Multimodal Personality Analysis for Candidate Screening

Abhishek Phalak
Computer Science Graduate student
University of Southern California
Los Angeles, CA 90007
aphalak@usc.edu

Piyush Patil
Computer Science Graduate Student
University of Southern California
Los Angeles, CA 90007
piyushdi@usc.edu

Mohammad Soleymani
Professor
University of Southern California
Los Angeles, CA 90007
soleymani@ict.usc.edu

1. PROBLEM DEFINITION

Assessing human personality is crucial as it can provide valuable insights into an individual's internal and external states, influencing a range of daily and professional behaviors. In recent years, researchers have developed automatic personality computing approaches that predict personality traits based on non-verbal audio-visual behaviors. However, these methods often require complex and dataset-specific pre-processing steps and models, making them challenging to implement. The analysis of apparent personality from short video sequences is particularly difficult in the field of computer vision and multimedia research. Our goal is to extract meaningful information from both visual and audio modalities of videos to accurately assess personality. This proposed solution aims to assist in the candidate screening process by providing valuable insights that can support informed decision-making.

The proposed solution involves the use of deep learning techniques to extract data from both visual and audio modalities of videos. This data is then analyzed using sophisticated algorithms to accurately assess an individual's personality. The results of this analysis can be used to gain insights into an individual's temperament, behavior patterns, emotional states, and other important personality traits.

The proposed solution can be particularly useful in candidate screening processes where it can provide valuable information that can be used to make informed decisions. By accurately assessing an individual's personality, employers can gain insights into their potential fit within the company culture and identify any potential issues that may arise. This can help in reducing turnover rates and improve overall employee satisfaction.

To summarize, our solution aims to accurately evaluate an individual's personality by extracting comprehensive information from both visual and audio modalities of videos. By doing so, we can provide valuable insights to support the candidate screening process, facilitate informed decision-making, and ultimately improve employee satisfaction.

2. LITERATURE REVIEW

In a study by Subramaniam et al. [8], two deep learning models were proposed to recognize first impressions based on audio and facial images. The first model used a volumetric 3D convolutional neural network to predict personality traits, while the second model utilized an LSTM-based deep neural network to capture temporal patterns in audio and visual features. To improve recognition accuracy, the study concatenated the extracted features from both models. Notably, the study found that analyzing only a few frames instead of the entire video and mining temporal patterns in audio and visual features using a stochastic approach were critical factors in identifying first impressions. The results also demonstrated that the LSTM-based model outperformed the 3D convolutional neural network by effectively incorporating temporal relations into the prediction process.

X. -S. Wei [2] proposes a Deep Bimodal Regression framework that leverages modified neural networks to extract and identify crucial visual cues for the visual modality. Additionally, the audio modality utilizes linear regression to extract audio features. The proposed modification involves discarding fully connected layers, thereby reducing the model size and dimensionality and accelerating training. The framework predicts scores on individual modalities, and the results are combined through ensemble regression scores using both early and late fusion. While the architecture was not able to solely focus on humans and considered the background, the use of RESNET would have enable this to be achieved.

In their study, Ventura [3] conducted an in-depth analysis of why CNN models are remarkably successful in solving complex personality trait inference problems. They combined current CNN model interpretability techniques with face detection and Action Unit (AU) recognition systems to perform quantitative studies. The results revealed that facial features provide the majority of the distinctive information for personality trait inference. Moreover, the internal CNN representations primarily focus on analyzing key facial regions such as eyes, nose, and mouth.

Heysem Kaya [4] presented a method for automatic job candidate screening using video CVs. The proposed method leverages multiple modalities, including facial expressions, head pose, gaze direction, and speech, to assess the job candidate's suitability for a particular job position. They then used decision trees to provide a transparent and interpretable way of making a final decision on the candidate's suitability. The use of decision trees also allowed for the interpretation of the decision-making process, which can be useful in providing explanations to job applicants or in ensuring fairness and transparency in the recruitment process.

