

Bi-Modal Regression for Apparent Personality Trait Recognition

Nishant Rai

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur

Workshop on Multimedia Challenges beyond Visual Analysis

ICPR '16, Cancun, Mexico

INTRODUCTION

- The task of the ChaLearn First Impressions Challenge is to quantify personality traits of users in short video sequences (Around 15s).
- We present our approach for predicting traits by combining multiple modality specific models.
- Our models include Deep Networks focusing on leveraging visual information in faces, networks focusing on supplementary background information and models using acoustic features.

- **Acoustic feature based Recognition:** In recent years, multiple audio representations have been suggested. Features such as MFCC [12] have given good results in tasks related to Emotion Recognition and Age/Gender estimation.
- **Deep Learning for Visual Features:** Deep Learning methods have attracted huge interest due to recent models achieving state of the art performance in both image and audio related tasks.
- **Automatic Personality Perception:** Personality analysis has been an area of vast research in the domain of psychology [5], [3], [10]. There have been multiple works exhibiting the effect of visual cues on formation of first impressions [1], [8].

METHODOLOGY

Reasonable to assume that first impressions depend on,

- Visual Information, such as,
 - Facial features
 - Facial emotions
 - Background Information
- Audio Information

Create modality specific models, which are later combined for the final prediction.

MODEL OVERVIEW

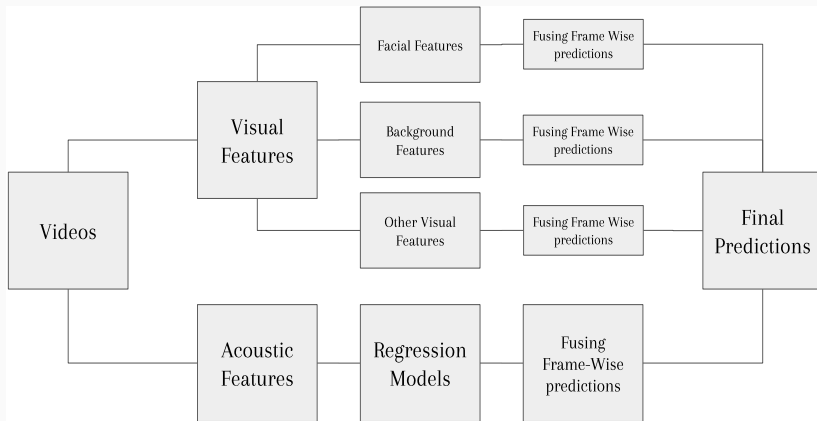


Figure: OVERVIEW OF OUR MODEL

VISUAL FEATURE BASED MODELS

- Based on the previous discussions, the features we explore are facial features and background information.
 - **Facial Features:** Take each frame one at a time and extract a smoothed facial bounding box. Used as inputs to a CNN based model.
 - **Background Information:** Use VGG-Net pre-trained on ImageNet [19] to get background representations by sampling random crops, extracting the penultimate activations and pooling them.
- The proposed framework can be divided into the following,
 - Visual Feature extraction
 - Network Architecture and Training
 - Frame-wise Prediction
 - Aggregating frame-wise predictions

VISUAL FEATURE EXTRACTION

- Facial Feature Extraction:
 - We detect faces in each frame of the video clip using the Face detector [7] provided in the DLIB library.
 - We then construct a box around the face after centering it appropriately after face registration. Also perform smoothing on the box coordinates¹.
 - This is followed by image normalization via Contrast Limited Adaptive Histogram Equalization (CLAHE) present in OpenCV.
- Background Feature Extraction:
 - We randomly sample multiple crops of size 224x224 from each frame.
 - We then 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 vector for each frame.

¹Details present in the paper

CLAHE NORMALIZATION



Figure: CLAHE NORMALIZATION ON FACES

- Facial Feature based Model:
 - As mentioned earlier we use a ConvNet for facial features.
 - The ConvNet takes batches of 100x100 images as input while performing random cropping, flipping and resizing at each epoch.
 - We experiment with both RGB and grayscale inputs. There were no significant differences¹.
- Background Feature based Model:
 - 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.
 - For feature extraction from each frame, we sample 20 random crops. While 10 random crops are chosen during training¹.

¹Details present in the paper

ACOUSTIC FEATURE BASED MODELS

- We use FFMPEG¹ for extracting audios from the original videos. Audio features¹ are extracted using the openSMILE ² framework.
- We extract multiple audio segments (2-3 seconds) from each clip, which gives us an increase in the number of training samples.
- Also allows us to use segment-wise cues to predict the final scores. Results much better than treating them as single samples.
- Experiment with various pooling approaches and regression algorithms. Finally ensembles of Decision Tree Regressors give the best results.

¹<http://ffmpeg.org/>

²<http://audeering.com/research/opensmile/>

¹The features we used are the same as the one used in INTERSPEECH 2010 Paralinguistic Challenge [9]

PREDICTION AGGREGATION

FUSING FRAME-WISE PREDICTIONS

- We perform aggregation of frame-wise predictions to arrive at fixed length video descriptions [6]. It is further used for computing the final predictions.
- We adopt the approach in [6] and tackle these problems by using expansion and averaging which are described below,
 - **Frame Averaging:** For long videos with high number of useful frames, we average the probability vectors of randomly chosen blocks of frames.
 - **Frame Expansion:** For short videos which contain low number of useful frames, we create copies of randomly selected frames till we get 15 frames.

