Fact sheet

Explainable Computer Vision Workshop
and
Job Candidate Screening Coopetition

Decision Tree based Mapping of Predicted Apparent Personality for Job Candidate Invitation

08/04/2017

1. Team details

1.1 **Team name:** BU-NKU
1.2 **Team leader name:** Heysem Kaya (Namık Kemal University)
1.3 **Team leader address, phone number and email:**
NKÜ Çorlu Mühendislik Fakültesi Silahtarağa Mahallesi Üniversite 1. Sokak No:13 59860 Çorlu / Tekirdağ TURKEY, +90 282 250 2346, kaya.heysem@gmail.com

1.4 **Rest of the team members:** Furkan Gürpınar (Boğaziçi University), Albert Ali Salah (advisor, Boğaziçi University)
1.5 **Team website URL (if any):**
1.6 **Affiliation:** Namık Kemal University & Boğaziçi University

2. Contribution details

2.1 **Title of the contribution:** Decision Tree based Mapping of Predicted Apparent Personality for Job Candidate Invitation
2.2 **General method description:**

- **Overview of the proposed approach:** We use the predictions of the system proposed for the first stage of this coopetition after binarization (based on their trainin set mean statistics) and learn a decision tree to map them to the binarized “invite-for-interview” variable. This way the model in which the decision made is clear/understandable and interpretable.

The overview of the system proposed in the quantitative stage can be summarized as follows. Here, we use four feature types from our ICPR 2016 Challenge system: LGBPTOP_face,

DCNN_face (VGG face DCNN model [1] fine tuned on FER-2013 [2] emotion corpus), IS13 (a standard set of acoustic features extracted using openSMILE tool with INTERSPEECH 2013 configuration file) and DCNN_Scene (extracted from the first frame of the video with VD19 pretrained DCNN [3]). The architecture can be summarized as:

RF(FF(LGBPTOP_face, DCNN_face), FF(IS13, DCNN_Scene))

where the FF(a,b) represents feature level fusion of a and b regressed using Extreme Learning Machine and Random Forest (RF) takes the predictions from the first level regressors (5 traits + 1 JCS score x 2 = 12 features) to predict each of six target variables.

Additionally, we annotate 4000 videos based on the first frames, so as to cast an automatic prediction for the gender of the person and use it in our explanation. To train a gender classifier we use IS13 (audio) and DCNN_face (video) features and fuse their predictions scores with equal weight. This gives a valdiation set classification accuracy of 99.33%.

- **The proposed method uses / takes advantage of personality traits?**

Yes, in both stages. First, in the final prediction of six target variables (OCEAN + interview) we stack the predictions from the two regressors using a Random Forest. Then, the binarized values of these predictions are used to map OCEAN traits to the “invite-for-interview” variable.

- **Coopetition: did you use any code or idea from other participants (shared in previous stage of the challenge)?**

  In our final system, we use only our shared codes from the quantitative stage of the challenge.

- **Total method complexity :** Low to mid level. LGBPTOP features were very high dimensional (100224) in our former work, here we reduced them to half by taking mean of two concatenated feature representations of the temporal halves of the video. The max complexity is that of face alignment and DCNN_face feature extraction. ELM based learning and regression is simple.

Time complexity: 0.17s for face alignment and 2*0.03s(for 2 networks) for feature extraction means 0.23s processing time per image. Functional encoding takes another 3 seconds per video. With around 415 frames per video, one video takes around 98 seconds to process.

- **Which pre-trained or external methods/models have been used (for any stage, if any):** VGG-Face and VGG-VD19 (both are available from http://www.vlfeat.org/matconvnet/pretrained/)
- **Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any):** VGG-Face [1] is fine tuned on FER 2013 [2] expression corpus to improve emotion based discrimination (as in recent works see e.g. [4]).
- **Qualitative advantages of the proposed solution:** We do not train a deep network but use a pretrained model to fine-tune it. Then we extract face features from this network. Computation of Scene and Audio features are purely based on available resources. LGBPTOP feature extraction codes are also available online. Thus, all feature extraction process is easily reproducible. The two level fusion scheme not only improve performance but also shows the importance of using personality first impression predictions for job candidate screening. The classifiers used are fast and robust.
- **Novelty degree of the solution and if is has been previously published:**

First stage: The features used in this study are from our winning contribution to ICPR

2016 challenge (second round for LAP FI). A similar pipeline is applied on Video based Emotion Recognition in the Wild, and the details are reported in [4].

