Continuous Affect Recognition from Multimodal Signals in Videos

CPSC 6300 Applied Data Science Spring 2023

Charanjit Singh
School of Computing
Clemson University
CU ID: C15246652
charans@g.clemson.edu

Parampreet Singh
School of Computing
Clemson University
CU ID: C19377466
paramps@g.clemson.edu

Vinod Ramavath
School of Computing
Clemson University
CU ID: C13139775
vramava@g.clemson.edu

ABSTRACT

In this project, we aimed to predict arousal and valence values from utterances of videos using the OMG Dataset. We used three machine learning models - AudioNet, VideoNet, and Joint Training - to extract audio and visual features from the data. These models were then fed to a FC Layer followed by tanh activation for arousal/valence response prediction training. We evaluated our models using the CCC metric and found that the Joint Training model gave the best results. Based on our analysis, we found that both audio and visual features were important predictors of arousal and valence values. Our results suggest that combining audio and visual features can improve the accuracy of prediction models for affective computing. Domain experts in affective computing can use our findings to better understand the relationship between audio and visual cues and emotional states. Finally, we discuss how our project could be improved with more data and additional models.

INTRODUCTION

In affective computing, it is essential to be able to reliably identify and interpret emotions from human speech and facial expressions. Application areas for affect identification include social robots, human-computer interaction, and mental health. In this experiment, our major goal was to determine whether it was possible to estimate arousal and valence values from video utterances using machine learning models.

Our study issue is driven by a desire to investigate how merging audio and visual information might enhance the precision of emotional computing models. Numerous practical uses for being able to properly forecast emotional states include marketing research, human-robot interaction, and mental health diagnosis.

We used the OMG Dataset in our project, which includes audio and visual recordings of participants while watching videos that evoke different emotional responses. The dataset consists of around 600 videos, each having on average 8-10 utterances, of which each lasts between 3 and 60 seconds. The data was annotated by multiple human raters for arousal and valence values, which we used as our response variable. The dataset was already publicly available, and we did not have to collect the data ourselves.

DATA REPRESENTATION

A. Unit of analysis – Individual video utterances. Each row represents a single utterance, and the columns contain features such as the transcript of the utterance, the video ID, and the predicted arousal and valence values.

B. Observations – There were around 6000 individual video utterances in the OMG Dataset subset that were used in this project for training, validating, and predicting arousal and valence values.

C. Time period – The OMG Dataset does not cover a specific time, as it consists of a collection of video recordings from various sources.

D. Response Variable – Predicted arousal and valence values for each video utterance.

Arousal: -1 Calm to +1 Alert
Valence: -1 Negative to +1 Positive

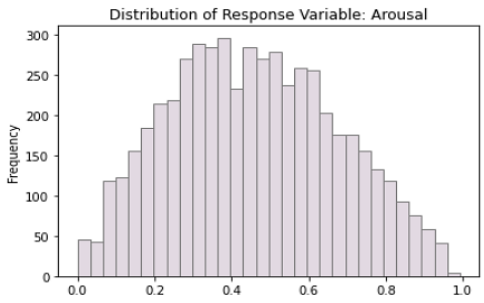

Figure 1: Distribution of Arousal Response

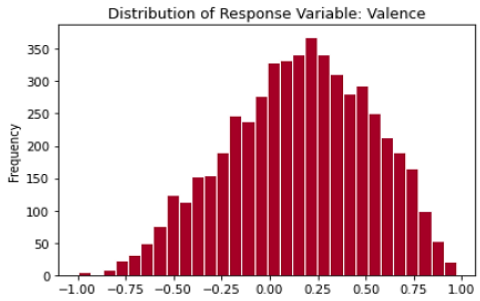

Figure 2: Distribution of Valence Response

DATA PREPROCESSING

This section details how we preprocess the audio and visual modalities from the provided OMG Dataset [1].

A. Acoustic Representation –

Since the audio files were not provided separately, we first converted all utterance snippets to WAV format files using MoviePy. These files were then used to get the Short-Term Fourier Transform (STFT) spectrum [2] using Librosa. Since each frequency bin in the spectrum is a complex number, we obtain STFT maps containing both the real and imaginary parts of the acquired STFT values.