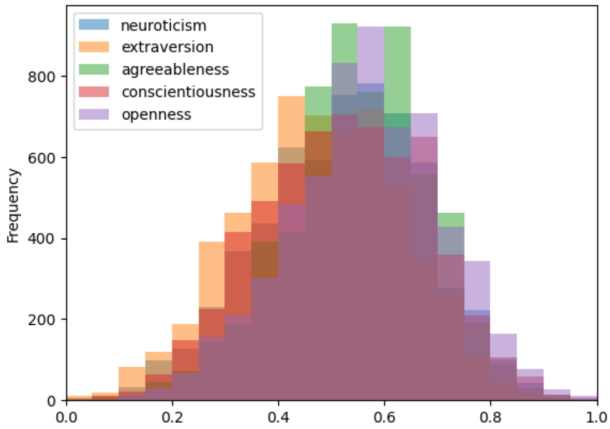
Yağmur Güçlütürk, [5] proposed a novel approach to predict personality traits using audio and visual modalities. The researchers developed an advanced audiovisual deep residual network that effectively combined features from both modalities and achieved state-of-the-art performance on a large-scale YouTube video dataset. The study showed that the visual modality was more important for predicting extraversion and openness, while the audio modality was more important for predicting neuroticism and agreeableness. The authors also analyzed the learned representations in the model and found that certain neurons in the convolutional layers were tuned to specific visual or auditory features related to personality traits.

3. DATA

For our analysis, we utilized the Chalearn First Impressions V2 dataset, which comprises over 10,000 clips extracted from more than 3,000 distinct YouTube videos. The clips feature individuals

facing and speaking in English directly to a camera, presenting a diverse range of genders, ages, nationalities, and ethnicities. This dataset was specifically designed for research on personality recognition, with a focus on first impressions.

The dataset used in this study was divided into three portions: a training set, a validation set, and a test set. The split ratio adopted was 3:1:1, with 60% of the data assigned to training, 20% to validation, and the remaining 20% to testing. Each video clip in the dataset was annotated with personality trait variables generated using Amazon Mechanical Turk (AMT). Specifically, the study focused on personality traits from the Five Factor Model, which is also referred to as the Big Five model.



The Five Factor Model is a widely-used framework in personality research that describes human personality along five dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Each video clip in our dataset has ground truth labels for these five traits, with values ranging from 0 to 1. Furthermore, each video has a transcription file in Unicode format, represented as a single dictionary. The keys in the dictionary are the names of the videos, and their corresponding values are the transcriptions of the audio in the video.

4. METHOD

4.1 Preprocessing

A. Method - I

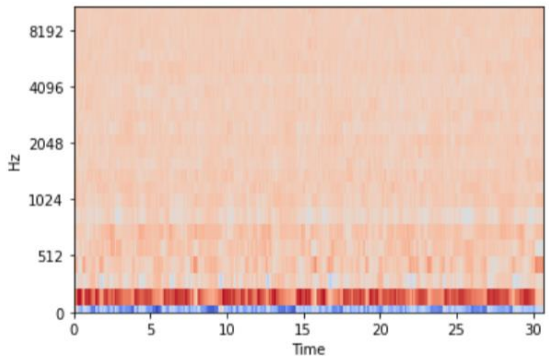
Audio preprocessing is a crucial step in preparing data for machine learning models, particularly for tasks such as speech recognition and music classification. In our approach, we extract audio data from video files using the open-source tool ffmpeg and convert it into NumPy arrays for further processing.

The first step in our audio preprocessing pipeline is computing Mel-Frequency Cepstral Coefficients (MFCC), a widely used feature representation for audio data. MFCCs capture important information about the spectral content of audio signals and have been shown to be effective in various audio classification tasks. To compute Mel Frequency Cepstral Coefficients (MFCCs), we utilized the librosa library, with a window size of 2048 samples and a hop size of 512 samples. We set n_mfcc to 24, which resulted in the calculation of 24 coefficients for each audio frame.

After computing MFCCs, we standardize the coefficients to have a mean of zero and a variance of one. Standardization helps to mitigate the effects of differences in scaling and signal intensity between audio samples and is a common preprocessing step in machine learning pipelines.

Finally, we pad and reshape the standardized MFCC coefficients to obtain a fixed size of (24, 1319). Padding ensures that all audio samples have the same length, which is necessary for feeding them into neural networks that require fixed-size inputs. Reshaping allows us to represent each audio sample as a 2D array, where each row corresponds to a different MFCC coefficient and each column corresponds to a time frame.

By following these preprocessing steps, we can convert raw audio data into a standardized, fixed-size format that can be easily fed into machine learning models. These steps are essential for ensuring that the audio data used for training and testing models is of high quality, free from biases, and can effectively capture the relevant features of the audio signals.

