# ChaLearn Looking at People @ ECCV 2016
# Apparent Personality Analysis: First Impressions –
# A Study of Multi-modal Apparent Personality Analysis

July 16, 2016

## 1 Team details

- Team name
  ucas

- Team members
  Mengyi Liu (mengyi.liu@vipl.ict.ac.cn)
  Yan Li (yan.li@vipl.ict.ac.cn)

- Affiliation
  Institute of Computing Technology, Chinese Academy of Sciences

## 2 Contribution details

- Title of the contribution
  A Study of Multi-modal Apparent Personality Analysis

- General method description
  Our method is based on a combination of visual and audio features. For visual features, we extract both static scene/face features via deep convolutional neural networks [1] (AlexNet [2, 3], VGGNet [4, 5], ResNet [6, 7]) and dynamic features via spatio-temporal descriptors, (HOG3D [8], LBP-TOP [9]). For audio features, we extract a batch of predefined feature set using an open-sources tool OpenSMILE [10]. In decision stage, we employ two kinds of linear regressors, i.e. Partial Least Squares regressor [11] and Support Vector Regressor [12], on all of the visual/audio features above respectively, and then fusing their decision scores for final predictions.

- References
  (Please refer to the end of this manuscript.)

# 3 Visual Analysis

## 3.1 Face Detection Stage

Open-source tool Face++ [13].

## 3.2 Face Landmarks Alignment Stage

Open-source tool Face++ [13].

# 4 Personality Trait recognition from Visual data

## 4.1 Data preprocessing

- Face image preprocessing
  All the detected face images should be normalized according to eye locations and resized to 64x64 for spatio-temporal descriptors and 224x224 for deep convolutional neural networks.
- Scene image preprocessing
  Scene images are original video frames. For feature extraction, all images are resized to 64x64 for spatio-temporal descriptors and 224x224 for deep convolutional neural networks.

## 4.2 Dynamic features

As the labels for both validation and test datasets are still blind to participants, we report our results based on a self-defined protocol using all 6,000 training videos which takes 5,400 videos as training and 600 videos as testing. Note that, the subjects from these two splits have no overlap with each other.

Table 1: Performance comparison of dynamic features.

| +pls | Accuracy | Extra | Agree | Consci | Neuro | Open |
|---|---|---|---|---|---|---|
| hog3d-face-8*8*16*8 | 0.8994 | 0.9031 | 0.8977 | 0.8965 | 0.8945 | 0.9003 |
| hog3d-scene-8*8*16*8 | 0.8742 | 0.8710 | 0.8792 | 0.8733 | 0.8686 | 0.8788 |
| lbptop-face-8*8*3*59 | 0.8979 | 0.9023 | 0.8986 | 0.8952 | 0.8934 | 0.9001 |
| lbptop-scene-8*8*3*59 | 0.8836 | 0.8796 | 0.8877 | 0.8829 | 0.8803 | 0.8873 |

## 4.3 Static features

Please refer to Table. 4 for the performance of static scene/face features.

Table 2: Performance comparison of static scene/face features.

| +pls | Accuracy | Extra | Agree | Consci | Neuro | Open |
|---|---|---|---|---|---|---|
| vggface-face-fc6 | 0.8992 | 0.9027 | 0.8998 | 0.8963 | 0.8962 | 0.9012 |
| vggface-face-pool5 | 0.8997 | 0.9040 | 0.8985 | 0.9012 | 0.8964 | 0.8997 |
| alexnet-cfw-face-fc6 | 0.8983 | 0.9031 | 0.8988 | 0.8971 | 0.8947 | 0.8980 |
| alexnet-cfw-face-pool5 | 0.9003 | 0.9041 | 0.8992 | 0.8998 | 0.8974 | 0.9007 |
| vgg-imagenet-face-fc6 | 0.8938 | 0.8887 | 0.8984 | 0.8978 | 0.8891 | 0.8950 |
| vgg-imagenet-scene-fc6 | 0.8933 | 0.8871 | 0.8973 | 0.8996 | 0.8864 | 0.8961 |
| resnet-imagenet-face-pool5 | 0.8961 | 0.8934 | 0.8973 | 0.9013 | 0.8910 | 0.8974 |
| resnet-imagenet-scene-pool5 | 0.8960 | 0.8915 | 0.8982 | 0.9025 | 0.8890 | 0.8986 |