B. Visual Representation –

For visual data preprocessing, we extracted frames from each video utterance using OpenCV [3]. Then we applied MTCNN [4] to those frames to align faces before feeding them to the SphereFace [5] architecture.

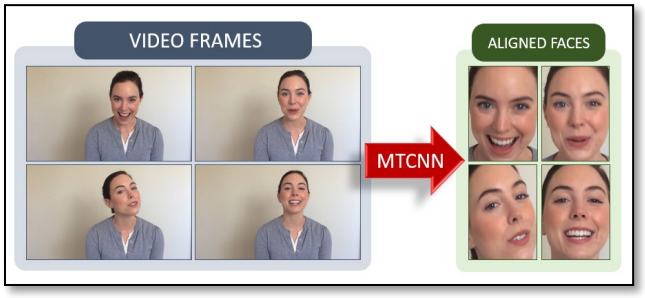

Figure 3: Visual Preprocessing

MODEL ARCHITECTURE [6]

For this project, we implemented the following models and compared their test error rates to finalize our choice of ML model.

A. AudioNet –

For processing the audio data, the STFT maps are used as input to a modified VGG16 [6] network. The original VGG16 network is a popular convolutional neural network architecture, which was trained on millions of images from the ImageNet dataset. However, the first layer of VGG16 is designed to accept 3-channel RGB images as input, whereas the STFT maps have a depth dimension of 2. Therefore, the first layer of VGG16 is modified to accept 2-channel input, which corresponds to the real and imaginary parts of the STFT maps.

The modified VGG16 network is used to extract a 7×9×512 feature map from the STFT maps. This feature map is then flattened and fed into two fully connected (FC) layers with ReLU activation and dropout regularization in between. The first FC layer has 4096 neurons, while the second FC layer has 512 neurons. These layers learn to map the high-dimensional STFT feature map to a lower-dimensional feature space that captures the relevant information for predicting arousal and valence.

If the network is trained solely on audio data, an additional FC layer with 2 output neurons and the Tanh activation function is applied to obtain the final arousal and valence values. The Tanh function maps the output of the last FC layer to the range of [-1, 1], which corresponds to the range of arousal and valence values.

B. VideoNet –

To make use of the temporal information present in the snippets of varying length in the dataset, the number of frames extracted may vary for each snippet. To efficiently utilize GPU memory, the authors sparsely sample N_V frames from each snippet. They divide each snippet into N_V segments, and then randomly select a single frame from each segment to obtain intermediate features (of dimension 512) using the SphereFace model.

These N_V features are then fed into a bidirectional LSTM (long short-term memory) model. The LSTM is designed to process sequences of input data, and it can selectively remember or forget information over long periods of time. It is bidirectional because it processes the sequence both forward and backward, allowing it to capture dependencies in both directions.

Finally, the output of the LSTM is passed through a temporal average pooling layer, followed by a fully connected layer and a Tanh activation function. The temporal average pooling layer aggregates the output of the LSTM across time, resulting in a fixed-length vector. The fully connected layer maps this vector to a 2-dimensional output, representing the arousal and valence scores of the input snippet. The Tanh function ensures that the output values are in the range of -1 to 1, which is a common range for representing emotion scores.

C. Joint Training –

The joint training scheme is designed to train ANet and VNet together. Firstly, ANet and VNet are trained separately. To train them together, NASTFT maps are sampled from every audio snippet and the output of the penultimate FC layer in ANet is averaged. Then, the ANet and VNet features are concatenated and fed into another FC layer followed by Tanh. This scheme is illustrated in Figure 1. By combining the audio and visual features in this way, the joint training can better capture the emotional states from both modalities.

