Bi-modal Regression for Apparent Personality Trait Recognition

Nishant Rai
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur
Uttar Pradesh, India, 208016

***Abstract*—The task of the ChaLearn Apparent Personality Analysis: First Impressions Challenge is to rate/quantify personality traits of users in short video sequences. Although the validity of personality judgments from short interactions is questionable, studies show the possibility of predicting attributed traits (First Impressions) using facial [15] and acoustic [13] features. The challenge introduces a newly constructed dataset which consists of manually annotated videos collected from YouTube. In this paper, we present our approach for predicting traits by combining multiple modality specific models. Our models include Deep Networks which focus on leveraging visual information in the given faces, Networks focusing on supplementary information from the background and models using acoustic features. We also discuss another approach for modeling traits as a combination of global and trait-specific variables. We explore methods for extracting fixed length descriptors of videos based on frame-level predictions. We also experiment with various methods for fusing model predictions. We observe that fusion achieves a considerable gain in accuracy over the best stand-alone model, possibly due to utilizing information from all modalities. The proposed method achieves an accuracy gain of approximately 18% above the provided challenge baseline.**

I. Introduction

In this paper, we propose a method to automatically recognize Apparent Personality traits (i.e. First Impressions), as part of the Chalearn: First Impressions Challenge, organized in conjunction with the IAPR International Conference on Pattern Recognition (ICPR 2016). The challenge dataset is a newly collected dataset consisting of 10K short videos from YouTube. They have been manually annotated with personality traits by AMT workers. Scores were assigned for each personality trait from the Big-Five Traits [8] i.e. Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. The challenge task is to predict trait scores (Between 0 and 1) given a short video clip (Average duration is 15s).

The dataset contains 10K short videos, out of which 6K are available for training. It can be seen that Apparent Personality traits depend on multiple factors such as behavior, appearance, speech and possibly context. Given the low number of samples, it is difficult to deal with the large variety of subjects and variety in conditions. Also, the video subjects are of different gender, age, nationality, and ethnicity, which makes the task even more challenging. Fortunately, the videos generally contain frontal poses, therefore reducing variance in pose.

We explore multiple modality-specific models such as Convolutional Neural Networks focusing on leveraging visual information in detected faces, networks focusing on supplementary background information and acoustic feature based models. We treat the trait prediction problem as a regression task, not assuming any relations between the traits. We also observe that the Big-Five traits are not independent of each other and are instead positively correlated. There have been many studies supporting this claim [18], [8]. Inspired from such work, we propose an alternative way to model traits as a combination of global and trait-specific factors.

These models help us arrive at frame/segment-level predictions. Given segment-wise predictions for each segment in a video, we represent each video as a collection of segment predictions. This is achieved by simply concatenating the predictions for each segment while using expansion and compression to arrive at a fixed length video representation [9]. We also explore multiple fusion and aggregation strategies for computing video level predictions from frame-wise cues. Using the aggregated video representation leads to significant improvements. The best stand-alone model, ignoring combinations with others is the acoustic feature based model. These modality-specific models are further combined using various methods. This further boosts the performance and gives us our best model which secured 3^{rd} position in the final evaluation.

Paper Organization: A brief overview of Prior work in this field is given in Section II. This is followed by Section III, which introduces the dataset used for the challenge and the associated difficulties. Our proposed models are described in Section IV along with details of the aggregation and fusion strategies. This is followed by Section V that summarizes the results. Conclusion and discussions are presented in Section VI.

II. Prior and Related Work

We briefly discuss related work for audio and visual based models used for recognition and classification. We also mention related work in the field of Automatic Personality Perception.

Acoustic feature based Recognition: There has been a surge in interest related to Audio Based Recognition and Classification. In recent years, multiple audio representations

have been suggested. Features such as Mel Frequency Cepstral Coefficients (MFCC) [12] have given good results in tasks related to Emotion Recognition, Age/Gender estimation, etc.

Deep Learning for Visual Features: Deep learning is a branch of machine learning based on algorithms that attempt to model high-level abstractions in data by using graphs with multiple processing layers, generally composed of multiple linear and non-linear transformations. Deep Learning methods have attracted huge interest due to recent models achieving state of the art performance in both image and audio related tasks. Recently, very deep networks have achieved SoA results in tasks including Image Classification [19], [6], Action/Object Recognition [17] and Semantic Segmentation [14]. Specifically, Convolutional Neural Networks (CNNs) have grown to be the method of choice for processing visual data. In our models, we use CNNs by training them on facial features and obtaining frame wise predictions, which are further fused for the final predictions.

