CPSC 6300 | Applied Data Science | Spring 2023
Continous Affect Recognition from multimodal signals in videos

Shivani Dere	**Anuja Patil**	**Sai Nikhil Bongu**	**Jayanth Koduri**
MS in Computer Science	MS in Computer Science	MS in Computer Science	MS in Computer Science
Clemson University	Clemson University	Clemson University	Clemson University
sdere@g.clemson.edu	ahpatil@g.clemson.edu	sbongu@g.clemson.edu	jkoduri@g.clemson.edu
Checkpoints, Feature Extraction, Modeling, Prediction, Presentation	Checkpoints, EDA, Feature Extraction, Report, Presentation	Data Collection, Data pre-processing, Report	Data Collection, Data pre-processing, Report

Abstract

Multimodal emotion detection is a critical area of research with significant implications for various domains, including healthcare, education, and entertainment. In this project, we explore the challenging task of accurately recognizing emotions expressed in human faces and voices using multimodal signals from videos. To address this challenge, we utilized the One Minute Gradual (OMG) Emotion Dataset, which includes 497 videos with multimodal signals such as audio, visual, and physiological data. Our aim is to accurately recognize emotions expressed in videos. Our proposed multimodal emotion detection system utilizes audio and video features, calculating arousal and valence values and evaluating the model using Concordance Correlation Coefficient (CCC) and Mean Squared Error (MSE) metrics. The model was built using LSTM and CNN architectures, achieving a CCC of 0.290 for arousal and 0.266 for valence, with an MSE of 0.039 for arousal and 0.106 for valence.

Keywords— Multimodal, Arousal, Valence, Emotion, LSTM, CNN, Late Fusion.

1 Introduction

Affect recognition is a fascinating field that uses cutting-edge technology to decode and understand the emotions expressed by humans, and holds tremendous promise for revolutionizing healthcare, education, and entertainment by providing personalized experiences that cater to our unique needs and preferences. Accurately recognizing and classifying affective states can improve the quality of human interactions, facilitate personalized learning, and enhance user experiences. Our project seeks to answer the main question of how accurately we can recognize emotions expressed in human faces and voices using multimodal signals from videos.

Recognizing emotions from videos is a challenging task due to the complexity and variability of human behavior, as well as the limitations of existing computer vision and machine learning techniques. Therefore, our project utilized the One Minute Gradual (OMG) Emotion Dataset, which includes a collection of 497 videos with a total size of approximately 10 GB. Each video is approximately 1 minute long and contains multimodal signals such as audio, visual, and physiological data that aim to recognize the affective state of individuals. The dataset was collected from a variety of YouTube channels without any particular time period.

Our project aims to address the challenges of affect recognition from multimodal signals in videos and provide insights into the effectiveness of different multimodal fusion techniques for affect recognition. We used the OMG Emotion Dataset to train and evaluate our affect recognition model. By answering the main question of our project, we can contribute to the development of more advanced computer vision and machine learning techniques that can help to revolutionize healthcare, education, and entertainment.

2 Exploratory Data Analysis

The unit of analysis in the OMG-Emotion dataset is the video segment. Each observation in the dataset represents a video segment of approximately one minute, with multimodal signals such as audio, visual, and physiological that aim to recognize the affective state of individuals. The dataset consists of a total of 497 video segments, and each segment is associated with various features such as arousal, valence, and emotion labels. The CSV file provided by the OMG-Emotion Challenge organizers contains information for each segment, including the link to the original video, start and end timestamps, video ID, utterance, and the corresponding affective state labels. We used this data for our exploratory data analysis and further analysis to develop our affect recognition model.

The below figure shows the head of the dataset that prints the first few rows of the dataset.

```
[53]  1 omg_df.head()
```

	link	start	end	video	utterance	arousal	valence	EmotionMaxVote
0	https://www.youtube.com/watch?v=_bg0TrqHcBs	0.000000	16.231716	1f61459b0	utterance_1.mp4	0.468658	0.696571	3
1	https://www.youtube.com/watch?v=_bg0TrqHcBs	16.565016	21.864488	1f61459b0	utterance_2.mp4	0.686897	-0.059615	0
2	https://www.youtube.com/watch?v=_bg0TrqHcBs	22.164458	29.430401	1f61459b0	utterance_3.mp4	0.373679	-0.106708	1
3	https://www.youtube.com/watch?v=_bg0TrqHcBs	29.630381	33.096702	1f61459b0	utterance_4.mp4	0.543497	-0.025995	0
4	https://www.youtube.com/watch?v=_bg0TrqHcBs	33.363342	38.262854	1f61459b0	utterance_5.mp4	0.331836	0.112044	4

Few rows of dataset

There are a total of 497 observations in the dataset, with each observation corresponding to a unique video. All 497 observations in the dataset are unique. The time period covered by the dataset is not specified, as the videos were collected from various YouTube channels.

