Multimodal Personality Recognition using Cross-Attention Transformer and Behaviour Encoding

Tanay Agrawal, Dhruv Agarwal, Michal Balazia, Neelabh Sinha, Francois F Bremond

Multimodal Personality Recognition using Cross-Attention Transformer and Behaviour Encoding

Tanay Agrawal[1], Dhruv Agarwal[1,3], Michal Balazia[1,2], Neelabh Sinha[1,4], and François Bremond[1,2]
[1]*INRIA Sophia Antipolis - Méditerranée, France*
[2]*Université Côte d'Azur, France*
[3]*Indian Institute of Information Technology, Allahabad, India*
[4]*Birla Institute of Technology and Science, Pilani, India*
{*firstname.secondname*}*@inria.fr*

Keywords: Multimodal Transformer, Multimodal Data, Feature Engineering, Personality Recognition

Abstract: Personality computing and affective computing have gained recent interest in many research areas. The datasets for the task generally have multiple modalities like video, audio, language and bio-signals. In this paper, we propose a flexible model for the task which exploits all available data. The task involves complex relations and to avoid using a large model for video processing specifically, we propose the use of behaviour encoding which boosts performance with minimal change to the model. Cross-attention using transformers has become popular in recent times and is utilised for fusion of different modalities. Since long term relations may exist, breaking the input into chunks is not desirable, thus the proposed model processes the entire input together. Our experiments show the importance of each of the above contributions.

1 Introduction

Personality is a combination of behavior, emotion, motivation, and thought patterns. Our personality greatly impacts our lives, defining choices, health along with our preferences and desires. Personality traits define a particular way of thinking, feeling, and behaving. Specifically, personality traits have been defined pertaining to individual well being and social-institutional outcomes like occupational choices, interpersonal relations, and success in various scenarios.

We make decisions using a two system model: rational and emotional. So, modelling the latter will help us build more accurate AI systems when it is coupled with the vast amount of research done on the former. The problem of personality recognition is complex and thus would require a lot of training data to get models usable in real-life. This is one reason that multimodal learning is very popular in this domain. First Impressions v2 (Junior et al., 2021) is a multimodal dataset for personality recognition and is used in this work. We utilise all the information available in the dataset – speech, body language, expressions and their surroundings along with their demographic information – and define a new behaviour encoding to facilitate learning. Deep learning backbones have been found to extract meaningful features. Generally, larger models give better features. Due to the high number of inputs, it is not possible to have large backbones for all. So we decide to compute the additional behaviour encoding to have better features even with a smaller backbone. In multimodal learning, we need to process each modality individually and also find how they are correlated. Thus, we also show how to merge the behaviour encoding with an existing baseline (Palmero et al., 2021). Temporal processing is also important, we use LSTMs to have a higher temporal resolution. Even a simple temporal processing model helps as there are multiple modalities and the embedding input to the temporal processor is very rich in information.

For defining personality, the big five personality traits are used. They are often referred to as OCEAN: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. These five traits represent broad domains of human behaviour and account for differences in both personality and decision making. Today, a major use of this model is by Human Resource practitioners to evaluate potential employees and marketers to understand the audiences of their products.

The focus on this problem is relatively new and there is limited work done till date. As discussed above, the task is complex so it requires careful processing of each input modality and their relations. Previous works either involves using a subset of modalities or only simple fusion techniques to fuse the modalities (Aslan and Güdükbay, 2019).

Another challenge in this domain is that annotations are generally provided by the participants through questionnaires or an online answering platform during or after the experiment sessions. They may also be annotated by third-party annotators, but personality is subjective so they are not always perfect. This makes the task even harder and further requires a method that can utilise all available modalities and formulate complex relations not only for each modality but also across modalities. Defining handcrafted inputs increases performance as there might be more direct correlation between them and other modalities or even the output as compared to the original inputs.

As stated above, ChaLearn First Impressions V2 challenge dataset (Ponce-López et al., 2016) which is publicly available is used (Palmero et al., 2021) for this work. The dataset consists of 10,000 videos of people facing and speaking to a camera. Videos are extracted from YouTube, they are mostly in high-definition (1280×720 pixels), and, in general, they have an average duration of 15 seconds with 30 frames per second. In the videos, people talk to the camera in a self-presentation context and there is a diversity in terms of age, ethnicity, gender, and nationality. The videos are labeled with personality factors using Amazon Mechanical Turk (AMT), so the ground truth values are obtained by using human judgment. For the challenge, videos are split into training, validation and test sets with a 3:1:1 ratio and we choose to use the same to compare the results.