FRAME AVERAGING

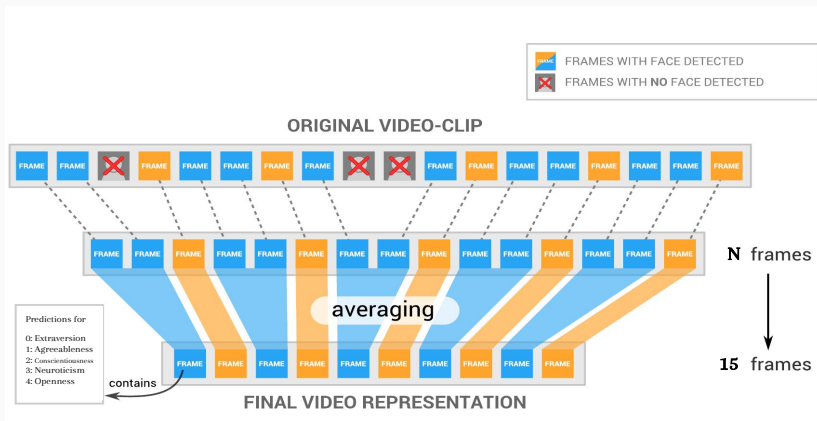


Figure: FRAME AVERAGING, FIGURE TAKEN FROM [6]

VIDEO LEVEL DESCRIPTORS

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

- **Trait specific Representations:** For each trait, only consider the corresponding scores for each frame and concatenate only these.
- **Sorted trait scores:** The trait score are now sorted, we provide the sorted vector as input. The motivation being that the position of the frames is not relevant in such a score based model. There was a slight improvement in the results.
- **Take-it-all:** Instead of taking only one trait we consider all the traits. This method allows the model to use the inter trait relationships for predictions. This performs considerably better.

FUSING MODEL PREDICTIONS

The methods explored for prediction fusion are described below,

- **Averaging:** Simple averaging of the predictions of all models, without assigning importance to any specific model.
- **Random Weight Search:** Weighted averaging of predictions. We include weights to represent this trait-model confidence. We use random search for choosing the hyper-parameters [2], [6].
- **Regression over Predictions:** We treat the fusion task as a regression problem. Where, the base-model predictions are the inputs and the ground truths are the target values.

RESULTS

DATASET DESCRIPTION

- Newly constructed dataset consisting of short video clips collected from YouTube
- The video clips contain a main subject, special care has been taken to ensure a unique major subject.
- Most of the faces are frontal and there is little camera movement making 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).

EVALUATION CRITERIA

- The score is computed as one minus the absolute distance between the predicted and true values.
- The mean accuracy for the Big-Five traits [4] is computed as,

$$\frac{1}{5N} \sum_{i=1}^5 \sum_{j=1}^N 1 - |true_value_{i,j} - predicted_value_{i,j}|$$

where N is the total number of testing samples.

QUANTITATIVE RESULTS¹

Models	Score	Models	Score
ColorCNN	0.8893	GrayCNN	0.8947
BackgroundFet	0.8895	FuseGrayBG	0.8895
ModelXAudioAvg ²	0.8970	AudioAvg	0.8964
AudioMinMax	0.8959	AudioOrig	0.8916

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

¹Legend - 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.

²Refer to the paper for details

QUANTITATIVE RESULTS¹

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: Video-Level Prediction Scores. Performance computed on the validation set provided in the 1st round

¹The columns in the table represent the Big-Five traits. Due to space constraints, only the first character of each trait is used to mention them.

QUANTITATIVE RESULTS¹

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: Full model comparison. Performance computed on the validation set provided in the 1st round of the challenge




¹The columns in the table represent the Big-Five traits. Due to space constraints, only the first character of each trait is used to mention them.

CONCLUSIONS




- We propose models which exploit both visual and audio modalities present in videos.
- Our model secured 3rd place in the final evaluation phase of the Apparent Personality Analysis track at the Chalearn LAP challenge.
- A notable feature of our framework is its low complexity, all base models have very low number of parameters compared to the enormous deep networks 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.

REFERENCES

-  Noura Al Moubayed et al. “Face-based automatic personality perception”. In: *Proceedings of the 22nd ACM international conference on Multimedia*. ACM. 2014, pp. 1153–1156.
-  James Bergstra and Yoshua Bengio. “Random search for hyper-parameter optimization”. In: *Journal of Machine Learning Research* 13.Feb (2012), pp. 281–305.
-  John M Digman. “Personality structure: Emergence of the five-factor model”. In: *Annual review of psychology* 41.1 (1990), pp. 417–440.

-  Jair-Escalante Hugo et al. “ChaLearn Joint Contest on Multimedia Challenges Beyond Visual Analysis: An overview”. In: *Proceedings of the ICPR Contest*. 2016.
-  Timothy A Judge et al. “The big five personality traits, general mental ability, and career success across the life span”. In: *Personnel psychology* 52.3 (1999), pp. 621–652.
-  Samira Ebrahimi Kahou et al. “Emonets: Multimodal deep learning approaches for emotion recognition in video”. In: *Journal on Multimodal User Interfaces* (2015), pp. 1–13.

REFERENCES III

-  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*. 2014, pp. 1867–1874.
-  Rizhen Qin et al. “Modern Physiognomy: An Investigation on Predicting Personality Traits and Intelligence from the Human Face”. In: *arXiv preprint arXiv:1604.07499* (2016).
-  Björn Schuller et al. “The INTERSPEECH 2010 paralinguistic challenge.” In: *InterSpeech*. Vol. 2010. 2010, pp. 2795–2798.



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". In: *Personality and Social Psychology Bulletin* 18.5 (1992), pp. 536–545.