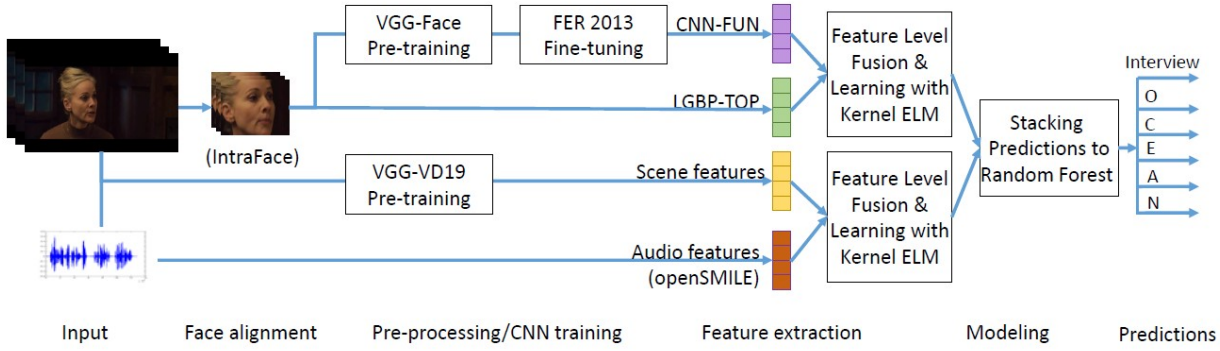
For the second stage, we use the predicted personality trait scores as meaningful high level features to train a decision tree for “invitation” target variable. This part of the work along with the fusion system employed for the quantitative stage are not published elsewhere.

2.3 **GitHub URL for the project:** **https://github.com/frkngrpnr/jcs_qual**

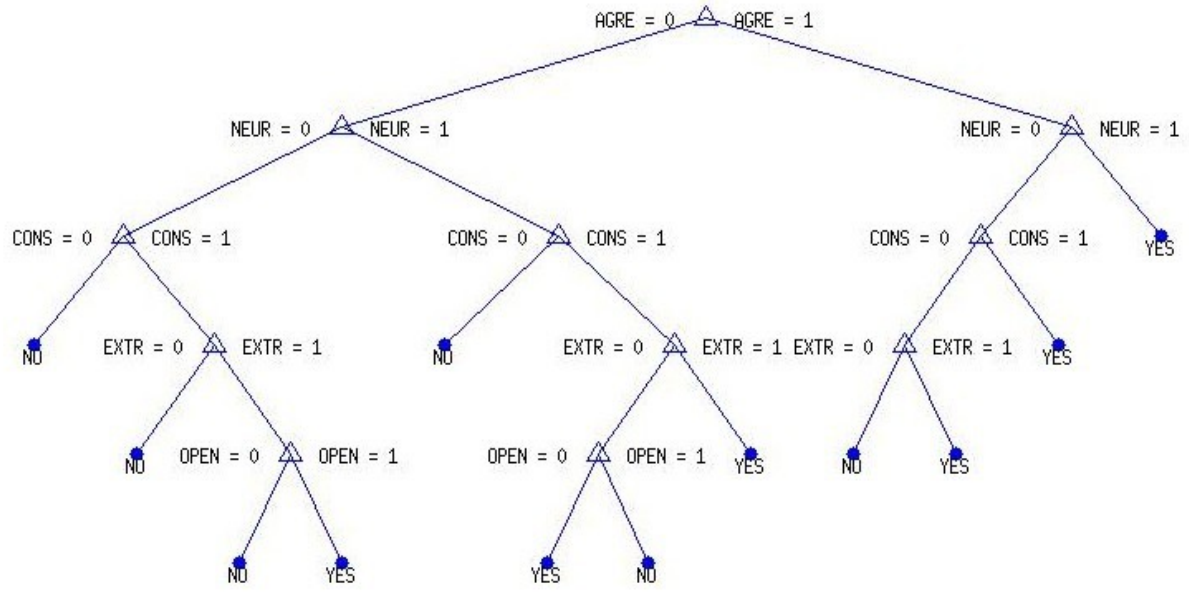
2.4 **References: See Section 9 for more detailed list.**

2.5 **Representative image / diagram of the method**

The first stage (fusion system for OCEAN and interview variable)



The second stage (training a decision tree to map binarized predicted OCEAN scores to binarized interview scores). 0/1 = Low/High personality score w.r.t training set mean. NO/YES= Reject/Invite for Interview



2.6 **Describe data preprocessing techniques applied:** Viola & Jones Face Detector [5] is used for face detection. IntraFace [6] is used for detecting the facial landmarks.