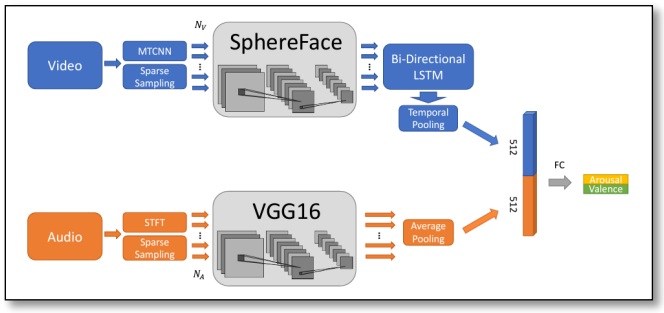

Figure 4: The workflow of the joint training architecture

The choice of using AudioNet and VideoNet models is appropriate for extracting relevant features from audio and visual data for the response variables of arousal and valence. Joint training of these models can capture more complex aspects of emotional expression by concatenating the visual and acoustic features.

EVALUATION METRICS

A. The Concordance Correlation Coefficient (CCC) –

The CCC is a measure of the agreement between two sets of measurements, in this case the true values and predicted values. The CCC ranges between -1 and 1, where values closer to 1 indicate higher agreement.

$$CCC = \frac{2\rho\sigma_y\sigma_{y'}}{\sigma_y^2 + \sigma_{y'}^2 + (\mu_y - \mu_{y'})^2}$$

where:

ρ : Pearson correlation coefficient between the true and predicted values.

$\sigma_y, \sigma_{y'}$: Standard deviations of the true and predicted values, respectively.

$\mu_y, \mu_{y'}$: Means of the true and predicted values, respectively.

B. Mean Squared Error (MSE):

The MSE is a measure of the average squared difference between the true and predicted values. The smaller the value of MSE, the better the predictions are.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{true,i} - y_{pred,i})^2$$

where:

$y_{true,i}$: Vector of true values

$y_{pred,i}$: Vector of predicted values

N : Number of samples in the data set

CCC COMPARISON WHEN USING ONLY AUDIO NET

MODEL	AROUSAL	VALENCE	TOTAL
Baseline	0.15	0.21	0.36
ANet	0.19	0.26	**0.45**

CCC COMPARISON WHEN USING ONLY VIDEO NET

MODEL	AROUSAL	VALENCE	TOTAL
Baseline	0.12	0.23	0.35
VNet	0.28	0.47	**0.75**

CCC COMPARISON WHEN USING AUDIO – VIDEO JOINT TRAINING

MODEL	AROUSAL	VALENCE	TOTAL
Baseline [Audio]	0.15	0.21	0.36
Baseline [Video]	0.12	0.23	0.35
Joint Training	0.30	0.48	**0.78**

The Joint Training model had the lowest CCC (Concordance Correlation Coefficient) error of 0.78, followed by VideoNet with a CCC error of 0.75 and AudioNet with a CCC error of 0.45.

This indicates that the Joint Training model performs the best in predicting arousal and valence values. It also benefits from incorporating features from both audio and visual data, allowing it to capture more nuances in emotional expression.

Therefore, for this project we selected the **Audio-Video Joint Training** model.

RESULTS

Following are 3 videos for which we predicted the arousal and valence values.

A. Video 1 –

It is a video of a person giving a speech with a serious tone.

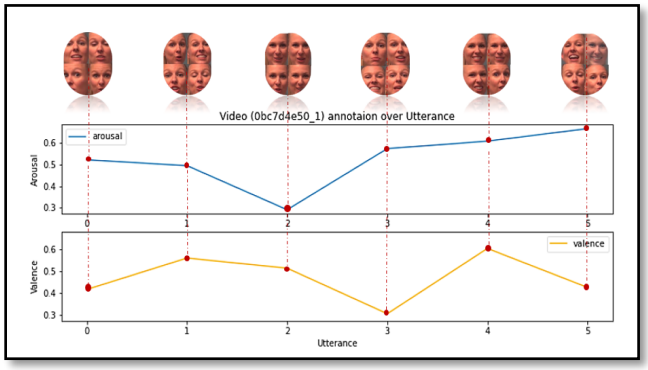

Figure 2: Predicted Arousal/Valence for Video 1

B. Video 2 –

It is a video of a person laughing, smiling and having a positive emotion throughout.