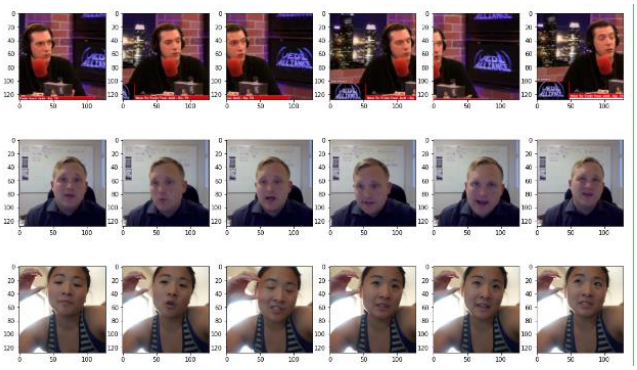


Video preprocessing is a crucial step in preparing data for machine learning models. In our approach, we extracted frames from videos in a random manner, ensuring that the selected frames are representative of the entire video content, rather than being biased towards specific segments or features.

Once the frames were extracted, we resized each RGB image to a fixed size of (248, 140) pixels, using bilinear interpolation to maintain the aspect ratio of the original image. Following the resizing process, we cropped a small window of size 128 X 128 pixels from each image, ensuring that the cropped window contained a face. In case a frame did not contain a face, it was dropped from the list of extracted frames.

The cropped RGB images were then stacked into a single NumPy array with a shape of (num_of_samples=6, 128,128,3). This stacked array represented the preprocessed video data, which was used as input to our machine learning models.

Overall, the video preprocessing steps involved random frame sampling, resizing, face detection and cropping, and stacking into a single array. These steps are crucial in ensuring that the data used for machine learning models is of high quality, representative of the entire video content.



This method had some limitations. The model could only be trained on a fixed set of images, which may not represent all variations in the video. Additionally, the model may not learn important features that were not captured in the extracted frames.

B. Method – II (Preprocessing using pretrained models)

1. Audio

The latest method of audio preprocessing uses a pre-trained model called YAMNet, which is a convolutional neural network (CNN) architecture. YAMNet has been trained on a large dataset of audio clips to learn important audio features such as speech, music, and ambient sounds. It is based on the MobileNet architecture, which is a lightweight and efficient CNN architecture that has been widely used in many computer vision and audio tasks.

YAMNet has been trained on a dataset called AudioSet, which consists of over 2 million audio clips of 527 different sound classes. The audio clips in this dataset vary in duration, background noise, and acoustic conditions, which makes it an ideal dataset for training a model to recognize audio under different conditions.

The input audio is preprocessed to fit the input shape required by the YAMNet model using librosa. This includes resampling the audio to 44100 Hz and extracting the Mel-spectrogram. The Mel-spectrogram is then passed through the YAMNet model to generate a set of embeddings. The resulting embeddings are flattened into a single vector and returned as a numpy array.

2. Video

The latest approach for video preprocessing involves using the pre-trained VGGFace model, which employs the vgg16 architecture. This convolutional neural network (CNN) has been pre-trained on a vast dataset of faces, allowing it to learn critical facial features such as gender, facial landmarks, and expressions. The VGGFace model is based on the widely-used VGG16 architecture, which has proven successful in numerous computer vision applications.

The VGGFace model has been trained on a dataset called FaceScrub, which consists of over 100,000 images of 530 different celebrities. The images in this dataset vary in lighting conditions, facial expressions, and poses, which makes it an ideal dataset for training a model to recognize faces under different conditions.

The input image is preprocessed to fit the input shape required by the VGGFace model using keras applications. This includes resizing and preprocessing using the VGG16 preprocessing function. The output is flattened to generate a one-dimensional feature vector.

This method has several advantages. Firstly, the model can be trained on a larger set of frames since all frames can be used for training. Additionally, the VGGFace model has already learned important features from a large dataset of faces, and thus, the model can better capture important face features.

4.2 Modelling

For audio modelling, the previous proposed architecture involves an audio subnetwork that processes the preprocessed audio data using a series of convolutional and dense layers. The input tensor is 4D, representing the audio signal, and is subjected to convolutional layers with batch normalization and max pooling to extract relevant features from the signal.

The architecture of the model consists of two convolutional layers. Each layer has 32 filters with a 3x3 size and uses ReLU activation. Following this, batch normalization is applied, and max pooling is performed with a pool size of 2x2. The output is then flattened to produce a feature vector that summarizes the entire audio sample, which is passed through a dense layer with 128 units and ReLU activation.

For video modelling, we created subnetwork for processing visual input data in the form of videos with the input shape (6,128,128,3) of video frames. The VGG16 convolutional neural network is used as a feature extractor to encode each frame in the preprocessed video data independently without fully connected layers at the top of VGG16 model. Instead, we have used last pooling layer as a fixed feature extractor. To capture the temporal dependencies between the frames, the output of the previous layer is fed into an LSTM layer with an output dimensionality of 64 layers. This allows the model to effectively capture the temporal dynamics of the input data.

