Combining Deep Facial and Ambient Features for First Impression Estimation

July 12, 2016

1 Team details

- **Team name:** BU-NKU
- **Team leader name:** Albert Ali Salah
- **Team leader address, phone number and email:** 34342 Bebek, Istanbul, Turkey, +90 212 359 7774, salah@boun.edu.tr
- **Team Members:** Furkan Gürpınar, Heysem Kaya,
- **Team website URL:** http://cmpe.boun.edu.tr
- **Affiliation:** Boğaziçi University & Namık Kemal University

2 Contribution details

- **Title of the contribution:** Combining Deep Facial and Ambient Features for First Impression Estimation
- **Final score:** 0.9113 on validation set
- **General method description:** Deep feature extraction from faces and scenes, feature-level fusion and regularized regression with Kernel Extreme Learning Machine [2]
- **References:** (see the last section)
- **Representative image / diagram of the method:** (see Figure 1)
- **Describe data preprocessing techniques applied (if any):** Input normalization with average training images of the networks, and instance-level min-max normalization of features.

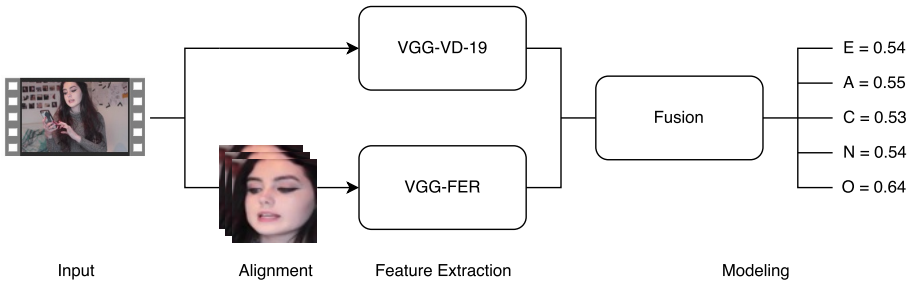

Figure 1: System Pipeline

3 Visual Analysis

3.1 Face Detection Stage

Viola & Jones Face Detector [5] is used for face detection.

3.2 Face Landmarks Alignment Stage

IntraFace [6] is used for detecting the facial landmarks.

3.2.1 Other techniques

We eliminate the roll angle by estimating it from the eye corner locations.

3.3 Facial expression recognition

3.3.1 Features / Data representation

We extract deep CNN features from a network that is VGG-Face [3], fine-tuned on FER-2013 dataset [1] for facial expression recognition. We represent each video with functional statistics computed over frames.

3.3.2 Learning strategy

We use kernel ELM for regularized regression.

4 Personality Trait recognition from Visual data

4.1 Features / Data representation

Other than the face features explained in Section 3.3.1, we employ VGG-VD-19 network [4] to extract features from the whole scene.

4.2 Learning strategy

We combine the two systems with both early (feature-level) and late (decision-level) fusion. We choose feature-level fusion because it provides higher accuracy in 10-fold cross validation on the training set.

5 Global Method Description

- **Total method 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 have been used (for any stage, if any)** [4] is used for feature extraction, and [3] is used as a base model for CNN fine-tuning.
- **Which additional data has been used in addition to the provided ChaLearn training and validation data (at any stage, if any)** : FER dataset in CNN fine-tuning
- **Qualitative advantages of the proposed solution** The proposed solution makes use of pre-trained models for feature extraction. Hence, it is parsimonious for source utilization. We also employ a fast and accurate classifier, namely ELM, for model learning. On the overall, the proposed solution is easily reproducible, simple and fast.
- **Results of the comparison to other approaches (if any)** The information on other approaches on this challenge are not available yet.
- **Novelty degree of the solution and if is has been previously published** The novelty of this contribution is twofold. First is combining deep ambient (scene) features with the deep facial features. The second contribution is using a DCNN that is pre-trained for face detection and fine tuned for emotion recognition. Here, we try to make use of the interrelation of facial expression (emotion) and perceived personality traits. The approach taken on this corpus was not published elsewhere.

6 Other details

- **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, MatConvNet library has been used with GPU parallelization using an Nvidia GeForce GTX 970 GPU.

- **Human effort required for implementation, training and validation?** No manual effort is required in any part of the pipeline.
- **Training/testing expended time?** As computed before, it took 11 days to process all 10,000 videos.
- **General comments and impressions of the challenge? What do you expect from a new challenge in face and looking at people analysis?** The challenges provide a common benchmark and protocol for evaluation and hence allow comparability as well as reproducibility of results. In this specific challenge, there was a confusion about the test set submissions. We would prefer that one team could only submit five prediction sets and the scores of the test submissions would be available to the submitting team. This way mitigates the risk of over-fitting and allows the team to test alternative hypotheses.

References

[1] 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.

[2] 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.

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

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

[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.