FaceMax : Personality Analysis from videos

Skand Vishwanath Peri
2014CSB1034
Department of Computer Science and Engineering, IIT Ropar, Punjab
Abhishek Jangra
2014EEB1041
Department of Electrical Engineering, IIT Ropar, Punjab

1 Introduction

Humans continuously perform evaluations of personality characteristics of others. First impressions on personality traits, despite being inaccurate, play a crucial role in many essential decisions in our everyday lives, such as the results of the elections, or court verdicts. These personality trait inferences are driven by informational cues with an evolutionary incentive.

Trait theories of personality have long attempted to pin down exactly how many personality traits exist. Earlier theories have suggested a various number of possible traits, including Gordon Allport's list of 4,000 personality traits, Raymond Cattell's 16 personality factors, and Hans Eysenck's three-factor theory. However, many researchers felt that Cattell's theory was too complicated and Eysenck's was too limited in scope. As a result, the five-factor theory emerged to describe the essential traits that serve as the building blocks of personality.

2 Dataset

All the papers whose approaches would be discussed have used the **ChaLearn and First Impressions Dataset**. The ChaLearn First Impressions provides a large corpus of annotated videos and it is one of the most popular benchmarks for apparent personality trait inference. The First Impressions dataset consists of 10,000 clips extracted from more than 3,000 different YouTube HD (720p) videos of people facing and speaking in English to a camera. These 10,000 clips are divided into three different subsets: 6, 000 clips for training, 2, 000 clips for validation and 2, 000 clips for testing. Each clip lasts 15 seconds. The videos are labeled according to the Big Five personality traits in a continuous 0 âĹŠ 1 scale. Figure 2 shows some video frames illustrating different personality traits scores.

3 Literature Review

In this section we are going to discuss the contributions of 4 papers in personality assessment.

3.1 Personality traits and job candidate screening via analyzing facial videos[3]

Bekhouche et al. decided that personality of a person can be derived from their facial expressions. During preprocessing, they extracted all the frames from the video. For each frame they apply Harr Cascasade Object Detector that used Viola Jones algorithm in order to detect the face region. Then they detect the face landmarks using ensemble of regression trees (ERT algorithm), where they use the location of eyes and nose to align the image using 2D homographic transformation.
After obtaining all the aligned faces, they apply 2 different texture descriptors.

- Local Phase Quantization (LPQ)- which can handle motion blur in the video
- Binarized Statistical Image Features (BSIF) - inspired by LBP and LPQ to improve accuracy in texture recognition.

The results of these two descriptor are represented by a 7 level pyramid. The features extracted from both pyramids are concatenated in one vector so that each video frame has its own feature vector as shown in Figure 1. To compute the feature vector for the whole video, they compute the mean of the features from all the frames. Then they feed this feature vector into five non linear Support Vector Regressors (SVR) to estimate the five scores corresponding to each personality. Then they feed these values to a Gaussian Process Regressor (GPR) to obtain the interview score which gives a score on a scale of 0-1 of how likely should the person be called for an interview.

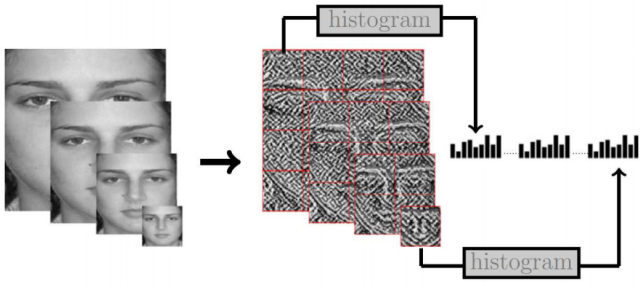

Figure 1: Figure showing the concatenation of features extracted from 7 Pyramid Levels.

3.2 Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition[4]

Gucluturk et al. in their paper considered the audio features as well unlike the previous paper which only considered the videos. Unlike the previous paper this approach does not require any feature engineering or visual analysis such as face detection, face landmark alignment or facial expression recognition. They also have a deep convolutional network in order to achieve good accuracy. The use Microsoft's ResNet architecture[7] for extracting audio as well as video features as shown in Figure 2.
Each training video clip was processed as follows:

- The audio data and the visual data of the video clip are extracted.
- A random 50176 sample temporal crop of the audio data is fed into the auditory stream. The activities of the penultimate layer of the auditory stream are temporally pooled.
- A random 224×224 pixels spatial crop of a random frame of the visual data is randomly flipped in the left/right direction and fed into the visual stream. The activities of the penultimate layer of the visual stream are spatially pooled.
- The pooled activities of the auditory stream and the visual stream are concatenated and fed into the fully-connected layer.
- The fully-connected layer outputs five continuous prediction values between the range [0, 1] corresponding to each trait for the video clip.

