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

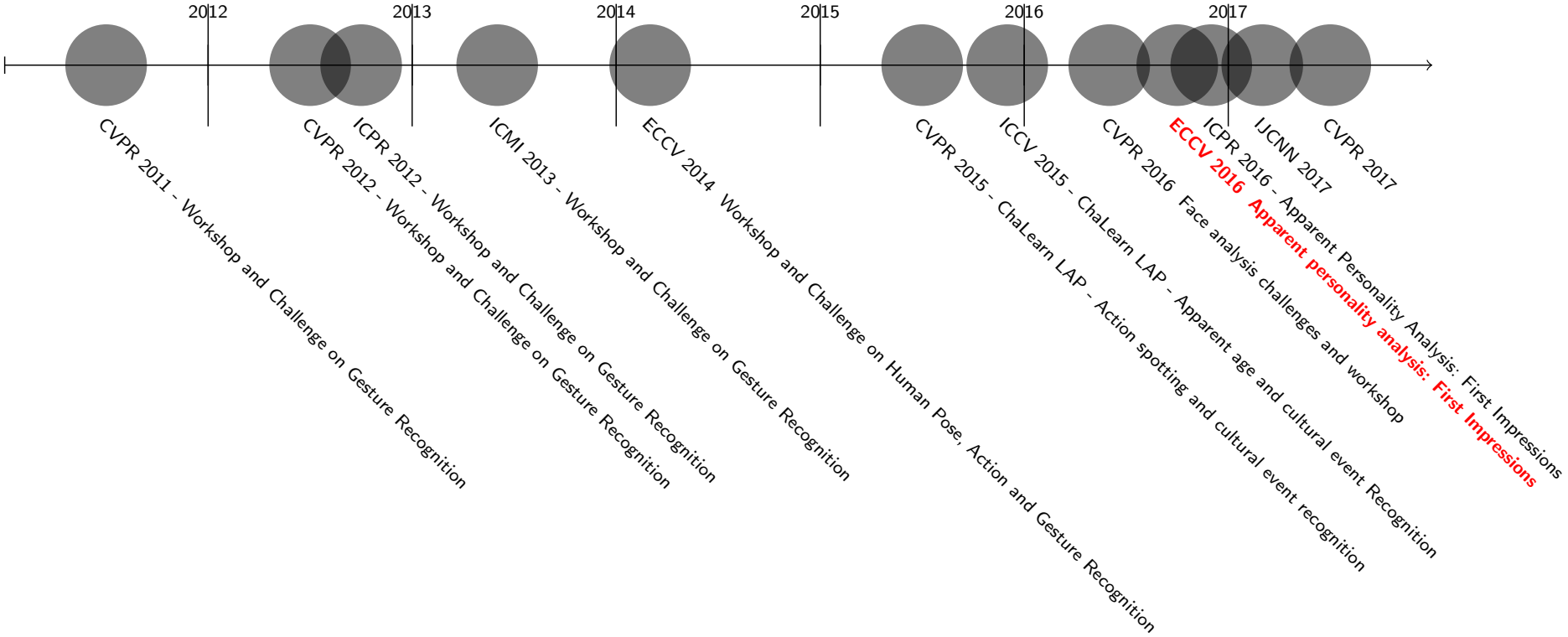
- ChaLearn is a non-profit organisation focusing on challenge organization in Machine Learning.
- ChaLearn Looking at People (ChaLearn LAP) is a branch of ChaLearn focusing on Human Behaviour Analysis.

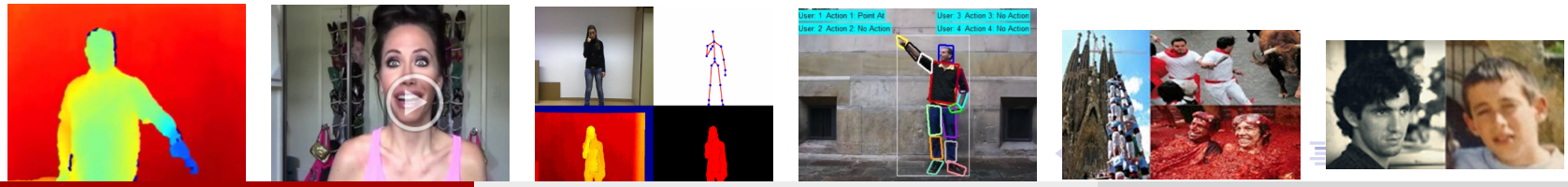
http://gesture.chalearn.org
http://sergioescalera.com/hupba/

Previous Activities of ChaLearn LAP

Associated special issues:

