A Multi-modal Personality Prediction System

Chanchal Suman[1], Aditya Gupta[2], Sriparna Saha[1], and Pushpak Bhattacharyya[1]

[1]Department of Computer Science & Engineering, Indian Institute of Technology Patna, India
[2]Department of Electrical Engineering, Indian Institute of Technology Patna, India
email: {1821cs11, aditya.ee17, sriparna}@iitp.ac.in, pushpakbh@gmail.com

Abstract

Automatic prediction of personality traits has many real-life applications, e.g., in forensics, recommender systems, personalized services etc.. In this work, we have proposed a solution framework for solving the problem of predicting the personality traits of a user from videos. Ambient, facial and the audio features are extracted from the video of the user. These features are used for the final output prediction. The visual and audio modalities are combined in two different ways: averaging of predictions obtained from the individual modalities, and concatenation of features in multi-modal setting. The dataset released in Chalearn-16 is used for evaluating the performance of the system. Experimental results illustrate that it is possible to obtain better performance with a hand full of images, rather than using all the images present in the video.

1 Introduction

Our personality impacts a lot on our lives, affecting our life choices, mental health, well-being, and desires. Thus, automatic prediction of one's personality has many applications such as enhanced personal assistants, recommender system, job screening, forensics, psychological studies, etc. (Mehta et al., 2019). The big-five personality traits (Digman, 1990) are the most popular measures used in the literature. They are Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness.

Most of the methods have utilized the CNN-based architectures for extracting the features from the images, mostly the facial features (Güçlütürk et al., 2016),(Gürpınar et al., 2016),(Biel et al., 2012). Mainly, researchers have used the late fusion strategy (averaging the predictions from all the modalities) for combining the results obtained from different modalities (audio, and video) (Güçlütürk et al., 2016, 2017; Gürpınar et al., 2016; Wei et al., 2017; Pianesi et al., 2008; André et al., 1999). There are very few works, which have employed early fusion strategy for developing the multimodal system (Yang et al., 2017; Kampman et al., 2018).

This has motivated us to develop a multimodal system which uses early fusion for combining the features generated from different modalities, and using the combined feature for final prediction. We have developed an audio-visual system, which extracts ambient features from video using ResNet (He et al., 2016), facial features using MTCNN (Zhang et al., 2016), and the audio features from the VGGish CNN (Hershey et al., 2017). Finally, those features are concatenated and then fed to the fully connected layer followed by a sigmoid layer for the final prediction. We have used the Chalearn-16 dataset for evaluating the performance of our system (Güçlütürk et al., 2016). An accuracy of 91.43% on the test data has been achieved using our proposed system.

The main contributions of this work are i) to the best of our knowledge, combined feature representation of images has been carried out for the first time for personality prediction. ii) The VGGish CNN has been used for extracting the audio features, and then those are used for the prediction. It has also been performed for the first time. From the results, it can be established that only some of the images extracted from different parts of the video are capable for successful training and testing of the model with good performance.

2 The Proposed Methodology

Our methodology consists of learning some features extracted from two different modalities, namely video and audio. They are discussed below:

2.1 Visual Modality

We have extracted two types of features from the visual modality, i) ambient, and ii) facial. Using

Proceedings of the 17th International Conference on Natural Language Processing, pages 317–322
Patna, India, December 18 - 21, 2020.

these two features, we have developed two different systems namely Amb-visual, and Fac-visual, respectively. Both of the architectures consist of three sub-parts, first is pre-processing step, second is the CNN architecture and the third is the final step to combine the features of the images to predict the big-five personality traits.

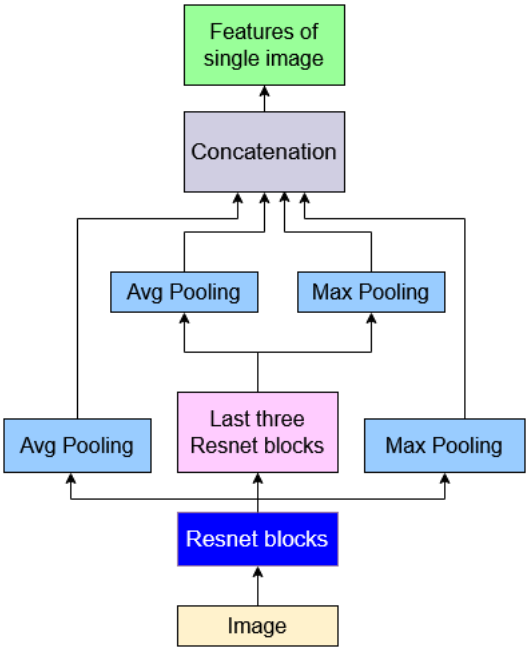

Figure 1: Extraction of Features from a Single Image using Deep Feature Extractor.