The validation or test video clip is processed as follows:

- The audio data and the visual data of the video clip are extracted.
- The entire audio data are fed into the auditory stream. The activities of the penultimate layer of the auditory stream are temporally pooled (see below note).
- The entire visual data are fed into the visual stream one frame at a time. The activities of the penultimate layer of the visual stream are spatiotemporally pooled (see below note).
- The pooled activities of the auditory stream and the visual stream are concatenated and fed into the fully-connected layer.
- The fully-connected layer outputs five continuous prediction values between the range [0, 1] corresponding to each trait for the video clip.

This architecture won the third place in ECCV 2016 workshop on Personality Traits Assessment.

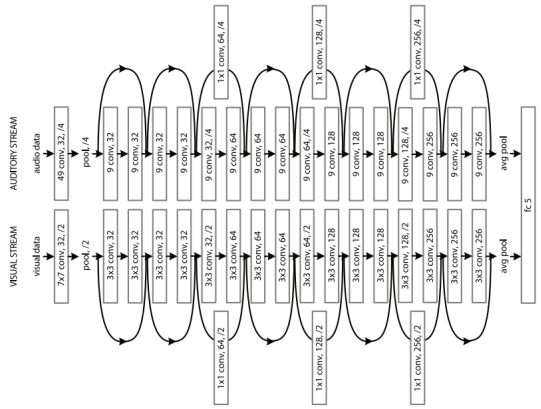

Figure 2: ResNet like network for Audio-Visual Features extraction and classification

3.3 Deep Bimodal Regression for Apparent Personality Analysis[6]

In DBR, for the visual modality, *Zhang et al.* modify the traditional convolutional neural networks for exploiting important visual cues. In addition, taking into account the model efficiency, we extract audio representations (mfcc and logfilterbank) and build the linear regressor for the audio modality. For combining the complementary information from the two modalities, we ensemble these predicted regression scores by both early fusion and late fusion.

3.3.1 Visual Feature extraction

From each video, they extract 6 images per second giving them around 100 images per video. After that images are labelled with same personality trait as the corresponding video. Now they train a deep regressor using CNN's for a personality analysis. They use a modified version of Descriptor Aggregator Network (DAN) called as DAN$^+$.
What distinguishes DAN from the traditional CNN is: the fully connected layers are discarded, and replaced by both average- and max-pooling following the last convolutional layers ($Pool_5$).Meanwhile, each pooling operation is followed by the standard l_2- normalization. After that, the obtained two 512-d feature vectors are concatenated as the final image representation. Thus, in DAN, the deep descriptors of the last convolutional layers are aggregated as a single visual feature. Finally, because APA is a regression problem, a regression (fc+sigmoid) layer is added for end-to-end training.
What distinguishes DAN$^+$ from the traditional CNN is: the deep convolutional descriptors of $ReLU_{5_2}$ are also incorporated in the similar aforementioned aggregation approach, which is shown in Fig. 3. Thus, the final image feature is a 2048-d vector. They call this end-to-end deep regression network as DAN$^+$.

3.3.2 Audio Feature extraction

In the audio modality, they choose logfilterbank and mfcc features as audio representations. After that they use a model consisting of fully connected layers followed by sigmoid activation to train the audio regressor. l_2 distance is used as the loss function to calculate the regression loss.

3.3.3 Combining Visual and Audio Features

They use the simply averaging method giving equal weightage to audio(mfcc and logfilterbank) and visual features.

$$FinalScore = \sum_{i=1}^{3} \frac{s_i}{3} \quad (1)$$

where s_i is a 5-dimensional vector consisting of 0-1 range values for each class.

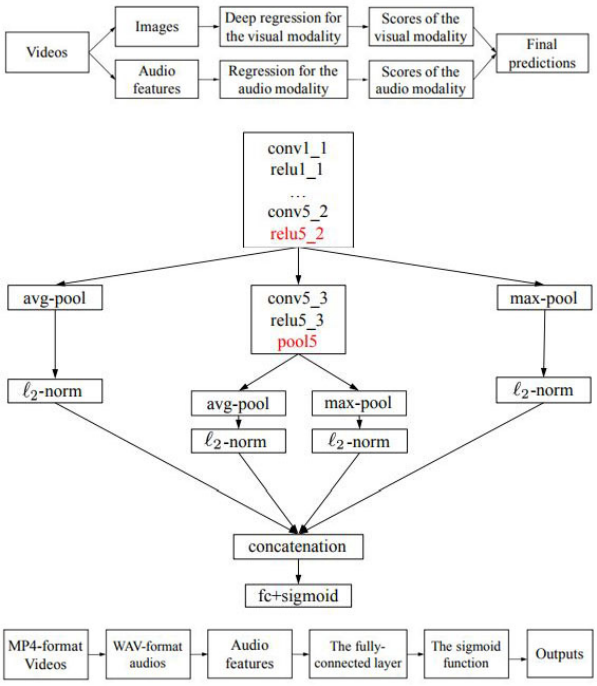

Figure 3: The top figure shows the over all network architecture, the DAN$^+$ architecture is shown in the middle and the architecture to extract the audio features is shown at the bottom of the figure.