Automatic Personality Perception: Personality analysis has been an area of vast research in the domain of psychology [8], [4], [18]. There has been a great deal of work exhibiting the effect of visual cues on formation of first impressions [1], [15]. Most studies in psychology focus on facial expressions as people frequently use facial characteristics as a basis for personality attributions. The authors of [3], [5] studied the human tendency to judge others based on their faces. They also discovered some important facial features important in forming first impressions. In addition, they revealed that humans can make valid inferences for at least four personality traits from facial features. In the computer science domain, research in Automatic Personality Perception has focused mainly on nonverbal behavior and online activities (such as tagging images on social media) [11]. The problem of automatically mapping audio-visual information to personality traits has been neglected in the computing literature. As far as we are aware, there has not been any work which considers Apparent Personality prediction through videos.

III. Dataset

The challenge dataset is a newly constructed dataset consisting of short video clips collected from YouTube, manually annotated with the apparent personality traits. As far as we are aware, this is the first such challenge specially focused on trait analysis with such a large dataset. For each video sample, RGB and audio information, along with the continuous ground-truth values for each of the Big-Five Personality Traits are provided. The continuous ground-truth values for each trait are computed by means of a calibration analysis using around 350,000 pairwise AMT annotations. The video clips contain a main subject who is at a safe distance from the camera. Special care has been taken to ensure a unique major subject. Most of the faces are frontal and there is not much camera movement which makes it a very clean dataset. The dataset is divided into three parts: the training set (6K videos), the validation set (2K videos) and the evaluation set (2K videos). The statistics for the trait-score distributions are given in table 1. They are well distributed and have a reasonable spread with the extremes being close to 0 and 1.

Statistics	Extra.	Agree.	Consient.
Mean	0.476	0.548	0.523
Std-Dev	0.152	0.136	0.155
Min	0.000	0.000	0.000
Max	0.925	1.000	0.971

Statistics	Neuro.	Open.
Mean	0.520	0.566
Std-Dev	0.154	0.147
Min	0.021	0.000
Max	0.979	1.000

TABLE I: Statistics for the Big-Five Personality traits

There were a few challenges we faced while working with the dataset. The variation in gender, age, nationality, ethnicity made the dataset quite challenging. Coupling this with the fact that first impressions depend on a multitude of factors makes the task extremely challenging with the small number of training samples. We perform aggressive augmentation and make use of pre-trained [19] models to counter this issue. We have observed that sometimes similar videos are given different scores. During the dataset construction, a large video is broken into multiple parts and then participants annotate them. In some cases different segments of the same video are given very different scores (Even though the segments are very similar). While this could surely be due to genuine reasons such as the subject behaving differently in different segments, manually checking the videos in question shows otherwise.

IV. Proposed Models for Personality Analysis

In this section, we introduce our proposed models for Apparent personality analysis which exploit information provided by various modalities. The overall structure of the model is given in Fig. 1. As shown in numerous studies, visual and speech information have a large effect on the formation of first impressions. Therefore, we construct multiple specialist models which take a particular modality and predict personality traits. Our model consists of multiple visual specialists along with a few audio specialists. We explore multiple approaches for combining the predictions of the base specialists to get the final prediction. Motivated by studies showing relationships between the big-Five traits, we also discuss an alternative way to model the traits as a mixture of global and trait specific components.

A. Visual Feature Based Models

We first discuss models which perform predictions based on visual features in the frames. The features which are explored are as follows,

1) *Facial features:* As discussed earlier, there is evidence pointing towards a strong relationship between facial appearance and first impressions. Therefore we construct models which use facial features to predict the scores.