2.1.1 Preprocessing

The pre-processing of the data is important for extracting and learning appropriate features. The first step was to extract the images from the videos using FFmpeg tool in RGB color frame, and then resizing into 224x224 pixels.

i) Extraction of ambient features: we have extracted six equally spaced images to represent the whole video thus required RAM has decreased to around 20 Gb.

ii) Extraction of facial features: For the facial modality we first detect and align faces using Multi Task CNN (MTCNN) (Zhang et al., 2016). Again as before, only six images at equal intervals were considered from the set of images having faces.

2.1.2 Model Architecture

Here, we have discussed the model architectures for the Amb-visual, and the Fac-visual.

- We experimented with different CNN architectures and found that ResNet (He et al., 2016) with 101 layers worked best for extracting features from images. The feature vectors from last three layers are considered as the latent features of the image. For better and robust feature extraction instead of performing just max-pooling, we have performed both max and average pooling after the last convolution layer. Just like skip connection, these two feature vectors are concatenated to the upper feature vector which is extracted from the final ResNet block. This is named as deep feature extractor and it is depicted in figure 1.

- The model architecture for extracting features from the Fac-visual is same, as the Amb-visual. The only difference is the usage of MTCNN (Zhang et al., 2016) for extracting faces before passing it into deep feature extractor.

Apart from the feature extraction part, all the other components are same for the Amb-visual and the Fac-visual architecture.

2.1.3 Final classifier

The representations, learnt from the deep feature extractor module are fed to the final classification layer for the output prediction. We experimented in three settings, for analysing the behaviour of the system.

M1: In the first method, all six images were labeled with their corresponding video's numbers. The previously extracted features of each image are passed to a final fully connected layer with sigmoid as the activation function. This layer gives us the output values between 0, and 1 for the big five personality traits. The loss value achieved for each of the images is added to the final loss for training. Finally, the trait values for the video are obtained by considering the mean values of six images.

M2: In the second method, we concatenated the features of six images as a final feature vector, representing the video's visual feature. This feature vector is then passed to the fully connected and sigmoid layer for getting the final trait values of the video. In this method, loss of each image is not considered separately.

M3: For the third method, we try to use the fact that a video is a time-series data. By using an LSTM (Hochreiter and Schmidhuber, 1997), we wanted to learn more better features so that image at time 't' could be represented using information from previous time steps as well. For that,

we passed the extracted features into an LSTM of appropriate hidden dimension with several layers. After that, the outputs of LSTM for different time-steps are collected, and then concatenated for getting the final feature vector of a video. This feature vector is then passed to a fully connected + sigmoid layer to extract trait values for a video.

2.2 Audio Modality

For the audio modality, the VGGish CNN (Hershey et al., 2017), along with below mentioned pre-processing steps are used. The pre-processing and architecture are explained in the following section.

2.2.1 Preprocessing

Firstly, all audios are re-sampled to 16 kHz mono. A spectrogram is computed using magnitudes of the Short-Time Fourier Transform with a window size of 25 ms, a window hop of 10 ms, and a periodic Hann window. A mel spectrogram is computed by mapping the above spectrogram to 64 mel bins covering the range 125-7500 Hz. Then log of mel spectrogram is computed with a small offset of 0.01 to stabilize mel spectrogram values. These features are then framed into non-overlapping examples of 0.96 second, where each example covers 64 mel bands and 96 frames of 10 ms each.

2.2.2 Model Architecture

After the pre-processing, a 2d feature array of shape 96x64 for each 1 second was obtained. Thus, for a 15 second video there are 15 such feature vectors. This feature vector is then passed to the VGGish CNN architecture, that has many 2d convolution layers. This CNN outputs a 128 length embedding for each second, that was further used to train a classifier for getting traits. Like before, we initialised the weights of our convolution filters with the pre-trained weights of VGGish CNN trained on large Youtube dataset for warm start of training.

After getting audio features, for each second of the video we have fifteen 128 length vectors, two methods were experimented for further combining the features for regression.