Summarising our contributions in this work, we introduce a handcrafted behaviour embedding that improves performance and reduces convergence time of the model. We modify the chosen baseline to incorporate new modalities (transcript and behaviour encoding) and also address missing temporal relations in it. We also achieve state of the art results for personality recognition on the chosen dataset.

2 Related Work

This section discusses the work done on personality recognition using different techniques and modalities. They can be broadly classified into the following categories.

2.1 Using Video

As in the case of most visual deep learning tasks, Convolutional Neural Networks (CNNs) are the most commonly used in the field of personality detection. Facial attributes can be an important factor in predicting social traits (Qin et al., 2016; Vernon et al., 2014). Impressions that influence people's behavior towards other individuals can be accurately predicted from videos (Gürpınar et al., 2016). Many researchers have experimented with different ways of capturing facial features such as in the form of Facial Action Coding System (FACS) which extracts action units such as raised eyebrows or blinking, and morphological features (Güçlütürk et al., 2017).

2.2 Using Audio

Using audio as the only input modality is not a popular choice for personality recognition. It is combined with video in most of the cases resulting in bimodal approaches. In the existing ones, audio features like Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Logfbank, other cepstral and spectral ones serve as inputs into regressors. Analyzing conversations (Valente et al., 2012) and the pitch, timber, tune and rhythm of the voice (Madzlan et al., 2014), it is possible to recognize the personality traits or predict the speaker attitudes automatically. These approaches demonstrate that audio information is important for personality.

2.3 Using Text

Looking at the textual modality, preprocessing is an important step. Generally, extracted features include Linguistic Inquiry and Word Count (LIWC) (Mikolov et al., 2013), Mairesse, Medical Research Council (MRC), which are then fed into standard classifiers or regressors. Learning word embeddings and representing them as vectors, either with GloVe, Word2Vec, GRU, LSTM or recently BERT, is also a very commonly followed approach. It was observed that combining text features with something else such as metadata and convolutions results in better performance paving the path to multimodal approaches.

Social networks provide rich textual data for the recognition of personality traits (Alam et al., 2013; Farnadi et al., 2016). Transcribed videos blogs and dialogues also provide useful information for this task (Nowson and Gill, 2014).

2.4 Multimodal approaches

Personality traits can be detected in self presentation videos based on the acoustic and visual, non-verbal features such as pitch, intensity, movement, head orientation, posture, fidgeting and eye-gaze. Zheng et al. (Zeng et al., 2009) shows body gestures, head movements, expressions, and speech lead to effective assessment of personality and emotion. According to Sarkar et al. (Sarkar et al., 2014), features such as audiovisual, text, demographic and sentiment features are important for our task.

Although multimodal approaches are commonly used to recognize personality traits, there does not exist a comprehensive method utilizing a considerable amount of informative features. Most of the multimodal approaches perform late fusion. Deep bimodal regression give state of the art results (He et al., 2015). Some other approaches with good results are (Gürpinar et al., 2016; **?**) and (Wei et al., 2018). Each modality features may be used together for personality prediction, this approach is called early fusion. Present research in the field aims to find efficient ways of feature extraction and combination. Few models which have dealt with trimodal fusion of features (Aslan and Güdükbay, 2019; Palmero et al., 2021). Emotion recognition is a closely related problem and has interesting approaches for multimodal data processing (Dai et al., 2021; Tsai et al., 2019). Our approach aims to utilise all possible information available and also some extra features computed similar to the ideas discussed in the beginning of this subsection.

3 The Proposed Framework

The approach uses face crops of the target person and relates it to body language, surroundings and speech using a transformer based architecture. Short-term temporal relations are processed in this way and longer temporal relations are established using LSTM. For transcript analysis, short term temporal relations are not very meaningful so the features for the entire input sequences are extracted using BERT (Devlin et al., 2019). Late fusion is then finally used for inferring the OCEAN personality traits. There are several stages in the proposed method and they are discussed in the following sections. Figure 1 shows the overview of the entire architecture.