# 5 Personality Trait recognition from Audio data

## 5.1 Data preprocessing

The video data should be first converted from “mp4” to “wav”. Then we employ openSMILE to extract a batch of speech-related features, e.g. MFCC, PLP-CC, Pitch, Voice quality, Formants, LPC, Line Spectral Pairs (LSP), Spectral Shape descriptors (cited from http://audeering.com/research/opensmile/download).

## 5.2 Audio features

Table 3: Performance comparison of static audio features (pls).

| +pls | Accuracy | Extra | Agree | Consci | Neuro | Open |
|---|---|---|---|---|---|---|
| emobase | 0.8825 | 0.8784 | 0.8998 | 0.8937 | 0.8957 | 0.9018 |
| emobase2010 | 0.8971 | 0.8947 | 0.8998 | 0.8937 | 0.8957 | 0.9018 |
| IS10-paraling | 0.8966 | 0.8938 | 0.8996 | 0.8930 | 0.8951 | 0.9016 |
| IS10-paraling-compat | 0.8969 | 0.8942 | 0.8997 | 0.8936 | 0.8952 | 0.9020 |

Table 4: Performance comparison of audio features (svr).

| +svr | Accuracy | Extra | Agree | Consci | Neuro | Open |
|---|---|---|---|---|---|---|
| emobase | 0.8966 | 0.8970 | 0.8969 | 0.8892 | 0.8975 | 0.9025 |
| emobase2010 | 0.8960 | 0.8940 | 0.8982 | 0.8898 | 0.8962 | 0.9016 |
| IS10-paraling | 0.8958 | 0.8947 | 0.8969 | 0.8900 | 0.8959 | 0.9016 |
| IS10-paraling-compat | 0.8958 | 0.8947 | 0.8969 | 0.8900 | 0.8959 | 0.9016 |

# 6 Multimodal Personality Trait recognition

## 6.1 Data Fusion Strategies

We perform decision level fusion among all of the visual and audio features. The final score obtained is 0.9091 (0.9110, 0.9054, 0.9105, 0.9063, 0.9121).

## 6.2 Complexity (time cost)

- CPU
  Intel(R) Core(TM) i7-3770 CPU @ 3.40GHz; Memory Size: 32GB
  Stages:
  Face Detection: 0.5s per frame;
  Face landmark localization (83 points): 0.3s per frame;

- GPU
  NVIDIA GeForce GTX TITAN Black; Operating System: Ubuntu 14.04
  Stages:
  GoogleNet pool5/7x7s1 feature extraction: 0.02s per frame;
  VGG fc6/fc7 feature extraction: 0.025s per frame;
  VGGface fc6/fc7 feature extraction: 0.027s per frame;
  CaffeNet fc6/fc7/pool5 feature extraction: 0.031 per frame;
  ResNet pool5 feature extraction: 0.033 per frame;

# References

[1] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Multimedia, ACM (2014) 675–678

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012) 1097–1105

[3] Li, Y., Wang, R., Liu, H., Jiang, H., Shan, S., Chen, X.: Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In: ICCV. (2015) 3819–3827

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

[5] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. Volume 1. (2015) 6

[6] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition. (2015) 1–9

[7] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. preprint arXiv:1512.03385 (2015)

[8] Klaser, A., Marszałek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC, British Machine Vision Association (2008) 275–1

[9] Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. PAMI **29**(6) (2007) 915–928

[10] Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Multimedia, ACM (2010) 1459–1462

[11] Wold, H.: Partial least squares. Encyclopedia of statistical sciences (1985)

[12] Smola, A.J., Schölkopf, B.: A tutorial on support vector regression. Statistics and computing **14**(3) (2004) 199–222

[13] Inc., M.: Face++ research toolkit. www.faceplusplus.com (December 2013)