1. Audio Modelling:

To capture temporal dependencies between video frames, the model feeds the output of the previous layer into an LSTM layer with 64 output dimensions. This allows the model to effectively capture the temporal dynamics of the input data. Meanwhile, the audio modeling CNN architecture consists of a 1-dimensional convolutional layer, followed by batch normalization and max pooling. The CNN extracts relevant features from the audio signal by applying a set of filters, while the batch normalization layer normalizes the output of the convolutional layer to reduce overfitting and improve generalization. Finally, the max pooling layer reduces the size of the feature maps, making the model more computationally efficient.

The use of 1-dimensional convolutional layers in the CNN allows the model to learn important temporal features from the audio signal. This is because the convolutional layer slides over the audio signal in a single direction, capturing the temporal relationships between adjacent audio samples.

2. Video Modelling:

For Video modeling, an LSTM was used to summarize the video with an output dimensionality of 64 layers. The LSTM takes as input a sequence of video frames and processes them sequentially, one frame at a time. The output of each LSTM cell is then fed into the next cell in the sequence, allowing the model to capture long-term dependencies across the video frames.

The architecture used in this approach for processing sequential data is based on LSTM (Long Short-Term Memory) cells, each of which has two primary components: a hidden state and a cell state. The hidden state captures the relevant information from the previous time step, while the cell state retains the memory information from the previous time step. This enables the LSTM to selectively store or discard information, making it highly effective for processing long-term dependencies in sequential data.

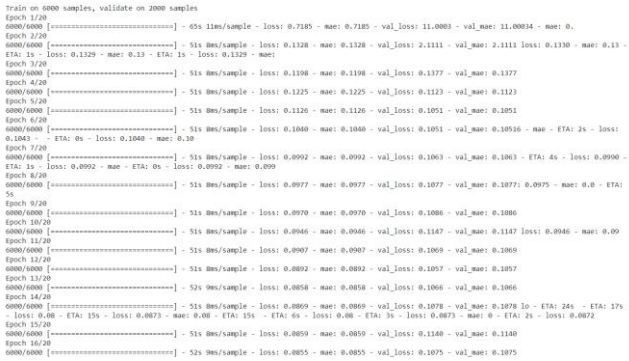
The output dimensionality of the LSTM was set to 64 layers, which means that the LSTM summarizes the video sequence into a 64-dimensional feature vector

3. Concatenation and Classification:

The outputs of the audio subnetwork and visual subnetwork were concatenated along the feature axis. Then, a dense layer with 256 units was applied to the concatenated network. Finally, the last dense layer produced a vector of 5 values which represents the

network's predictions for the input data. In addition, we have incorporated an early stopping callback in our approach to prevent overfitting and enhance the accuracy of the model. This callback enables the model to halt the training process early when the validation loss no longer improves, thereby preventing the model from continuing to learn unnecessary or redundant information.

RESULTS



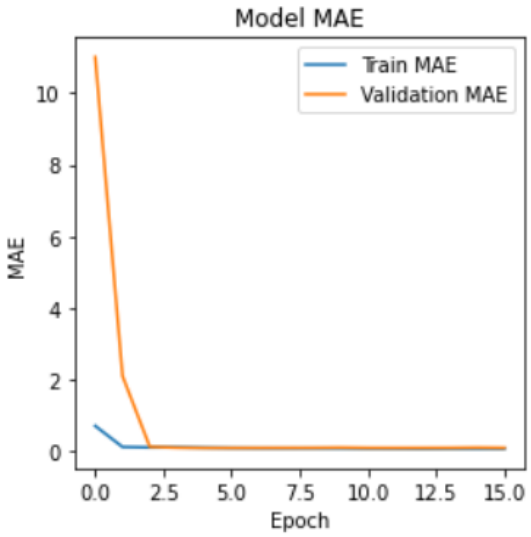
In our initial approach, the model achieved a training loss of 0.184 and a validation loss of 1.1473, indicating overfitting and a lack of an early stopping callback implementation. However, in our revised approach, we implemented an early stopping callback and utilized pre-trained audio model YAMNet and pretrained video model VGGFace to extract features. This resulted in a significant improvement, with the model achieving a training loss of 0.0855 and a validation loss of 0.1075.