3.1 Preparing the Input and Feature Extraction

The audiovisual data is pre-processed in a similar manner as in (Palmero et al., 2021). 32 frames with a stride of 2 are taken for video based inputs and R(2+1)D (Tran et al., 2018) is used to extract spatio-temporal features. Stride is modified depending on the frame rate of the video to keep the time span of the chunks roughly the same. Audio clip with the same time span is converted to a tensor for input in the same way as in the method used in VGGish (Hershey et al., 2017). BERT is used for extracting features of the transcript. The method for computing behaviour encoding is discussed in the next section in detail. Demographic data – age, gender, ethnicity and attractiveness – are also used. The value are either one hot encode or normalised to the range $[0,1]$. Table 1 and (Escalante et al., 2019) give the details of each element of the feature vector.

Note that Attractiveness is only available for people with Caucasian ethnicity, but since ∼86% of the people in the dataset are Caucasian, it is utilised and the default Attractiveness value of 0 is set for people of other ethnicities.

Demographic variable	Dimension
Ethnicity	3D (one hot encoding)
Gender	2D (one hot encoding)
Age	1D
Attractiveness	1D

Table 1: Dimensions of demographic metadata

3.2 Behaviour Encoding

We compute behaviour encoding for 13 actions: head tilt, thrust, bob, lips in, mouth corner, frown, small mouth, wrinkle, crouch, lean forward, fold arms, hand to face, hand to mouth. For detecting the individual behaviors, we use a rule based approach on the skeleton and facial key points. For extracting the key points (skeleton and face), we use LCRNet (Rogez et al., 2019) and OpenFace (Baltrusaitis et al., 2016).

In each frame, we infer a detection confidence for all behaviors in the scale of 0–1, where 1 represents complete confidence in presence of the behavior and 0 represents complete confidence in absence of the behavior. This is done by extracting a specific feature x and transforming it through a sigmoid function

$$f_\sigma(x) = \frac{1}{1+e^{-\lambda_\sigma(x-c_\sigma)}}$$

with parameters of center c_σ and multiplier λ_σ. As shown in Table 3, each behavior is characterized by