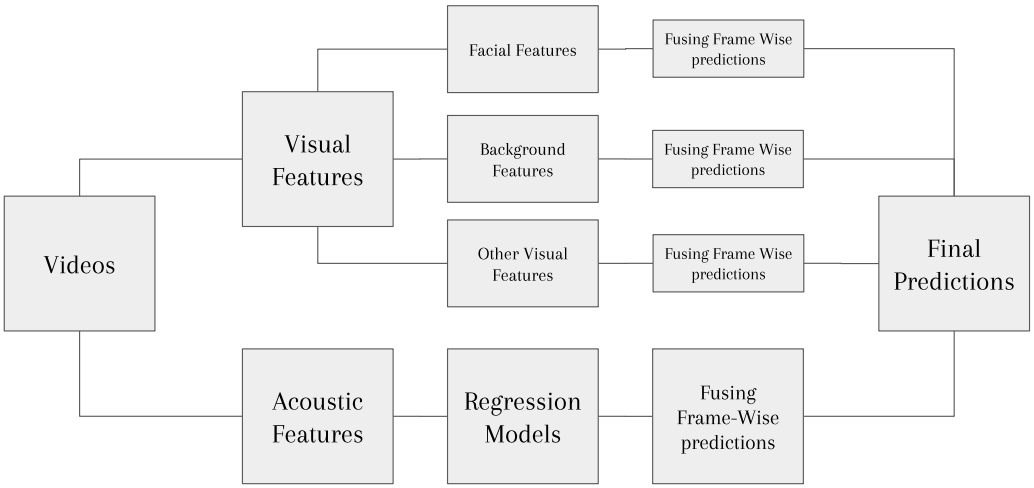

Fig. 1. Overview of our model. We extract the two major modalities present in the original videos i.e. the visual and audio modality. After learning multiple models based on different types of features, we fuse their final predictions by performing late fusion.

We take each frame one at a time and extract a smoothed facial bounding box. A detailed description of the extraction scheme is given later. We use this image as our input to a CNN trained to predict the personality-trait scores for that frame.

2) *Background Information:* We also experiment with models which utilize background information given in each frame. It is reasonable to expect first impressions to be affected by the background information too. For example, a person (gamer) with a lot of posters and other games in the background would most likely give a different impression than a person (cook) with a kitchen in the background. We use VGG-Net pre-trained on ImageNet [19] to get background representations by sampling random crops, extracting the penultimate activations and pooling them.
3) *Facial and Background features:* Both features contain information complementary to each other. Therefore, we propose methods to utilize this complementarity. We try two variants for this; In the first one, we perform late fusion of the model predictions. The second method consists of a Multi-Layer Perceptron which takes the mid layer activations of the previous models as input. The activations from both models are simply concatenated. The first variant leads to a minor increase in performance while the second method results in poor performance than the original models.

We use Convolutional Neural Networks (ConvNets) for the Facial and Background features as they have been shown to be extremely effective in vision related tasks [9], [19].

The proposed Visual-Modality framework can be divided into the following parts,

1) Visual Feature extraction
2) Network Architecture and Training
3) Frame-wise Prediction
4) Aggregating frame-wise predictions

The last stage is shared with other models as well and is therefore discussed at the end of the section.

Visual Feature Extraction:

Consider the extraction of facial features. Initially, we detect faces in each frame of the video clip using the Face detector [10] provided in the DLib[1] library. We construct a box around the face after centering it appropriately after face registration. To avoid abrupt changes in the bounding box size, we perform box smoothing. It is done by computing 2-sided running averaging (window size 8). This is followed by image normalization via Contrast Limited Adaptive Histogram Equalization (CLAHE) present in OpenCV[2].

For background features, we randomly sample multiple crops of size 224x224 from each frame. We do not enforce exclusiveness between them, although that could be an effective way to include more information while simultaneously avoiding repetition. We consider the activations of the penultimate layer of VGG-net after feeding each crop into the network. Finally, we pool all the activations which gives us a 4096D representation for each frame. We also experiment with average and max pooling.

Fusing facial and background features is performed via simple concatenation of the model activations. Instead of simply concatenating the original inputs, we choose to concatenate the mid layer activations of the above two models. It allows us to avoid high memory costs and over-fitting issues.

Network Architecture and Training:

- *Facial Feature Based Models:* As mentioned earlier we use a ConvNet for facial features. We consider two models, one expects color images while the other expects grayscale ones. The ConvNet takes batches of 100x100

[1]http://dlib.net/
[2]http://opencv.org/

images as input while performing random cropping and resizing at each epoch. The images are then randomly flipped horizontally with a probability of 0.5. These simple methods allows us to extract more data from the small training set and also avoid over-fitting.
The first ConvNet architecture which takes RGB images has 5 stages. The initial two stages include a Conv. layer containing 32 3x3 filters followed by a max pooling layer. The next two stages contain Conv. layers with 64 3x3 filters followed by max pooling. The last stage consists of a fully connected layer with 256 hidden states connected to five units, further followed by softmax representing the trait scores. The activation function used in the network is rectified linear unit (ReLU).
The other ConvNet architecture which takes grayscale images also has 5 stages containing different layers. The initial two stages include a Conv layer with 48 3x3 filters followed by a max-pooling layer. The next two stages consist of Conv layers with 96 3x3 filters followed by avg-pooling. The rest remains the same.

