intermixed with text, data, and results. We are working on integrating into CodaLab worksheets a system of reward mechanisms suitable to keep challenge participants engaged.

As mentioned in the introduction, the First Impressions challenge is part of a larger project on Speed Interviews for job hiring purposes. Some of our next steps will consist in including more modalities that can be used together with audio-RGB data as part of a multimedia CV. Examples of such modalities include handwritten letters and/or traditional CVs.

Acknowledgments

We are very grateful for the funding provided by Microsoft Research without which this work would not have been possible, and for the kind support provided by Evelyne Viegas, director of the Microsoft AI Outreach project, since the inception of this project. We also thank the Microsoft CodaLab support team for their responsiveness and particularly Flavio Zhingri. We sincerely thank all the teams who participated in ChaLearn LAP 2016 for their interest and for having contributed to improve the challenge with their comments and suggestions. Special thanks to Marc Pomar for preparing the annotation interface for Amazon Mechanical Turk (AMT). The researchers who joined the program committee and reviewed for the ChaLearn LAP 2016 workshop are gratefully acknowledged. We are very grateful to our challenge sponsors: Facebook, NVIDIA and INAOE, whose support was critical for awarding prizes and travel grants. This work was also partially supported by Spanish projects TIN2015-66951-C2-2-R, TIN2012-39051, and TIN2013-43478-P and received additional support from the Laboratoire d'Informatique Fondamentale (LIF, UMR CNRS 7279) of the University of Aix Marseille, France, via the LabeX Archimede program, the Laboratoire de Recherche en Informatique of Paris Sud University, INRIA-Saclay and the Paris-Saclay Center for Data Science (CDS). We thank our colleagues from the speed interview project for their contribution, and particularly Stephane Ayache, Cecile Capponi, Pascale Gerbail, Sonia Shah, Michele Sebag, Carlos Andujar, Jeffrey Cohn, and Erick Watson.

References

1. Ambady, N., Bernieri, F.J., Richeson, J.A.: Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. Adv. Exp. Soc. Psycho. **32** (2000) 201–271
2. Berry, D.S.: Taking people at face value: Evidence for the kernel of truth hypothesis. Social Cognition **8**(4) (1990) 343
3. Hassin, R., Trope, Y.: Facing faces: studies on the cognitive aspects of physiognomy. JPSP **78**(5) (2000) 837
4. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol. Bull. **111**(2) (1992) 256

5. Willis, J., Todorov, A.: First impressions making up your mind after a 100-ms exposure to a face. PSS **17**(7) (2006) 592–598
6. Ozer, D.J., Benet-Martinez, V.: Personality and the prediction of consequential outcomes. Annu. Rev. Psychol. **57** (2006) 401–421
7. Roberts, B.W., Kuncel, N.R., Shiner, R., Caspi, A., Goldberg, L.R.: The power of personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. PPS **2**(4) (2007) 313–345
8. Posthuma, R.A., Morgeson, F.P., Campion, M.A.: Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. Pers. Psychol. **55**(1) (2002) 1–81
9. Huffcutt, A.I., Conway, J.M., Roth, P.L., Stone, N.J.: Identification and meta-analytic assessment of psychological constructs measured in employment interviews. JAP **86**(5) (2001) 897
10. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. TAC **5**(3) (2014) 273–291
11. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. JAIR **30** (2007) 457–500
12. Ivanov, A.V., Riccardi, G., Sporka, A.J., Franc, J.: Recognition of personality traits from human spoken conversations. In: INTERSPEECH. (2011) 1549–1552
13. Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., Zancanaro, M.: Multimodal recognition of personality traits in social interactions. In: ICMI, ACM (2008) 53–60
14. Batrinca, L.M., Mana, N., Lepri, B., Pianesi, F., Sebe, N.: Please, tell me about yourself: automatic personality assessment using short self-presentations. In: ICMI, ACM (2011) 255–262
15. Batrinca, L., Lepri, B., Mana, N., Pianesi, F.: Multimodal recognition of personality traits in human-computer collaborative tasks. In: ICMI, New York, NY, USA, ACM (2012) 39–46
16. Mana, N., Lepri, B., Chippendale, P., Cappelletti, A., Pianesi, F., Svaizer, P., Zancanaro, M.: Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In: ICMI Workshop, ACM (2007) 9–14
17. Lepri, B., Subramanian, R., Kalimeri, K., Staiano, J., Pianesi, F., Sebe, N.: Connecting meeting behavior with extraversion - a systematic study. TAC **3**(4) (2012) 443–455
18. Polzehl, T., Moller, S., Metze, F.: Automatically assessing personality from speech. In: ICSC, IEEE (2010) 134–140
19. Biel, J.I., Teijeiro-Mosquera, L., Gatica-Perez, D.: Facetube: predicting personality from facial expressions of emotion in online conversational video. In: ICMI, ACM (2012) 53–56
20. Sanchez-Cortes, D., Biel, J.I., Kumano, S., Yamato, J., Otsuka, K., Gatica-Perez, D.: Inferring mood in ubiquitous conversational video. In: MUM, ACM (2013) 22
21. Abadi, M.K., Correa, J.A.M., Wache, J., Yang, H., Patras, I., Sebe, N.: Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos. FG (2015)
22. Ponce-López, V., Escalera, S., Baró, X.: Multi-modal social signal analysis for predicting agreement in conversation settings. In: ICMI. ICMI '13, New York, NY, USA, ACM (2013) 495–502

