ChaLearn Looking at People: First Impressions
Apparent Personality Analysis Challenge

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, Sergio Escalera

Ciprian A. Corneanu

ECCV, 2016
Outline

1. Introduction
   - Context of the Challenge
   - Apparent Personality Analysis Challenge

2. Apparent Personality Analysis Challenge
   - Dataset and Evaluation
   - Results
   - Winning Methods

3. Conclusions
Outline

1 Introduction
   • Context of the Challenge
   • Apparent Personality Analysis Challenge

2 Apparent Personality Analysis Challenge
   • Dataset and Evaluation
   • Results
   • Winning Methods

3 Conclusions
ChaLearn is a non-profit organisation focusing on challenge organization in Machine Learning.

ChaLearn Looking at People (ChaLearn LAP) is a branch of ChaLearn focusing on Human Behaviour Analysis.

http://gesture.chalearn.org
http://sergioescalera.com/hupba/
Previous Activities of ChaLearn LAP

Associated special issues:
- 2011-2014, JMLR special issue
- 2016, International Journal of Computer Vision: Looking at People
Upcoming Activities

Challenges and Workshops:

- ICPR, 2016 - Challenge and Workshop on apparent personality analysis: First Impressions (Second Round)
- CVPR, 2017: Challenge and Workshop on explainable computer vision and pattern recognition in first impressions.
- IJCNN, 2017: Challenge and Workshop on explainable computer vision and pattern recognition in first impressions.

Special Issues:

- 2016-2017, Pattern Recognition, guest editor special issue: Articulated Motion and Deformable Objects.  **November, 15th, 2016**

More info at:  [http://gesture.chalearn.org](http://gesture.chalearn.org)
Outline

1 Introduction
   • Context of the Challenge
   • Apparent Personality Analysis Challenge

2 Apparent Personality Analysis Challenge
   • Dataset and Evaluation
   • Results
   • Winning Methods

3 Conclusions
Apparent Personality Analysis Challenge

- **Objective**
  - The goal of the competition was to automatically evaluate apparent personality traits from videos of subjects speaking in front of a camera. The traits follow the Big Five model: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism.

- **Data**
  - A novel data set consisting of 10,000 short clips from YouTube videos has been made publicly available.
  - Amazon Mechanical Turk (AMT) for building ground truth.
  - To alleviate calibration problems, we used pairwise comparisons between videos, and variable levels were reconstructed by fitting a Bradley-Terry-Luce model with maximum likelihood.

- **Evaluation**
  - Prediction submission and scoring through CodaLab, an open-source platform. [http://codalab.org](http://codalab.org)

- **Participation**
  - The competition attracted, over a period of 2 months (May 15th - July 15th), 86 participants grouped in several teams. 9 teams entered the final phase.
Outline

1. Introduction
   - Context of the Challenge
   - Apparent Personality Analysis Challenge

2. Apparent Personality Analysis Challenge
   - Dataset and Evaluation
   - Results
   - Winning Methods

3. Conclusions
Data Collection

Figure: Data collection and processing.

Final Data Corpus:

- 10,000 clips of 15 secs each.
- 41.6 hours of total footage. Aprox. 4.5 millions frames.
- 321,684 pairwise video annotations from AMT.
### Data Overview

#### Video data preparation

<table>
<thead>
<tr>
<th>Data</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downloaded videos</td>
<td>13,951* (HD 720p @ 30 FPS)</td>
</tr>
<tr>
<td>Remaining videos</td>
<td>8,581**</td>
</tr>
<tr>
<td>Sampled videos per channel</td>
<td>3 (at most)</td>
</tr>
<tr>
<td>Sampled clips per video</td>
<td>6 (at most)</td>
</tr>
<tr>
<td>Clip length</td>
<td>15 seconds</td>
</tr>
<tr>
<td>Candidate clips</td>
<td>32,139†</td>
</tr>
</tbody>
</table>

#### Final set of clips

<table>
<thead>
<tr>
<th>Data</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total duration of clips</td>
<td>41.6 hours (4.5M frames)</td>
</tr>
<tr>
<td>Unique channels (originating)</td>
<td>2,764; {1 : 2,584, 2 : 161, 3 : 19}§</td>
</tr>
<tr>
<td>Mean no. clips per video</td>
<td>3.27</td>
</tr>
<tr>
<td>Duration of originating videos</td>
<td>608.7 hours</td>
</tr>
<tr>
<td>Total no. views of originating videos</td>
<td>More than 115M; {0-100 : 27.64%, 100-1K : 34.15%, 1K-10K : 22.68%, 10K-100K : 11.44%, &gt;100K : 4.08%}∥</td>
</tr>
<tr>
<td>Originating videos' avg. rating</td>
<td>4.6/5.0; {1 : 8, 2 : 11, 3 : 43, 4 : 1340, 5 : 1,395}</td>
</tr>
<tr>
<td>Originating videos' keywords (top 20)</td>
<td>'Q&amp;A', 'q&amp;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'</td>
</tr>
</tbody>
</table>

---

§ 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:** Video data preparation and final data set statistics.
Ground Truth Estimation

- We used Amazon Mechanical Turk (AMT).
- For better calibration 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.
- Cardinal scores were obtained by fitting a Bradley-Terry-Luce (BTL) model.

