Fact sheet

Explainable Computer Vision Workshop

and

Job Candidate Screening Coopetition

Stacking Multimodal Predictions to Random Forest for Predicting First Impressions and Job Candidate Screening

Date: 10/02/2017

1. Team details

1.1 **Team name:** BU-NKU
1.2 **Team leader name:** Albert Ali Salah
1.3 **Team leader address, phone number and email:**
Boğaziçi Universitesi Bilgisayar Mühendisliği Bölümü, 34342, Bebek İstanbul / TURKEY
Tel: +90 212 359 (7774) E-mail: salah@boun.edu.tr

1.4 **Rest of the team members:** Heysem Kaya (Namık Kemal University), Furkan Gürpınar (Boğaziçi University)
1.5 **Team website URL (if any):** http://cmpe.boun.edu.tr
1.6 **Affiliation:** Boğaziçi University & Namık Kemal University

2. Contribution details

2.1 **Title of the contribution:** Stacking Multimodal Predictions to Random Forest for Predicting First Impressions and Job Candidate Screening
2.2 **Final score:** on val set **0.919782** (interview), **0.917021** (average big-5 traits)
2.3 **General method description:** 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.

- **Overview of the proposed approach:** First level ELM regression followed by stacking to RF. Feature level fusion, weighted score fusion and stacking strategies are used.
- **The proposed method uses / takes advantage of personality traits?** Yes, in the second level the personality trait predictions is shown to improve stacking performance
- **Coopetition**: **can your code be shared among other participants for the second stage of the challenge?** The codes used to extract the features are already available from https://github.com/frkngrpnr/lapfi Yes, our new prediction system that uses the aforementioned four audiovisual features can also be shared.

- **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.

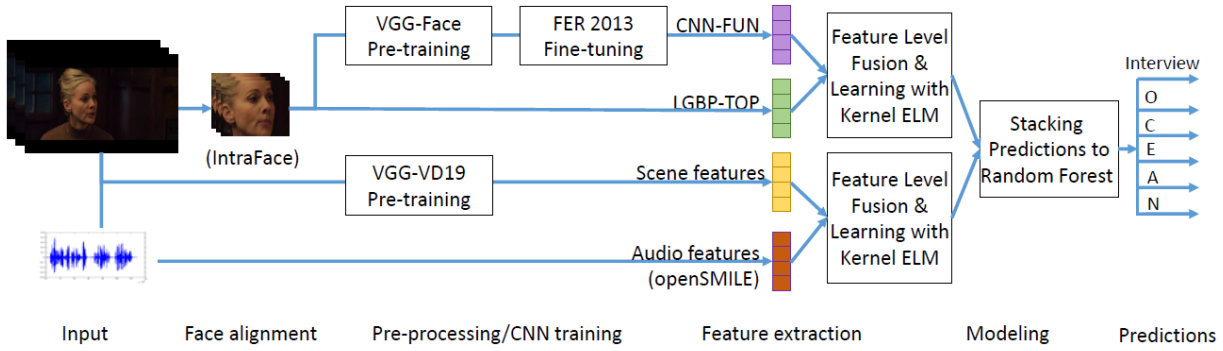
- **Which pre-trained or external methods/models have been used (for any stage, if any):** VGG-Face and VGG-VD19 (both are available from http://www.vlfeat.org/matconvnet/pretrained/)
- **Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, 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.
- **Results of the comparison to other approaches (if any):** Our results seem to advance the big-5 scores of the winning entry in the second round challenge (on val set from 0.9147 to 0.9170). Currently, the interview score ranks the first in CodeLab (on the val set).
- **Novelty degree of the solution and if is has been previously published:** 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]. Note that in [4] we use a random weighted fusion strategy, whereas here we use Random Forests (since we do not have the val set labels, try to optimize the system on the out-bag error).

2.4 **GitHub URL for the project:** https://github.com/frkngrpnr/jcs

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

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



2.7 **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 Personality Trait recognition from Visual data

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 Personality Trait recognition 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 Personality Trait recognition from ASR/text data
Note: We tried a simple Bag of Words representation followed by feature selection however it did not work well. Thus, our final system does not use a linguistic model.
6.1 **Features / Data representation:** BoW after removing punctuations and stopwords (the list is taken from http://www.ranks.nl/stopwords). 10808 Bow words left after purification.
6.2 **Dimensionality reduction:** Samples Versus Labels CCA Filter [10], top ranking 5900 features
6.3 **Model:** Linear projection (learning weights using a linear kernel)
6.4 **Learning strategy:** Kernel ELM (min-max normalization and instance level L2-Norm)
6.5 **Other techniques:** None
6.6 **Method complexity:** Computation of the wordbag from the training set (6000 instances)takes 52 seconds, Bow representation takes 75 seconds for training+validation sets (0.0094 seconds per instance). SLCCA is applied only once and therefore is fast.

7 Multimodal Personality Trait recognition
7.1 **Data Fusion Strategies**
We take two level fusion known as stacking in the literature. In the first level we apply feature level fusion: face sub-system combines face features (LGBPTOP and VGGFER_face), and the scene+acoustic sub-system combines IS13 and VGG-VD19 features. These are modeled using Linear Kernel ELM. The 12 dimensional outputs of the first level (one for each target variable x 2 regressors) are then given to Random Forest regressor with 100 trees. The trees are grown with a random subset of features (each with one third of the original features = 4 dimensions) and a random subset of instances (sampled with replacement). This scheme is experimentally shown to improve performance over the simple weighted fusion alternative all dimensions except agreeableness, where we observed better performance with simple weighted fusion.
8 Other details
8.1 **Language and implementation details (including platform, memory, parallelization requirements):** The whole 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.
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?**

We would prefer that one team could only submit five prediction sets for the test set and the scores of the test submissions would be available to the submitting team at the time of submission. Moreover, instead of taking the final submission, it could be possible to take the submission with the highest accuracy for the final performance of each team.

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.

[10] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller. CCA based feature selection with application to continuous depression recognition from acoustic speech features. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* (pp. 3729-3733). 2014.