- *Background Features:* As discussed earlier, we compute 4096D representations for each frame using the penultimate activations of VGG-Net. The final model is a fully connected NN with a single hidden layer containing 128 nodes. This layer also acts as our representation for the feature-fusion model discussed later. For feature extraction from each frame, we sample 20 random crops. During training, we randomly choose 10 crops out of the set computed in the previous stage. We then pool them appropriately to arrive at the background features for each frame. This enables us to get large number of training samples also preventing over-fitting.
- *Fusing Facial and Background features:* We fuse the mid layer activations of both the models and perform prediction using them as features. We experiment with a simple linear classifier but the results are disappointing. The poor performance may be due to mid layer activations not being representative or poor model hyper-parameters.

Frame-wise Prediction:
We use the above discussed models to compute the frame-level predictions. Results for frame-level performance have been provided in section V. The best performing model is the Grayscale Image CNN, which could be due its higher complexity. Feature extraction for testing is performed in the same way as discussed above.

B. Acoustic Feature Based Models

We first consider the extraction of audio features which is a crucial component of the pipeline. Good audio features have helped improve performance in numerous recognition and classification tasks. We use FFmpeg[3] for extracting audios from the original videos. Audio features are extracted using the openSMILE [4] framework. The features we used are the same as the one used in INTERSPEECH 2010 Paralinguistic Challenge [16]. The set contains 1582 features which result from a base of 34 low-level descriptors (LLD) with 34 corresponding delta coefficients appended, and 21 functionals applied to each of these 68 LLD contours (1428 features). Please refer to [16] for more details.

Initially, we consider each audio clip (Around 15s in duration) as a training sample. But the results were not very promising. Therefore, we adopt another approach in which we extract multiple audio segments (2-3 seconds) from each clip. This gives us an increase in the number of training samples. It also allows us to use segment-wise cues and use them to predict the final scores. The corresponding results are considerably better than the earlier method.

We also experiment with multiple pooling methods such as min-max pooling (Concatenating the min and max pools) and average pooling. We experiment with multiple algorithms (Such as SVRs, Random Forests, etc) for using these features, finally Ensembles of Decision Tree Regressors give the best results. As mentioned earlier, we compute segment wise predictions which are further used for the final inference.

C. Alternative Model for Traits

Correlation	Extra.	Agree.	Cons.	Neuro.	Open.
Extra.	1.00	0.710	0.589	0.796	0.774
Agree.	0.711	1.00	0.658	0.733	0.649
Cons.	0.589	0.658	1.000	0.720	0.583
Neuro.	0.796	0.733	0.720	1.00	0.774
Open.	0.774	0.649	0.583	0.774	1.000

TABLE II: Correlation of the Big-5 traits

We also propose another model for prediction of personality traits. There are many studies claiming that the Big-Five personality traits aren't independent of each other. This claim is also supported by the correlation matrix shown in table 2. As we can see, all traits are well correlated with each other. Therefore, motivated by this we propose the following: Suppose we quantify the existence/strength of a particular trait, then the value will actually be a combination of two components i.e. a global component and a trait specific component. Using this model, we aim at learning each trait specific model using a global component along with a trait specific component. To predict five personality traits, instead of creating five trait specific models. We create six models, where one represents the global component and the other five represent the trait-specific components. The global-component is trained to predict the average of all the traits. While the trait specific components learn to output the deviation of the trait score from the average. Effectively,
$trait\_score = global\_score + trait\_specific\_score$

We refer to this model as model **X**. We have not performed extensive tests using this model, but we show results for audio features in section V.

[3]http://ffmpeg.org/
[4]http://audeering.com/research/opensmile/

D. Fusing Frame-Wise Predictions

As discussed earlier, we perform aggregation of frame-wise predictions to arrive at fixed length descriptions of each video [9]. This representation is further used for training another model which gives us the final predictions. Table IV in Section V shows that such fusion leads to a significant increase in performance. For our experiments, we decide to consider 15 frames and concatenate their cues to arrive at a representation for each video clip.