Figure: Web interface used for pairwise labelling of the data.
### Ground Truth Estimation

<table>
<thead>
<tr>
<th>Agreeableness</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentic</td>
<td>0.9230</td>
</tr>
<tr>
<td>Self-interested</td>
<td>0.9340</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td></td>
</tr>
<tr>
<td>Organized</td>
<td>0.9708</td>
</tr>
<tr>
<td>Sloppy</td>
<td>0.9514</td>
</tr>
<tr>
<td>Extraversion</td>
<td></td>
</tr>
<tr>
<td>Friendly</td>
<td>0.9158</td>
</tr>
<tr>
<td>Reserved</td>
<td>0.9252</td>
</tr>
<tr>
<td>Neuroticism</td>
<td></td>
</tr>
<tr>
<td>Comfortable</td>
<td>0.9585</td>
</tr>
<tr>
<td>Uneasy</td>
<td>0.9791</td>
</tr>
<tr>
<td>Openness</td>
<td></td>
</tr>
<tr>
<td>Imaginative</td>
<td>0.9777</td>
</tr>
<tr>
<td>Practical</td>
<td>0.9582</td>
</tr>
</tbody>
</table>

**Figure:** Example videos with clearly perceived traits.

Outline

1 Introduction
   • Context of the Challenge
   • Apparent Personality Analysis Challenge

2 Apparent Personality Analysis Challenge
   • Dataset and Evaluation
   • Results
   • Winning Methods

3 Conclusions
Final Results

Evaluation Metric: mean accuracy over all traits and videos. Accuracy for each trait is defined as

\[ A = 1 - \frac{1}{N_t} \sum_{i=1}^{N_t} \left| g_i - p_i \right| / \sum_{i=1}^{N_t} \left| g_i - \bar{g} \right| \]  

(1)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Team Name</th>
<th>Extraversion</th>
<th>Agreeableness</th>
<th>Conscientiousness</th>
<th>Neuroticism</th>
<th>Openness</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>NJU-LAMDA</td>
<td>0.4215 ± 0.0146</td>
<td>0.3450 ± 0.0210</td>
<td>0.4497 ± 0.0145</td>
<td>0.4087 ± 0.0171</td>
<td>0.3876 ± 0.0171</td>
<td>0.4025 ± 0.0169</td>
</tr>
<tr>
<td>0.5</td>
<td>evolgen</td>
<td>0.4358 ± 0.0164</td>
<td>0.3318 ± 0.0178</td>
<td>0.4295 ± 0.0126</td>
<td>0.4069 ± 0.0238</td>
<td>0.3920 ± 0.0181</td>
<td>0.3992 ± 0.0178</td>
</tr>
<tr>
<td>0.5</td>
<td>DCC</td>
<td>0.3987 ± 0.0217</td>
<td>0.3236 ± 0.0157</td>
<td>0.4310 ± 0.0153</td>
<td>0.4091 ± 0.0116</td>
<td>0.3740 ± 0.0184</td>
<td>0.3873 ± 0.0165</td>
</tr>
<tr>
<td>0.5</td>
<td>ucas</td>
<td>0.4180 ± 0.0129</td>
<td>0.3123 ± 0.0111</td>
<td>0.4128 ± 0.0168</td>
<td>0.3891 ± 0.0134</td>
<td>0.3811 ± 0.0118</td>
<td>0.3827 ± 0.0132</td>
</tr>
<tr>
<td>0.5</td>
<td>BU-NKU</td>
<td>0.4416 ± 0.0188</td>
<td>0.2990 ± 0.0175</td>
<td>0.4324 ± 0.0217</td>
<td>0.3586 ± 0.0156</td>
<td>0.3651 ± 0.0162</td>
<td>0.3794 ± 0.0180</td>
</tr>
<tr>
<td>0.5</td>
<td>pandora</td>
<td>0.3771 ± 0.0150</td>
<td>0.3008 ± 0.0187</td>
<td>0.3770 ± 0.0156</td>
<td>0.3767 ± 0.0211</td>
<td>0.3670 ± 0.0200</td>
<td>0.3597 ± 0.0181</td>
</tr>
<tr>
<td>0.5</td>
<td>Pilab</td>
<td>0.2825 ± 0.0142</td>
<td>0.2464 ± 0.0214</td>
<td>0.2581 ± 0.0124</td>
<td>0.2897 ± 0.0142</td>
<td>0.2977 ± 0.0166</td>
<td>0.2749 ± 0.0158</td>
</tr>
<tr>
<td>0.5</td>
<td>Kaizoku</td>
<td>0.1620 ± 0.0314</td>
<td>0.1848 ± 0.0242</td>
<td>0.2183 ± 0.0299</td>
<td>0.1885 ± 0.0313</td>
<td>0.2353 ± 0.0179</td>
<td>0.1978 ± 0.0270</td>
</tr>
<tr>
<td>0.5</td>
<td>ITU-SiMiT</td>
<td>0.1847 ± 0.0067</td>
<td>0.1953 ± 0.0106</td>
<td>0.1750 ± 0.0082</td>
<td>0.1990 ± 0.0099</td>
<td>0.1915 ± 0.0091</td>
<td>0.1891 ± 0.0089</td>
</tr>
</tbody>
</table>

\[^1\] \( p_i \) are the predicted scores, \( g_i \) are the ground truth scores, with the sum running over the \( N_t \) test videos, and \( \bar{g} \) is the average ground truth score over all videos.
1. Introduction
   - Context of the Challenge
   - Apparent Personality Analysis Challenge