4 Interpreting CNN Models for Apparent Personality Trait Regression[5]

Ventura et al., questiones *Zhang et al's.* approach of considering audio and visual features. They raised 2 questions.

- Is audio require to classify the traits?
- Is the whole image required to achieve the results of simply using only the face would give better results?

After doing experiments they concluded that by using only visual features they could get extremely close to the state-of-the-art accuracy and also clearly showed that using face image was outperforming the network trained with the whole image.
In their architecture, they used the DAN$^+$ architecture proposed by *Zhang et al.* in their paper but also added a novelty by considering **Class Activation Map** (CAM)[9] to classify the traits. CAM is used to visualize class-specific discriminative regions. They modified the DAN$^+$ architecture to avoid the max and average pooling after the $relu5_2$ layer, instead they went off to generate the CAM hence avoiding the concatenation layer in the usual DAN$^+$ as shown in Figure 5.

They also do an analysis on Action Units for Personality Traits Prediction as their last set of experiments in which they focus on evaluating the influence of shown emotion for the problem of personality trait inferences. They use openface[2] library in order to extract the Action Units of

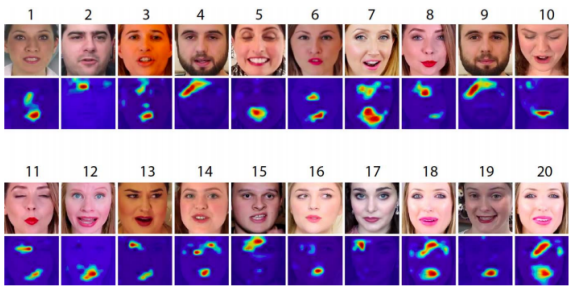

Figure 4: Discriminative localization (class activation maps) obtained for the 20 images that give the highest predicted value for the agreeableness personality trait in the test subset

.

each face. They use the AU activation as a 17- dimensional feature vector, and trained a simple linear classifier on this data. This simple model yields an accuracy close to 0.886 with this reduced set of features. This result suggests that there is a dual informational cue when inferring social traits from facial images.

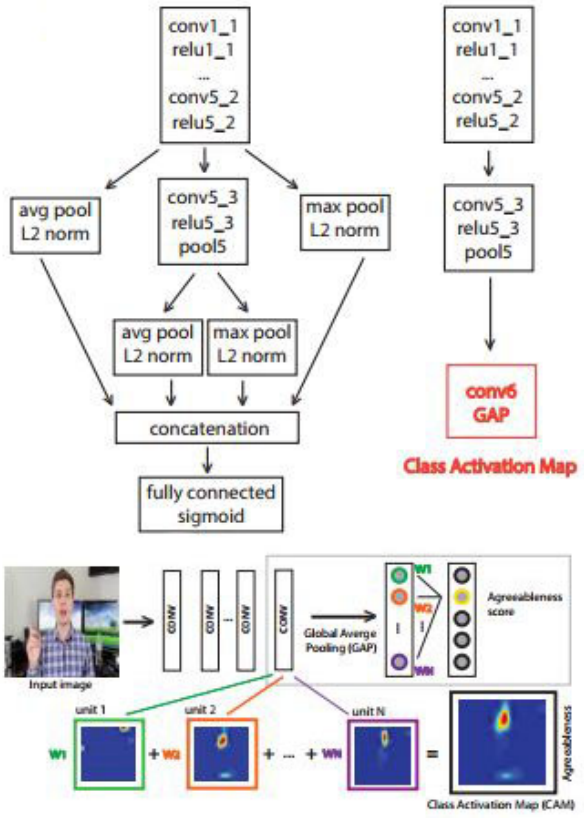

Figure 5: CNN architecture used for video modality in [23]; b) Their modification of the CNN architecture to add the Class Activation Map (CAM) module. The bottom picture shows the classification from the CAM using GAP (Global Average Pooling)

.