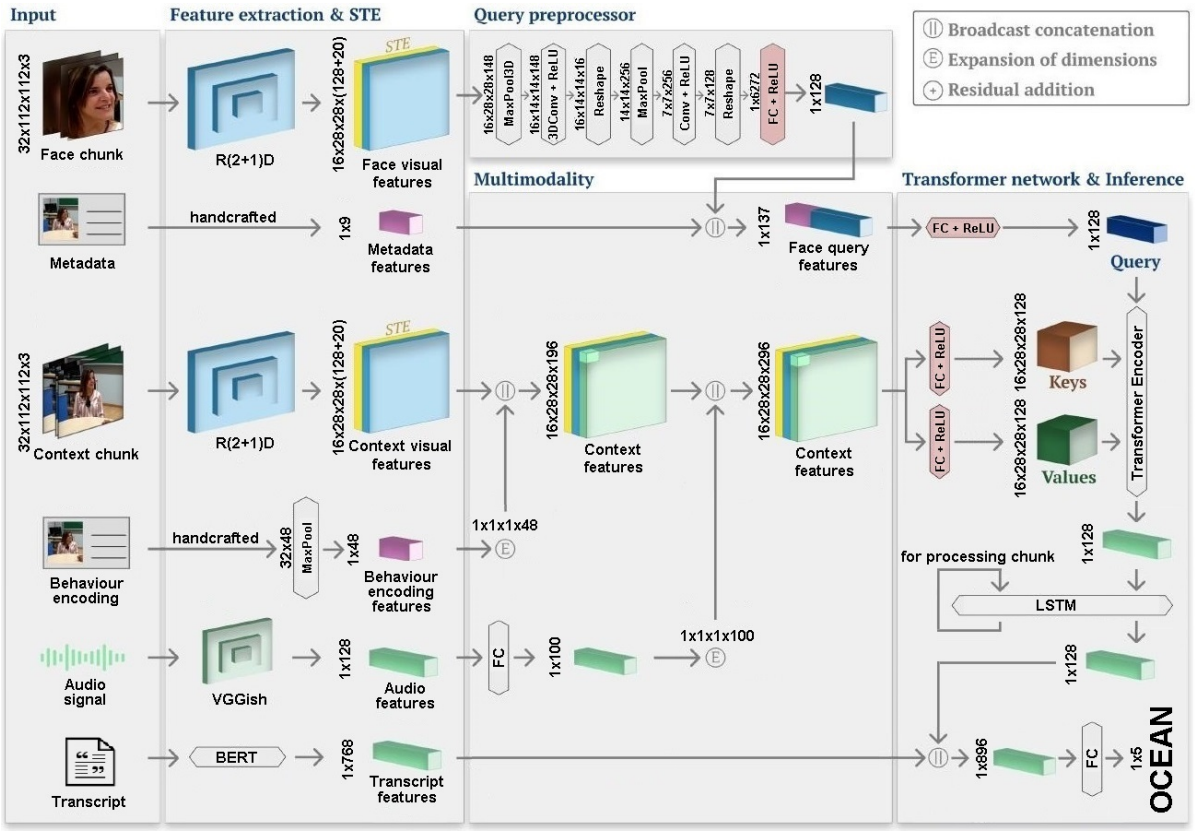

Figure 1: Proposed method to infer self-reported personality (OCEAN) traits from multimodal data. Input consists of visual (face and context chunks), audio (raw chunks), metadata of the target person, handcrafted behaviour encoding and transcript of the audio. Feature extraction is performed by a R(2+1)D network for the visual chunks, VGGish for audio and BERT for the transcript. The visual features from the R(2+1)D's 3rd residual block are concatenated to spatiotemporal encodings (STE). The VGGish's audio features and handcrafted metadata features are incorporated to visual context/query features and the result transformed to the set of Query, Keys, and Values as input to the Transformer encoder. The output of the transformers are sequentially passed to an LSTM chunkwise. The transcript features from BERT are concatenated with these and finally fed to a fully-connected (FC) layer to regress per-video OCEAN scores.

behavior	extracted feature x	c_σ	λ_σ
head tilt	head roll angle	10	1
thrust	derivative of translation vector along z axis when derivative along other directions is less than 10 and direction of derivative in previous and next frame is the same	$-25cm/s$	1
bob	derivative of pitch angle when derivative of yaw angle is less than 20 and direction of derivative in previous and next frame is the same	$-50deg/s$	1
lips in	FACS action unit Lip Suck	-	-
mouth corner	FACS action unit Lip Stretcher	1.2	6
frown	FACS action unit Brow Lowerer	1.2	6
small mouth	FACS action unit Lip Tightener	1.2	6
wrinkle	FACS action unit Nose Wrinkler	1.2	6
crouch	distance between knees and head	$30cm$	-0.35
lean forward	z coordinate on distance between root and shoulders	$10cm$	4
fold arms	alternate distance between elbows and wrists when y coordinate of both elbows are less than $10cm$	$20cm$	-0.5
hand to face	distance between wrists and head	$35cm$	-0.5
hand to mouth	distance between wrists and head minus $10cm$ on y axis	$25cm$	-0.5

Table 2: Detection methods of 13 behaviors.

a specific extracted feature and the two sigmoid function parameters.

3.3 Positional Encoding for the Transformer

Positional encodings are important to be added to the input with transformer based models as they make the model order invariant. Sinusoidal encodings are common, but we choose to use learned encodings in our experiments. As we need to process in both space and time, we need an encoding for both. We initialize encodings for both and use a two layer fully connected network for learning them. Then they are broadcast concatenated to each other resulting in an encoding which can be concatenated to the input.

3.4 Preparing Inputs for the Transformer

Features extracted from face crops of the complete frame input are further processed and are used as the query for the transformer. To factor its relation with the rest of the information in the complete frame and audio inputs, they are processed to be used as key and value. The face features are passed through the following layers to get the input query:

1. 3D max pooling layer with a kernel size and stride of 2 for height and width dimensions, and 1 for the temporal dimension.
2. 3D convolution layer with kernel size of 1 for all dimensions and 16 kernels.
3. ReLU activation followed by reshaping to merge temporal and channel dimensions.
4. 2D max pooling layer with a kernel size and stride of 2 for height and width dimension.
5. 2D convolution layer with kernel size of 2 for all dimensions and 128 kernels.
6. ReLU activation followed by flattening.
7. A fully connected layer to change the shape to 128, followed by ReLU activation and dropout $p = 0.2$ layer.