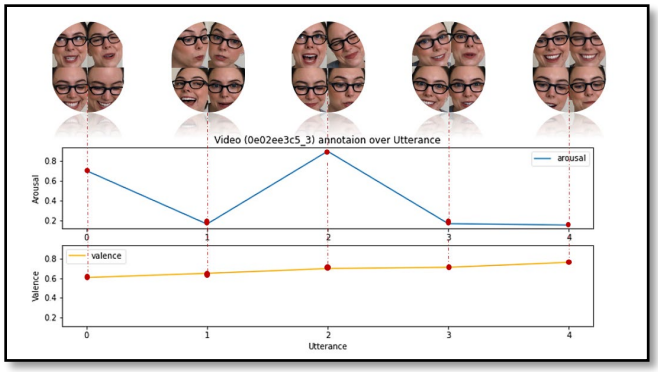

Figure 3: Predicted Arousal/Valence for Video 2

C. Video 3 –

It is a video of a person happy and excited at start, then expressing anger and sadness through the middle of video and is again happy at the end.

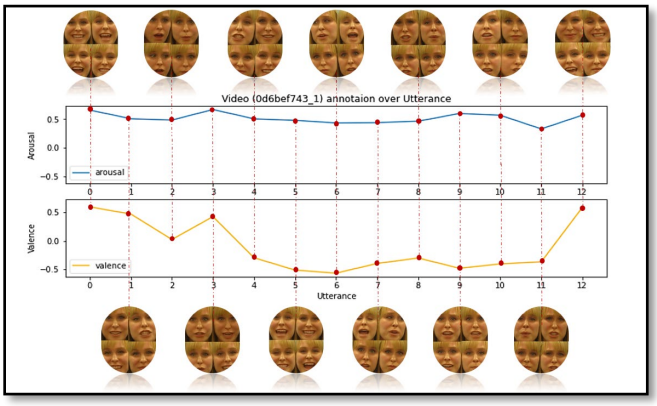

Figure 4: Predicted Arousal/Valence for Video 3

CONCLUSION

In this project, we analyzed the OMG dataset to predict the arousal and valence values from video utterances. We used three models: AudioNet, VideoNet, and Joint Training, and evaluated them using the CCC metric. Our Joint Training model gave the best results followed by VideoNet and then AudioNet. We also performed exploratory data analysis to better understand our data and made data cleaning and preprocessing decisions based on our observations.

The main question our project sought to answer was whether we can predict the arousal and valence values from video utterances. Based on our analysis, we can say that we can predict these values with good accuracy using audio and visual features.

Domain experts in the field of affective computing can learn from our project that it is possible to predict arousal and valence values from video utterances using audio and visual features. Our results could inform their work by providing a framework for developing more accurate and reliable models for predicting affective states from video utterances.

If we had more time and resources, we could improve our project in several ways. One way would be to gather more data to increase the size of our dataset. We could also explore different data preprocessing techniques to see how they affect the performance of our models. Additionally, we could experiment with more complex models or ensemble methods to improve the accuracy of our predictions.

ACKNOWLEDGMENTS

We would like to express our gratitude to everyone who contributed to the success of this project. First and foremost, we would like to thank Dr Nina Christine Hubig and the course TAs for their guidance, support, and encouragement throughout the project. We would also like to thank OMG Emotion Challenge team who provided us with access to the data used in this project. Without their support, this project would not have been possible.

In addition, we would like to thank our fellow team members for their hard work and dedication to this project. Each member played a vital role in the project's success, and we are grateful for their contributions.

REFERENCES

[1] Pablo Barros, Nikhil Churamani, Egor Lakomkin, Henrique Siqueira, Alexander Sutherland, and Stefan Wermter. The OMG-emotion behavior dataset. arXiv preprint arXiv:1803.05434, 2018.

[2] A. Nagrani, J.S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In INTERSPEECH, 2017.

[3] G. Bradski. The OpenCV Library. Dr. Dobb's Journal of Software Tools, 2000.

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

[5] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In CVPR, volume 1, 2017.

[6] Peng, S., Zhang, L., Ban, Y., Fang, M., & Winkler, S. (2018). A deep network for arousal-valence emotion prediction with acoustic-visual cues. arXiv preprint arXiv:1805.00638.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In BMVC, 2014.