Note that there are multiple issues at arriving at a fixed length vector. There may be videos which have very few useful frames, while some videos may have more than the desired number (15) of frames. We adopt the approach in [9] and tackle these problems by using expansion and averaging which are described below,

- *Frame Averaging:* For long videos which have more than the required number of frames, we average the probability vectors of randomly chosen blocks of frames. Effectively contracting the video to fit into the 15-frame representations.
- *Frame Expansion:* For videos which contain lower number of useful frames than required, we create copies of randomly selected frames till we get 15 frames. Effectively expanding the video to fit the desired size.

We explore the following strategies for arriving at video level descriptors,

1) *Trait specific Representations*: For each trait, only consider the corresponding scores for each frame and concatenate only these. Therefore, if we choose to combine 15 frames then the resultant vector will also be 15-dimensional. Note that we get different vectors for different traits.
2) *Sorted trait scores*: Another possible method is a minor variant of the method discussed above. We still take the related scores but now sort them and provide the sorted vector as input. The motivation being that the position of the frames is not relevant in such a score based model. Therefore we sort the predictions and feed this as input. There was a slight improvement in the results.
3) *Take-it-all*: In this approach, instead of taking only one trait we consider all the traits. This method allows the model to use the inter trait relationships for predictions. Thus for each trait, we get a 75 dimensional vector from 15 frames. This performs considerably better than the methods discussed before.

E. Fusing Model Predictions

In this section, we discuss the methods used in the final stage of our framework i.e. fusing predictions from all models. The methods we explore are given below,

1) Averaging: The first method involves simple averaging of the predictions of all models, without assigning importance to any specific model. Even though we don't penalize poor performing models, this leads to a significant improvement in performance.

2) Random Weight Search: A simple improvement to the average method can be the inclusion of per-trait-per-model weights. We can expect different models to have varying performances across different traits. We include weights to represent this trait-model confidence. Finally, we have $5\text{x}numModels$ weights to learn. Recent work has shown that random search for hyper-parameter optimization can be an effective strategy [2], [9], even when the number of hyper-parameters is moderate. Final prediction is simply the weighted sum of the predictions. This further improves the model performance.

3) Regression over Predictions: Our best performing method involves treating the fusion task as a regression problem. Where, the base-model predictions are the inputs and the ground truths are the target values. This also allows the model to assign negative weights to a few models. The results are given in section V.

V. Results

A. Evaluation Criteria

For evaluation, given a video and the corresponding traits values, the accuracy is computed as one minus the absolute distance between the predicted and true values. The mean accuracy for the Big-Five traits [7] is computed as,

$$\frac{1}{5N}\sum_{i=1}^{5}\sum_{i=1}^{N} 1 - |true\_value_{i,j} - predicted\_value_{i,j}|$$

where N is the total number of testing samples.

B. Performance Evaluation

We first discuss the segment-level performance of our proposed base models. The results are summarized in table 3. The segment-level performance is computed using a held back subset of the training data.

Legend - GrayCNN: Gray-Image CNN, ColorCNN: Color Image CNN, ModelXAudioAvg: Model X applied with audio features (After Avg pooling), AudioMinMax: Audio features with min-max pooling, AudioOrig: Audio features without pooling, the rest are self explanatory. The columns table III, IV and V represent the Big-Five traits. Due to space constraints, only the first character of each trait is used to mention them.

We observe that models based on visual modality consistently perform poor compared to the audio based ones. This may be due to the usage of weaker models while using visual features. It might also be due to the fact that audio segments contain more temporal information compared to individual frames. A possible way to incorporate temporal information is to take blocks of frames and use 3D convolutions over them. As we mentioned earlier, the fusion of background and facial features is not able to outperform the base models. Giving rise to the possibility of either poor intermediate representations or poor hyper-parameters. We also notice that our proposed Model X performs slightly better than its counterparts. Stricter experiments are required to confirm the effectiveness of Model X in general. It is also worth while to note that although

Models	Score	Models	Score
ColorCNN	0.8893	**GrayCNN**	0.8947
BackgroundFet	0.8895	**FuseGrayBG**	0.8895
ModelXAudioAvg	**0.8970**	**AudioAvg**	0.8964
AudioMinMax	0.8959	**AudioOrig**	0.8916

TABLE III: Frame/Segment-Wise prediction scores (Avg.)