Demographic metadata is concatenated to the obtained feature vector and is passed through a fully connected and ReLU layer to obtain a 128 dimensional query vector.

Behaviour encoding is broadcast concatenated to complete frame features which already contain spatio-temporal positional encoding. The audio features are projected into a 100 sized feature vector using a fully connected layer and broadcast concatenated with the above obtained complete frame features. These are passed through separate fully connected and ReLU layers to obtain keys and values for the transformer.

3.5 Transformer, Temporal Processing and Fusion with Transcript Features

The transformer consists of only the encoder with 2 attention heads and stacked 3 times, that is, 3 layers. The hidden dimension is 128. The transformer processes roughly 2.5 seconds of the input in one forward pass. These chunks are passed through two stacked LSTM blocks to find long-term temporal relations. The hidden state after the last chunk is passed, is concatenated with transcript features and passed through linear, ReLU and dropout layers to obtain the 5 personality trait values.

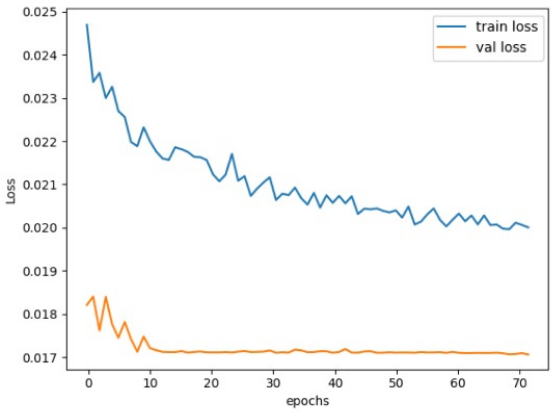

Figure 2: MSE Loss curves for w/o Transcript Ablation experiment

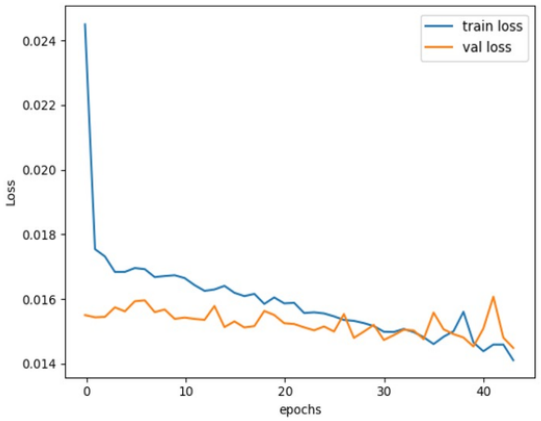

Figure 3: MSE Loss curves for our proposed approach

4 Experiments and Results

4.1 Training Details

To reduce training time, the parameters of backbones R(2+1)D, VGGish and BERT are frozen and are not updated during backpropagation. There is one exception to this, we finetune our model with the weights unfrozen as behaviour encoding helps improve the performance of the backbone also and to exploit that, we finetune our model for 20 epochs. One RTX 6000 GPU is used for training and the batch size is set to 8. Learning rate used is 10^{-5} with the scheduler "ReduceLROnPlateau", patience 5 and factor 0.5. Further details of the experiments are given in section 4.3. Figure 2 and 3 show training graphs of two different experiments, an ablation study with the proposed framework without transcript and the proposed framework, respectively. It is interesting to note that adding modalities decreases the number of epochs required for convergence as shown by these two figures.

4.2 Evaluation Protocol

The evaluation metric is chosen to be the same as that of the ChaLearn challenge where the dataset was released. The OCEAN traits have five classes which are rated in the range $[0,1]$. The challenge (Ponce-López et al., 2016) defines mean accuracy A over all predicted personality trait values as

$$A = 1 - \frac{1}{N}\sum_{i=1}^{N}|t_i - p_i| \quad (1)$$

where t_i are the ground truth scores and p_i are the predicted scores for personality traits summed over N videos.

4.3 Results and Ablation Studies

We compare our result to the previous state of the art and also perform ablation studies to show the need for all modalities present. Table 3 enumerates these results.