3. Visual Analysis

3.1 **Features / Data representation:** Face representation by LGBPTOP and FER fine tuned VGG-Face. Over the video, face features are summarized using five functionals (including

mean, std, and coefficients of second order polynomial fit to the contour). Scene representation by VGG-VD19 on the first frame of the video.
3.2 **Dimensionality reduction:** None
3.3 **Model:** Linear projection
3.4 **Learning strategy:** Kernel ELM [7] (after min-max normalization)
3.5 **Other techniques:** None
3.6 **Method complexity:** Mid level (see time complexity under Section 2.3)

4 Explainability from Visual data
The visual analysis (signal processing and machine learning) is the same as **Section 3**.
4.1 Features / Data representation
4.2 Dimensionality reduction
4.3 Model
4.4 Learning strategy
4.5 Other techniques
4.6 Method complexity

5 Explainability from Audio data
5.1 **Features / Data representation:** openSMILE [8] acoustic features using INTERSPEECH 2013 [9] baseline set.
5.2 **Dimensionality reduction:** None
5.3 **Model:** Linear projection (learning weights using a linear kernel)
5.4 **Learning strategy:** Kernel ELM (feature z-normalization and instance level L2-Norm)
5.5 **Other techniques:** None
5.6 **Method complexity:** The openSMILE tool is implemented in C++ , and audio processing is real time (8-10 times faster compared to the utterance length)
6 Explainability from ASR/text data
This modality is not used in the final system.
6.1 Features / Data representation
6.2 Dimensionality reduction
6.3 Model
6.4 Learning strategy
6.5 Other techniques
6.6 Method complexity

7 Multimodal Explainability
7.1 **Data Fusion Strategies**
The predictions from the face, scene and audio models are stacked to RF for final trait+interview prediction, which are then mapped using a decision tree for explainability.

8 Other details
8.1 **Language and implementation details (including platform, memory, parallelization requirements):** The feature extraction system is implemented in MATLAB R2015b on a 64-bit Windows 10 PC with 32GB RAM, Intel i7-6700 CPU. For fine-tuning and feature extraction with

CNNs, Mat-ConvNet library has been used with GPU parallelization using an Nvidia GeForce GTX 970 GPU. The machine learning routines and explainability parts were implemented in MATLAB 2014a on 64- bit Ubuntu 14.04 Workstation with 16 GB RAM.

8.2 **Human effort required for implementation, training and validation?:** No manual effort is required in any part of the pipeline.

8.3 **Training/testing expended time?:** As computed before, it took **11 days** to process all 10,000 videos. (**Note:** We will provide the extracted features to speed up the reproduction of results).

8.4 **General comments and impressions of the challenge? what do you expect from a new challenge in face and looking at people analysis?**

A baseline system (baseline features, scripts etc.) can be provided to the challenge coopetitors.

9 References

[1] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In British Machine Vision Conference, 2015.

[2] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea, J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests, 2013.

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[4] H. Kaya, F. Gürpınar, A. A. Salah, Video-Based Emotion Recognition in the Wild using Deep Transfer Learning and Score Fusion, Image and Vision Computing, Available online 4 February 2017, http://dx.doi.org/10.1016/j.imavis.2017.01.012.

[5] P. Viola and M. J. Jones. Robust real-time face detection. International journal of computer vision, 57(2):137–154, 2004.

[6] X. Xiong and F. De la Torre. Supervised Descent Method and Its Application to Face Alignment. In IEEE Conference on Computer Vision and Pattern Recognition, pages 532–539, 2013.

[7] [3] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 42(2):513–529, 2012.

[8] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In Proc. of the intl. conf. on Multimedia, pages 1459–1462. ACM, 2010. [Tool is freely available from: http://audeering.com/research/opensmile/]

[9] 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 INTERSPEECH, Lyon, France, 2013, pp. 148–152.