5 Methodology

In contrast to *Ventura et.al* we feel that the background of the video does have some correlation with the person's traits. First of all we extracted 30 frames from the each of the 15 second videos given in the training and validation set. Then we ran OpenFace on each of the frames to extract the face from the image. In order to consider the background, we used a Gaussian blur to blur the face out from the frame. We felt that simply blacking out the facial part in the frame would not be a good data to the VGG network. We did the following 3 experiments:

5.1 HOG Feature

HOG feature for each face in every frame was extracted and the element wise *max* was taken as the final feature. Lets call this feature as **HOGMax** feature. The HOGMax feature vector was of 6084 dimension. After this, a simple MLP of 3 layers with relu activation was trained.

5.2 VGG-Face Feature

VGG-Face feature was extracted for each of the faces and similary to HOGMax, element wise *max* was computed over these features for a video and the final feature was considered for further fine tuning. Let's call this as **VGG-FaceMax** feature.

5.3 VGG-Face + VGG Feature

Similar to the above feature VGG-FaceMax was obtained for all the videos and in addition to this VGG16 feature (pretrained on imagenet) was obtained for the background image. Since in all the videos the background was more or less same so we just took one frame from each video and blurred out the face and passed it into the VGG16 pretrained model. We call this feature as **VGG-FaceMax-BG** feature.

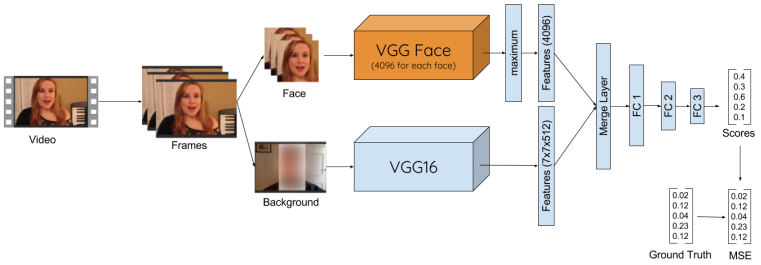

Figure 6: Proposed architecture with VGG-FaceMax and VGG16 features concatenated.

.

We used Pytorch[1] a Deep-Learning Library for training our networks. The loss function for all the experiments is the mean squared error with L2 regularizer in the FC layers to prevent overfitting. We used Adam [8] optimizer for stochastic optimization.

5.4 Comparision Metric

As mentioned in the Chalearn Dataset description, to calculate the accuracy the following formula was used.

$$\frac{1}{5}\sum_{i=1}^{5}\sum_{j=1}^{N} 1 - |trueValue_j^i - predictedValue_j^i| \quad (2)$$

6 Results

We compare our 3 methods with *Bekhouche et al.*, Gucluturk et al., *Zhang et al.*, *Ventura et al.* (image) and *Ventura et al.* (face) where image is the complete image and face is the model that is only trained on faces extracted.

MA - Mean Accuracy, A - Agreeableness, C - Conscientiousness, E - Extraversion, N - Neuroticism, O-Openness.

Author	MA	A	C	E	N	O
Bekhouche et al.	0.9115	0.9103	0.9137	0.91551	0.9082	0.910
Gucluturk et al.	0.9109	0.9101	0.9377	0.9107	0.9089	0.9110
Zhang et al.	0.9130	0.9126	0.9166	0.9133	0.9100	0.9123
Ventura et al.(image)	0.909	0.905	0.911	0.919	0.905	0.909
Ventura et al.(face)	0.912	0.912	0.914	0.915	0.907	0.910
HOGMax (Ours)	0.8537	0.8690	0.8147	0.8683	0.8549	0.8616
VGG-FaceMax-BG (Ours)	0.8926	0.8992	0.8889	0.8986	0.8789	0.8975

7 Conclusion

The results reflect that Computer Vision tackles this problem quite nicely. Out method i.e. FaceMax gives accuracy close to 90%. Performance is also acceptable, it took around 1 hour to train on a set of 6000 videos *with features already extracted*. From our tests we can conclude that background does provides relevant information about a personality (unless tests are conducted in controlled environment).

8 Further Extension

- Also include audio data - can prove to be helpful just like background .

- Increase the number of frames extracted per video (31 currently) - more information to work with.
- Try different features like Fischer Vector and LBP-top.

9 References

[1] http://pytorch.org/, 2016.

[2] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[3] Bekhouche et al. Personality traits and job candidate screening via analyzing facial videos. *CVPR 2017*, .

[4] Gucluturk et al. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. *ECCV 2016*, .

[5] Ventura et al. Interpreting cnn models for apparent personality trait regression. *CVPR 2017*, .

[6] Zhang et al. Deep bimodal regression for apparent personality analysis. *ECCV 2016*, .

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

[9] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.