Models	Avg.	E.	A.	C.	N.	O.
ColorCNN	0.895	0.893	0.905	0.890	0.891	0.896
GrayCNN	0.899	0.900	0.907	0.894	0.896	0.900
AudioAvg	**0.902**	**0.900**	**0.908**	**0.895**	**0.901**	**0.904**

TABLE IV: Video-Level Prediction Scores. Performance computed on the validation set provided in the 1^{st} round

the frame level scores are not very promising, video level performance is much better.

Table 4 gives results for the video level predictions after fusion, as we can see frame-level fusion results in a significant improvement in performance. The audio models continue to perform better than the visual. This may be attributed to the higher performance in the early stages. We can see that our model performs best for Agreeableness and Openness.

Finally, table 5 contains the performance of the model combinations i.e. the complete model. As we can see, random weights are outperformed by the regressor based fusion. Rand-X refers to random search fusion of X base models, Reg-X refers to regression based fusion of X base models.

Models	Avg.	E.	A.	C.	N.	O.
Rand-4	0.9036	0.9027	0.9096	0.8981	0.9018	0.9058
Rand-6	0.9043	0.9040	0.9100	0.8994	0.9024	0.9060
Reg-6	0.9074	0.9075	0.9123	0.9010	0.9067	0.9094
Reg-7	**0.9082**	**0.9077**	**0.9125**	**0.9034**	**0.9073**	**0.9100**

TABLE V: Full model comparison. Performance computed on the validation set provided in the 1^{st} round of the challenge

VI. Conclusions and Future Work

In this paper, we propose models for apparent personality analysis which exploit both visual and audio modalities present in videos. We also show that combining cues from both results in much better performance compared to the individual models. Our model secured 3rd place in the final evaluation phase of the Apparent Personality Analysis track at the Chalearn LAP challenge, held in association with ICPR 2016. A notable feature of our framework is its low complexity, all base models have very low number of parameters (All of them contain less that 0.6 million parameters) considering the enormous vision and audio models used currently. Consequently, training times and computation costs are extremely low.

As future work, we would like to study the effect of including more complex models in our framework. We would also like to explore integrating Model X with other modalities and architectures.

References

[1] Noura Al Moubayed, Yolanda Vazquez-Alvarez, Alex McKay, and Alessandro Vinciarelli. Face-based automatic personality perception. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1153–1156. ACM, 2014.

[2] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305, 2012.

[3] Justin M Carré, Cheryl M McCormick, and Catherine J Mondloch. Facial structure is a reliable cue of aggressive behavior. *Psychological Science*, 20(10):1194–1198, 2009.

[4] John M Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990.

[5] Michael P Haselhuhn and Elaine M Wong. Bad to the bone: facial structure predicts unethical behaviour. *Proceedings of the Royal Society of London B: Biological Sciences*, page rspb20111193, 2011.

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

[7] Jair-Escalante Hugo, Ponce-Lopez Victor, Wan Jun, Riegler Michael, Chen Baiyu, Clapes Albert, Escalera Sergio, Guyon Isabelle, Baro Xavier, Halvorsen Paal, Muller Henning, and Larson Martha. Chalearn joint contest on multimedia challenges beyond visual analysis: An overview. In *Proceedings of the ICPR Contest*, 2016.

[8] Timothy A Judge, Chad A Higgins, Carl J Thoresen, and Murray R Barrick. The big five personality traits, general mental ability, and career success across the life span. *Personnel psychology*, 52(3):621–652, 1999.

[9] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, pages 1–13, 2015.

[10] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1867–1874, 2014.

[11] Leqi Liu, Daniel Preotiuc-Pietro, Zahra Riahi Samani, Mohsen E Moghaddam, and Lyle Ungar. Analyzing personality through social media profile picture choice. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[12] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[13] Gelareh Mohammadi and Alessandro Vinciarelli. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing*, 3(3):273–284, 2012.

[14] George Papandreou, Liang-Chieh Chen, Kevin Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. *arXiv preprint arXiv:1502.02734*, 2015.

[15] Rizhen Qin, Wei Gao, Huarong Xu, and Zhanyi Hu. Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. *arXiv preprint arXiv:1604.07499*, 2016.

[16] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A Müller, Shrikanth S Narayanan, et al. The interspeech 2010 paralinguistic challenge. In *InterSpeech*, volume 2010, pages 2795–2798, 2010.

[17] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[18] Phillip R Shaver and Kelly A Brennan. Attachment styles and the" big five" personality traits: Their connections with each other and with romantic relationship outcomes. *Personality and Social Psychology Bulletin*, 18(5):536–545, 1992.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.