Model	**Accuracy**					
	O	C	E	A	N	Mean
Aslan and (Aslan and Güdükbay, 2019)	.9166	.9214	.9208	.9189	.9162	.9188
DCC (Güçlütürk et al., 2016)	.9117	.9113	.9110	.9158	.9091	.9122
evolgen (Subramaniam et al., 2016)	.9130	.9136	.9145	.9157	.9098	.9138
Gurpinar et al. (Gürpinar et al., 2016)	.9141	.9141	.9186	.9143	.9123	.9147
PML (Bekhouche et al., 2017)	.9138	.9166	.9175	.9166	.9130	.9170
BU-NKU (Kaya et al., 2017)	.9169	.9166	.9206	.9161	.9149	.9170
Our proposed model	**.9291**	**.9258**	**.9272**	**.9288**	.9210	**.9263**
Baseline: w/o behaviour encoding and transcript	.8959	.8996	.8987	.8938	.8932	.8962
w/o behaviour encoding	.9095	.9094	.9112	.9133	.9041	.9095
w/o transcript	.9013	.8992	.8988	.9041	.8996	.9006
w/o LSTM	.8892	.8532	.9131	.9024	**.9315**	.8978
w/o metadata	.9260	.9212	.9234	.9249	.9168	.9225

Table 3: Experiments and Results; O: Openness, C: Conscientiousness, E: Extroversion, A: Agreeableness, and N: Neuroticism

We achieve state of the art results as we utilise all the available information and also compute an additional behaviour embedding to facilitate learning. This method of computing a behaviour encoding can

be utilised in a variety of use-cases and we predict that it will help in reducing training time and improving results in other areas, such as action recognition also.

The ablation study proves the efficacy of our approach, showing the importance of using different input modalities and the difference in results is significant. All the different models discussed below are trained in parallel and not sequentially, that is, the later models were not finetuned from the initial ones. The first approach includes the baseline model and has the same inputs modified as per the dataset details. The baseline has all the inputs except the behaviour encoding and the transcript. This experiment is to establish our own baseline results to compare against.

We see the results of the model without behaviour encoding. There is roughly 1.8% decrease in accuracy which shows that behaviour encoding facilitates in the prediction of personality.

We also observe the performance of the model without transcript. This shows a similar trend as behaviour encoding - there is a slightly less decrease in accuracy but the difference is minute.

For finding the performance of the model with LSTM, we keep everything the same but take the median value across chunks for each video to get the output. Without LSTM, the model behaves erratically. As expected the accuracy decreases for most classes. But, for the class neuroticism, the best results are without LSTM. One explanation is the high variability in the inputs where neuroticism is high and the LSTM which tries to identify a pattern across chunks does not perform very well.

The last experiment is without metadata about the target person. There is not much difference in results as compared to the other inputs, but we still see a reduction in performance. So, demographic data about a person affects personality too. Some bias in the data is the most probable cause but since the dataset is large, in our opinion this is not the case and the inference drawn holds.

5 Conclusions

In this work, we show that a model for personality recognition will benefit from more modalities and data as input. We propose a new handcrafted behaviour encoding where each element is the probability of a low level action relevant to the task. We show the effectiveness of all the inputs in the data through ablation studies. We also give our opinion on the trends shown in the ablation studies. Owing to the interdisciplinary nature of the project, there are numerous additions that will further improve performance. From intuition, there are some which might improve performance by a higher margin than others. Using better backbones for feature extraction would be interesting. We use the same ones as in the baseline we choose but there are existing models with better performance for similar tasks that can be utilised. Transformers have been shown to perform better than LSTMs. In the future, we will try to increase temporal scale of attention in the transformer rather than using a separate module for combining information across chunks. This might tackle the problem that is seen with neuroticism as discussed in section 4.3. One of the major drawbacks of multimodal data is that preprocessing takes a lot of time. Thus, it will be interesting to explore Knowledge Distillation to allow the model to utilise one or a subset of modalities and give a similar performance but with lesser inputs. We would also like to test our approach on other big scale multimodal datasets, when they are available in the future. This area of work has a lot of applications in healthcare which we are exploring and hope that this work leads to advancement in the area. We also hope that it motivates other people to work on this interesting problem.

REFERENCES

Alam, F., Stepanov, E., and Riccardi, G. (2013). Personality traits recognition on social network -facebook.

Aslan, S. and Güdükbay, U. (2019). Multimodal video-based apparent personality recognition using long short-term memory and convolutional neural networks.