- 2011-2014, JMLR special issue
- 2016, International Journal of Computer Vision: Looking at People
- 2014, 2016 IEEE Transactions on Pattern Analysis and Machine Intelligence: Multi-modal Human Pose Recovery and Behavior Analysis M2HuPBA





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.
  - FG, 2017: Challenge and Workshop on Face Analysis: Emotion Recognition and Head Pose Estimation.
- **Special Issues**:
  - 2016-2017, IEEE Transactions on Affective Computing, guest editor special issue: Artificial Personality Analysis. **November, 15th, 2016**
  - 2016-2017, Pattern Recognition, guest editor special issue: Articulated Motion and Deformable Objects. **November, 15th, 2016**
  - 2017, IEEE Transactions on Pattern Analysis and Machine Intelligence, guest editor special issue: The Computational Face. **February, 15th, 2017**

More info at: 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**
- **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

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: **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.

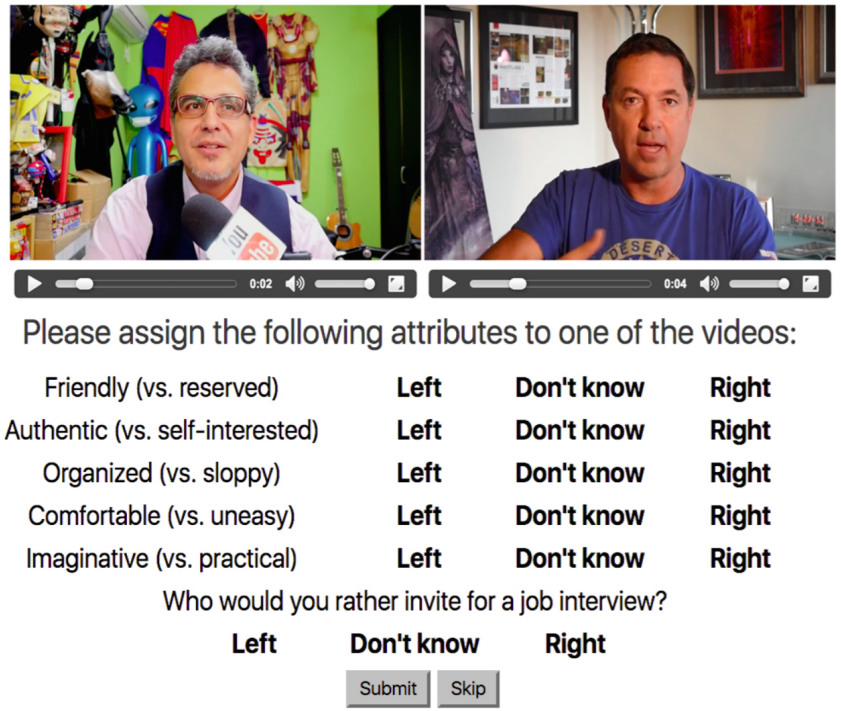


Figure: Web interface used for pairwise labelling of the data.

Ground Truth Estimation

Figure: Example videos with clearly perceived traits.

B. Chen, S. Escalera, I. Guyon, V. Ponce - Lopez, N. Shah, M. Oliu. Overcoming Calibration Problems in Pattern Labeling with Pairwise Ratings: Application to Personality Traits. Proceedings of the European Conference in Computer Vision (ECCV), Workshop on Apparent Personality Analysis, 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

Final Results

Evaluation Metric: mean accuracy over all traits and videos. Accuracy for each trait is defined as [1]:

$$A = 1 - \frac{1}{N_t} \sum_{i=1}^{Nt} |g_i - p_i| / \sum_{i=1}^{Nt} |g_i - \overline{g}| \quad (1)$$

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

[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 $\overline{g}$ is the average ground truth score over all videos

Outline

1. Introduction
   - Context of the Challenge
   - Apparent Personality Analysis Challenge
2. Apparent Personality Analysis Challenge
   - Dataset and Evaluation
   - Results
   - Winning Methods
3. Conclusions

Participant Methods

	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	early
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: **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 preprocesing (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.

	Pretraining	**Preprocessing**	**Modality**				**Fusion**
			Audio		**Video**		
			R[2]	**L**[3]	**R**	**L**	
NJU-LAMDA	VGG-face	-	logfbank	NN	CNN	CNN	late
evolgen	-	face alignment	spectral	RCNN	RCNN	RCNN	early
DCC	-	-	ResNet	ResNet+FC	ResNet	ResNet+FC	early

Table: **Overview of the winning methods.**

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. Güçlütü, U. Güçlürk, M. van Gerven, R. van Lier. *Prize: 500$ and travel award.*