2. Apparent Personality Analysis Challenge
   - Dataset and Evaluation
   - Results
   - Winning Methods

3. Conclusions
<table>
<thead>
<tr>
<th></th>
<th>Pretraining</th>
<th>Preprocessing</th>
<th>Audio R¹</th>
<th>Preprocessing</th>
<th>Video R¹</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NJU-LAMDA</td>
<td>VGG-face</td>
<td>-</td>
<td>logfbank³</td>
<td>NN</td>
<td>CNN</td>
<td>CNN</td>
</tr>
<tr>
<td>evolgen</td>
<td>-</td>
<td>face alignment</td>
<td>spectral</td>
<td>RCNN¹⁰</td>
<td>RCNN¹⁰</td>
<td>RCNN¹⁰</td>
</tr>
<tr>
<td>DCC</td>
<td>-</td>
<td>-</td>
<td>ResNet</td>
<td>ResNet+FC</td>
<td>ResNet+FC</td>
<td>early</td>
</tr>
<tr>
<td>ucas</td>
<td>VGG, AlexNet, ResNet</td>
<td>face alignment</td>
<td>spectral</td>
<td>PSLR⁴, SVR⁵</td>
<td>CNN(face/scene)</td>
<td>PSLR⁴, SVR⁵</td>
</tr>
<tr>
<td>BU-NKU</td>
<td>VGG-face, FER2013</td>
<td>face alignment</td>
<td>-</td>
<td>-</td>
<td>CNN(face/scene)</td>
<td>KELM⁶</td>
</tr>
<tr>
<td>pandora</td>
<td>-</td>
<td>face alignment</td>
<td>LLD⁸</td>
<td>Bagged Regressor</td>
<td>CNN(face/scene)</td>
<td>CNN</td>
</tr>
<tr>
<td>Pilab</td>
<td>-</td>
<td>-</td>
<td>spectral</td>
<td>RF regressor</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kaizoku</td>
<td>-</td>
<td>-</td>
<td>MFCC⁹/CNN</td>
<td>CNN</td>
<td>CNN</td>
<td>CNN</td>
</tr>
<tr>
<td>ITU-SiMiT</td>
<td>VGG-face, VGG-16</td>
<td>face detection</td>
<td>-</td>
<td>-</td>
<td>CNN(face/scene)</td>
<td>SVR⁵</td>
</tr>
</tbody>
</table>

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

- **Pretraining:** The first place uses external data to pretrain the model and fine tunes on the First Impression data.
- **Preprocessing:** All teams use some kind of basic preprocessing (frame pooling). The evolgen team extracts the face from the video stream.
- **Multimodality:**
  - The approaches of all three winning methods are multimodal, use separate streams for audio and video and apply neural networks for both streams.
  - The second and third methods both use end-to-end training, fusing the audio and video streams with fully-connected layers.
- **Dynamic analysis:** First and third team perform a per frame analysis. Second place uses RCNN (LSTM) to capture sequential information.

### Table: Overview of the winning methods.

<table>
<thead>
<tr>
<th></th>
<th>Pretraining</th>
<th>Preprocessing</th>
<th>Modality</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>NJU-LAMDA</td>
<td>VGG-face</td>
<td>-</td>
<td>logfbank</td>
<td>late</td>
</tr>
<tr>
<td>evolgen</td>
<td>-</td>
<td>face alignment</td>
<td>spectral</td>
<td>early</td>
</tr>
<tr>
<td>DCC</td>
<td>-</td>
<td>-</td>
<td>ResNet</td>
<td>early</td>
</tr>
</tbody>
</table>

Víctor Ponce-López, Baiyu Chen, Marc Oliu, ChaLearn Looking at People: First Impression ECCV, 2016 18 / 20
Conclusions

Achievements

- Large dataset published with 10,000 videos labelled with personality traits following the Big-Five personality model.
- Results demonstrate feasibility of automatic apparent personality detection from video/audio self presentations.

Participation

- 2 months of competition, 86 participants, 9 teams submitted final results.

Winning Methods

- Feature learning (via deep learning methods) dominates the analysis, but pretrained models are widely 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.
Awards

- **First Place**: **Team NJU_LAMDA**: C.L. Zhang, H. Zhang, X.S. Wei, J.Wu. *Prize: 1,500$ and travel award.*

- **Second Place**: **Team evolgen**: A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, A. Mittal. *Prize: 1,000$ and travel award.*

- **Third Place**: **Team DCC**: Y. Güçlü, U. Güçlü, M. van Gerven, R. van Lier. *Prize: 500$ and travel award.*