Audio-M1: In the first method, these features are passed into few layers of LSTMs of appropriate hidden layers and then the outputs of LSTM are concatenated to get a final feature vector of whole audio. Then a fully connected layer with sigmoid activation function is applied.

Audio-M2: In the second method, instead of using LSTMs, we simply concatenate the audio features and then apply a fully connected layer and sigmoid. This network is called deep audio feature extractor.

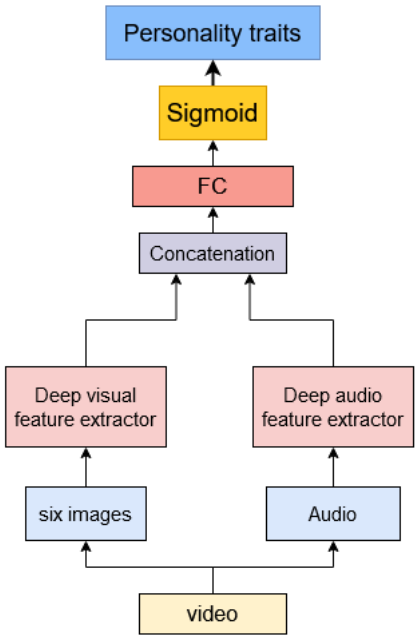

Figure 2: The Proposed Multi-modal System

2.3 Combining Modalities

We experimented with two approaches for combining the modalities. They are: 1) Average: we simply took the average of predicted trait values of different modalities. ii) Concat: we concatenated the obtained features of different modalities and then applied fully connected layer with sigmoid. It is depicted in Figure 2.

Table 1: Performance of the Proposed Model on Validation Data

Modality	Accuracy (in %)
Audio-M1	90.29
Audio-M2	90.64
Amb-visual (M1)	91.27
Amb-visual (M2)	91.19
Amb-visual (M3)	90.65
Fac-visual (M1)	90.90
Fac-visual (M3)	90.51
Amb-visual+audio (Concat)	91.44
Amb-visual+audio (Average)	91.56
Amb-visual+Fac-visual+audio (Average)	91.62

Table 2: Class-wise accuracy (in %) values of the proposed model on test data

	Average	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Ours (Fusion(average))	91.43	91.53	91.29	92.06	91.18	91.07
PersEmoN (Zhang et al., 2019)	91.7	92.0	91.4	92.1	91.4	91.5
NJU-LAMDA (Wei et al., 2017)	91.3	91.3	91.3	91.7	91.0	91.2
evolgen (Subramaniam et al., 2016)	91.2	91.5	91.2	91.2	91.0	91.2
DCC(Güçlütürk et al., 2016)	91.1	91.1	91.0	91.4	90.9	91.1
ucas (**?**)	91.0	91.3	90.9	91.1	90.6	91.0

3 Results and Comparison

In this section, we have discussed the results obtained for our proposed approach. We have compared these results with the existing works too.

3.1 Results and Discussion

The ECCV ChaLearn LAP 2016 data, is used for experimentation, having $10,000$ videos. Out Of the $10,000$ videos, $6,000$ videos are used for training phase and $2,000$ videos for both validation and testing (Wei et al., 2017). The accuracy is used as the performance measure, for evaluating the proposed system.

In all the modalities, models without LSTM layers performed better than the ones with them as shown in table 1. Validation accuracies with LSTM are 90.65, 90.51, 90.29 for Amb-visual, Fac-visual, and the Audio modality, respectively, whereas the corresponding accuracies without LSTM are 91.27 , 90.90 and 90.64, respectively. In the Amb-visual, and the Fac-visual, the highest validation accuracies were achieved by using average and max pooling on ultimate and penultimate ResNet layers and concatenating these features of six images (M2). For M1 (averaging the prediction of all the six images), we achieved an accuracy of 91.19 for Amb-visual. We didn't evaluate the M1 for Fac-visual, as the accuracy of M1 for Amb-visual is lesser than M2.

Similar to the Visual modality, the method based on LSTM layers (Audio-M1), proved disadvantageous as validation accuracy with LSTM layers is 90.29 and without LSTM (Audio-M2) is 90.64.