23. Ponce-López, V., Escalera, S., Pérez, M., Janés, O., Baró, X.: Non-verbal communication analysis in victim-offender mediations. PRL **67, Part 1** (2015) 19 – 27 Cognitive Systems for Knowledge Discovery.
24. Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H.J., Hamner, B.: Results and analysis of the chalearn gesture challenge 2012. In: WDIA. (2012) 186–204
25. Guyon, I., Athitsos, V., Jangyodsuk, P., Escalante, H.J.: The chalearn gesture dataset (cgd 2011). Mach. Vis. Appl. **25**(8) (2014) 1929–1951
26. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.J.: Multi-modal gesture recognition challenge 2013: Dataset and results. ICMI Workshop (2013) 445–452
27. Escalera, S., Baró, X., Gonzàlez, J., Bautista, M.A., Madadi, M., Reyes, M., Ponce-López, V., Escalante, H.J., Shotton, J., Guyon, I.: Chalearn looking at people challenge 2014: Dataset and results. ECCV Workshop **1** (2014) 459–473
28. Baró, X., Gonzalez, J., Fabian, J., Bautista, M.A., Oliu, M., Escalante, H.J., Guyon, I., Escalera, S.: Chalearn looking at people 2015 challenges: Action spotting and cultural event recognition. In: CVPR Workshop, IEEE (2015) 1–9
29. Escalera, S., Fabian, J., Pardo, P., Baró, X., Gonzalez, J., Escalante, H.J., Misevic, D., Steiner, U., Guyon, I.: Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In: ICCV Workshop. (2015) 1–9
30. Escalera, S., Torres, M., Martinez, B., Baró, X., Escalante, H.J., et al.: Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In: CVPR Workshop. (2016)
31. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV **57**(2) (2004) 137–154
32. Lang, A., Rio-Ross, J.: Using amazon mechanical turk to transcribe historical handwrit- ten documents. (2011)
33. Watts, D.J., Strogatz, S.H.: Collective dynamics of'small-world'networks. Nature **393**(6684) (1998) 409–10
34. Humphries, M., Gurney, K., Prescott, T.: The brainstem reticular formation is a small-world, not scale-free, network. PRSL-B **273**(1585) (2006) 503–511
35. Bradley, R., Terry, M.: Rank analysis of incomplete block designs: The method of paired comparisons. Biometrika **39** (1952) 324–345
36. Chen, B., Escalera, S., Guyon, I., Ponce-López, V., Shah, N., Oliu, M.: Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits. (Submitted to ECCV LAP challenge workshop, 2016)
37. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv (2015)
38. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
39. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8) (1997) 1735–1780
40. Wei, X.S., Luo, J.H., Wu, J.: Selective convolutional descriptor aggregation for fine-grained image retrieval. arXiv (2016)
41. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. Volume 1. (2015) 6