Team	Agreeableness	Conscientiousness	Extraversion	Neuroticism	Openness
BU-NKU	0.913731	0.919769	0.921289	0.914613	0.917014
Baseline	0.91123	0.915228	0.91122	0.910378	0.911123
PML	0.910312	0.913775	0.91551	0.908297	0.910078
ROCHCI	0.903216	0.894914	0.90266	0.901147	0.904709
FDMB	0.891004	0.865975	0.878842	0.863237	0.874761

The CVPR 2017 challenge on predicting personality traits from social media data attracted a number of teams with different approaches. Among them, the BU-NKU team achieved the highest accuracy scores for Agreeableness, Conscientiousness, Extraversion, Neuroticism, and Openness. Their model's accuracy was 0.913731 for Agreeableness, 0.919769 for Conscientiousness, 0.921289 for Extraversion, 0.914613 for Neuroticism, and 0.917014 for Openness. As a result, we have used their model as our baseline model in this study.

Accuracy	**0.8352**
Extraversion	0.8376
Agreeableness	0.8402
Conscientiousness	0.8491
Neuroticism	0.8132
Openness	0.8360

The results achieved in the experiment show that the model has an overall accuracy of 83.52%. The model performed relatively well on all the personality traits with a minimum accuracy of 81.32% for Neuroticism and a maximum accuracy of 84.91% for Conscientiousness. Extraversion, Agreeableness, and Openness had accuracies of 83.76%, 84.02%, and 83.60%, respectively. These results indicate that the model has the potential to accurately predict personality traits based on social media data.



5. TEAM MEMBERS CONTRIBUTION

- Abhishek - Literature survey, preprocessing audio data, building audio model, preprocessing audio data using YAMNet pretrained model, building combined model, presentation, report.
- Piyush - Literature survey, preprocessing video data, building video model, preprocessing video data using VGG16 pretrained model, building combined model, presentation, report.

6. REFERENCES

[1] Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., Mittal, A. (2016). Bi-modal First Impressions Recognition Using Temporally Ordered Deep Audio and Stochastic Visual Features. In: Hua, G., Jégou, H. (eds) Computer Vision – ECCV 2016 Workshops.

[2] X. -S. Wei, C. -L. Zhang, H. Zhang and J. Wu, "Deep Bimodal Regression of Apparent Personality Traits from Short Video Sequences," in IEEE Transactions on Affective Computing, vol. 9, no. 3, pp. 303-315, 1 July-Sept. 2018

[3] Ventura, D. Masip and A. Lapedriza, "Interpreting CNN Models for Apparent Personality Trait Regression," 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 2017

[4] Heysem Kaya, Furkan Gurpinar, and Albert Ali Salah. 2017. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 1–9.

[5] 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. Springer, 349–358.

[6] Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yağmur Güç, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio CS Jacques, Meysam Madadi, et al. 2020. Modeling, Recognizing, and Explaining Apparent

Personality from Videos. IEEE Transactions on Affective Computing(2020).

[7] Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yagmur Gucluturk, Umut Guclu, Xavier Baró, Isabelle Guyon, Julio Jacques Junior, Meysam Madadi, et al. 2018. Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. arXiv preprintarXiv:1802.00745(2018).

[8] TGI Fernando et al. 2016. Persons' personality traits recognition using machine learning algorithms and image processing techniques. Advances in Computer Long Paper ICMI '20, October 25–29, 2020, Virtual Event, Netherlands Science: an International Journal 5, 1 (2016), 40–44

[9] Furkan Gürpinar, Heysem Kaya, and Albert Ali Salah. 2016. Multimodal fusion of audio, scene, and face features for first impression estimation. In 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE, 43–48.

[10] N Madzlan, Jingguang Han, Francesca Bonin, and Nick Campbell. 2014. Towards automatic recognition of attitudes: Prosodic analysis of video blogs. Speech Prosody, Dublin, Ireland (2014), 91–94.

[11] Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In European Conference on Computer Vision. Springer, 400–418.

[12] Fabio Valente, Samuel Kim, and Petr Motlicek. 2012. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In Thirteenth annual conference of the international speechcommunication association.

[13] Richard JW Vernon, Clare AM Sutherland, Andrew W Young, and Tom Hartley. 2014. Modeling first impressions from highly variable facial images. Proceedings of the National Academy of Sciences 111, 32 (2014), E3353–E3361.

[14] Firoj Alam, Evgeny A Stepanov, and Giuseppe Riccardi. 2013. Personality traits recognition on social network-facebook. In Seventh International AAAI Conference on Weblogs and Social Media.

[15] Timur R Almaev and Michel F Valstar. 2013. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In 2013 Humaine association conference on affective computing and intelligent interaction. IEEE, 356–361.