Table 3: Comparison with Other Works

Modality	**Our Method**	NJU-LAMDA (Wei et al., 2017)	PersEmoN (Zhang et al., 2019)
Visual	91.13	91.16	**91.7**
Audio	**90.16**	89.50	–
Video+ Audio (Avg)	**91.43**	91.30	–

Since, audio is a time series data, LSTMs should have increased the accuracy. But this is not observed in the obtained results. It proves that LSTM was futile and has only led to overfitting.

We have applied two different approaches for combining the two modalities, i) Average, and ii) Concat. By averaging the two best performing modalities (Amb-visual, and audio), a validation accuracy of 91.56% is attained. Concatenation of the features generated from different modalities (Amb-visual, audio) resulted in a validation accuracy of 91.44%. After that, we calculated the performance using averaging of predictions of all the three modalities, and achieved an accuracy of 91.62%. Average accuracies of 91.13%, 90.16%, and 91.43% are achieved using the best models for video(Amb-visual), audio(Audio-M2), and the fusion (Average), respectively on test data.

3.2 Comparison with Other Works

The best performing system on the Chalearn-16 dataset is developed by the (Zhang et al., 2019).

Emotion and personality, both features are fused together in (Zhang et al., 2019), for analysing the effects of emotion on personality prediction. The methodology developed by (Wei et al., 2017) is the second best performing work.

We tried two different approaches for combining the visual and the audio features, averaging and the concatenation. The averaging of the predictions generated by three modalities has attained an accuracy of 91.43%. The detailed class-wise accuracy for each of the class and the comparative results are shown in table 2, and 3 respectively. It can be seen that, our proposed approach (video+audio(average)) attains better performance than the method proposed in (Wei et al., 2017). We have achieved an accuracy of 91.43% for the visual modality, while 91.30% is reported by (Wei et al., 2017). This shows that, only handful of images extracted from different parts of the video are enough for successful training and testing with good performance. The researchers in (Wei et al., 2017), used 100 images for making the visual system, while we have extracted only 6 images. For the audio modality, an increment of 1.14% is attained with respect to the existing one. But our developed system, could not outperform the performance of the multi-task based methodology. The reason can be, the incorporation of emotion features in their model.

From the obtained results, it can be concluded that multi-modality helps as the concatenation of features extracted from visual and audio improves the accuracy in comparison to the single one.

4 Conclusion and Future Work

Personality prediction reveals the overall characteristics of a user. In this work, we have proposed a deep multi-modal system for personality prediction given a video. It extracts features from a video, and those are then used in the neural network setting for the final prediction. From the obtained experimental results, we can conclude the following : i) only handful of images from different parts of video are enough for successful training and testing with good performance, ii) averaging the predictions of different modalities yields better performance than the simple concatenation of the modalities in the multi-modal setting.

We are planning to improve the fusion strategy for combining the different modalities. We will try to use emotion features and different types of attention mechanisms like weighted attention, self-attention etc. for combining the modalities.

Acknowledgments

Sriparna Saha would like to acknowledge the support of SERB WOMEN IN EXCELLENCE AWARD 2018 for conducting this research. This research is also supported by Ministry of Electronics and Information Technology, Government of India.

References

Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist. 1999. Integrating models of personality and emotions into lifelike characters. In *International Workshop on Affective Interactions*, pages 150–165. Springer.

Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. Facetube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 53–56.

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

Yağmur Güçlütürk, Umut Güçlü, Xavier Baro, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel AJ Van Gerven, and Rob Van Lier. 2017. Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, 9(3):316–329.

Yağmur Güçlütürk, Umut Güçlü, Marcel AJ van Gerven, and Rob van Lier. 2016. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European Conference on Computer Vision*, pages 349–358. Springer.

Furkan Gürpınar, Heysem Kaya, and Albert Ali Salah. 2016. Combining deep facial and ambient features for first impression estimation. In *European Conference on Computer Vision*, pages 372–385. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Onno Kampman, Elham J Barezi, Dario Bertero, and Pascale Fung. 2018. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 606–611.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. 2019. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27.

Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. 2008. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60.

Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. 2016. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European Conference on Computer Vision*, pages 337–348. Springer.

Xiu-Shen Wei, Chen-Lin Zhang, Hao Zhang, and Jianxin Wu. 2017. Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315.

Karen Yang, S Mall, and N Glaser. 2017. Prediction of personality first impressions with deep bimodal lstm.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Le Zhang, Songyou Peng, and Stefan Winkler. 2019. Persemon: A deep network for joint analysis of apparent personality, emotion and their relationship. *IEEE Transactions on Affective Computing*.