Baltrusaitis, T., Robinson, P., and Morency, L.-P. (2016). Openface: An open source facial behavior analysis toolkit. pages 1–10.

Bekhouche, S. E., Dornaika, F., Ouafi, A., and Taleb-Ahmed, A. (2017). Personality traits and job candidate screening via analyzing facial videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1660–1663.

Dai, W., Cahyawijaya, S., Liu, Z., and Fung, P. (2021). Multimodal end-to-end sparse model for emotion recognition.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

Escalante, H. J., Kaya, H., Salah, A. A., Escalera, S., Gucluturk, Y., Guclu, U., Baro, X., Guyon, I., Junior, J. J., Madadi, M., Ayache, S., Viegas, E., Gurpinar, F., Wicaksana, A. S., Liem, C. C. S., van Gerven, M. A. J., and van Lier, R. (2019). Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., Davalos, S., Moens, M.-F., and De Cock, M. (2016). Computational personality recognition in social media. *User Modeling and User-Adapted Interaction*, 26.

Gürpınar, F., Kaya, H., and Salah, A. A. (2016). Combining deep facial and ambient features for first impression estimation. In Hua, G. and Jégou, H., editors, *Computer Vision – ECCV 2016 Workshops*, pages 372–385, Cham. Springer International Publishing.

Gürpinar, F., Kaya, H., and Salah, A. A. (2016). Multimodal fusion of audio, scene, and face features for first impression estimation. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 43–48.

Güçlütürk, Y., Güçlü, U., Pérez, M., Escalante, H. J., Baró, X., Andujar, C., Guyon, I., Junior, J. J., Madadi, M., Escalera, S., Van Gerven, M. A., and Van Lier, R. (2017). Visualizing apparent personality analysis with deep residual networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 3101–3109.

Güçlütürk, Y., Güçlü, U., van Gerven, M. A. J., and van Lier, R. (2016). Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. *Computer Vision – ECCV 2016 Workshops*, page 349–358.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R., and Wilson, K. (2017). Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Junior, J. C. S. J., Lapedriza, A., Palmero, C., Baro, X., and Escalera, S. (2021). Person perception biases exposed: Revisiting the first impressions dataset. pages 13–21.

Kaya, H., Gürpinar, F., and Salah, A. A. (2017). Multimodal score fusion and decision trees for explainable automatic job candidate screening from video cvs. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1651–1659.

Madzlan, N., Han, J., Bonin, F., and Campbell, N. (2014). Automatic recognition of attitudes in video blogs - prosodic and visual feature analysis.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality.

Nowson, S. and Gill, A. J. (2014). Look! who's talking? projection of extraversion across different social contexts. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, page 23–26, New York, NY, USA. Association for Computing Machinery.

Palmero, C., Selva, J., Smeureanu, S., Junior, J. C. S. J., Clapes, A., Mosegui, A., Zhang, Z., Gallardo, D., Guilera, G., Leiva, D., and Escalera, S. (2021). Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–12.

Ponce-López, V., Chen, B., Oliu, M., Corneanu, C., Clapés, A., Guyon, I., Baró, X., Escalante, H. J., and Escalera, S. (2016). Chalearn lap 2016: First round challenge on first impressions -dataset and results. *European Conference on Computer Vision*.

Qin, R., Gao, W., Xu, H., and Hu, Z. (2016). Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face.

Rogez, G., Weinzaepfel, P., and Schmid, C. (2019). Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1.

Sarkar, C., Bhatia, S., Agarwal, A., and Li, J. (2014). Feature analysis for computational personality recognition using youtube personality data set. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, page 11–14, New York, NY, USA. Association for Computing Machinery.

Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., and Mittal, A. (2016). Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *ECCV Workshops*.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Valente, F., Kim, S., and Motlicek, P. (2012). Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus. In *Proc. Interspeech 2012*, pages 1183–1186.

Vernon, R. J. W., Sutherland, C. A. M., Young, A. W., and Hartley, T. (2014). Modeling first impressions from highly variable facial images. *Proceedings of the National Academy of Sciences*, 111(32):E3353–E3361.

Wei, X.-S., Zhang, C.-L., Zhang, H., and Wu, J. (2018). Deep bimodal regression of apparent personality traits from short video sequences. *IEEE Transactions on Affective Computing*, 9(3):303–315.

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.