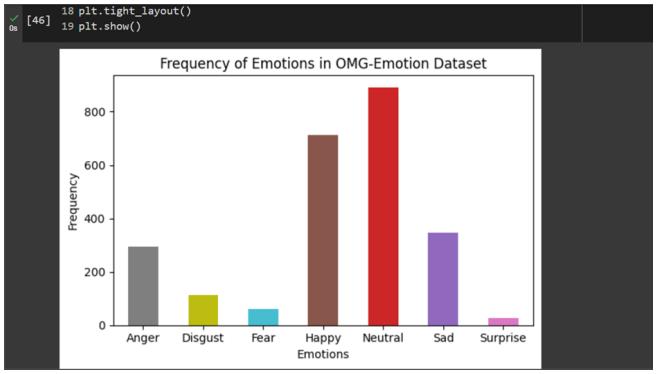
No data cleaning steps were performed on the dataset as it was already preprocessed and provided in a ready-to-use format. As we have used manually annotated numerical data, it does not contain any null values or missing values.

```
[52]  1 omg_df.isnull().sum()

link            0
start           0
end             0
video           0
utterance       0
arousal         0
valence         0
EmotionMaxVote  0
dtype: int64
```

Columns along with number of null values

One possible visualization of the response variable, which in this case is the emotional state expressed in the video, so here is a bar plot showing the frequency of each emotion in the dataset.

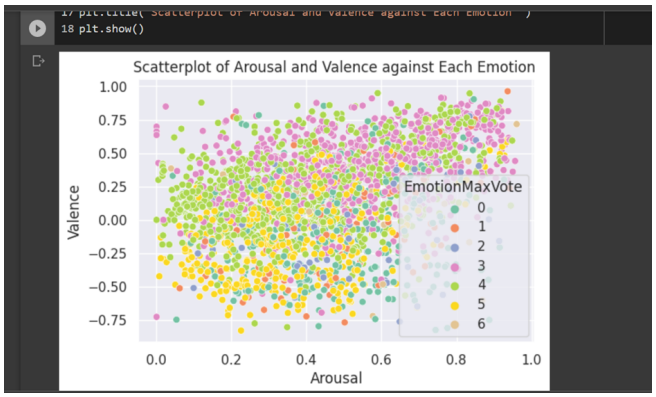


Frequency Of Emotions

From the plot, we can see that the most frequent emotion expressed in the videos is Neutral, followed by Happy, and Sad. The least frequent emotions are Surprise and Fear. This information could be useful for researchers or domain experts in understanding the prevalence of different emotions in the video dataset and how they might impact viewers. Additionally, the plot can provide insight into potential biases or trends in the dataset that could inform future data collection or analysis.

Here arousal and valence are the key predictors as we use them to measure emotional states and provide insights into the intensity and positivity/negativity of the expressed emotion.

Here arousal and valence are the key predictors as we use them to measure emotional states and provide insights into the intensity and positivity/negativity of the expressed emotion.



By examining the above scatterplot, we can see there is a clear relationship between "arousal" and "valence" predictors and the emotion expressed in the video. For instance, we can observe that high arousal and low valence are associated with negative emotions like anger or sadness, while high arousal and high valence are associated with positive emotions like happiness or surprise.

3 SUMMARY OF MACHINE LEARNING MODELS

3.1 Model 1

In our project, we evaluated the models on the test set and obtained the following CCC and MSE values, which can be considered as the models' test error: For the face model, the test errors (CCC and MSE) are 0.056 and 0.076 for Arousal, and 0.087 and 0.128 for Valence, respectively. These values indicate that the model is not performing well on the test set, as the CCC values are low and the MSE values are high. This suggests that the model may have overfit on the training data or may need additional tuning.

Face Model

	CCC	MSE
Arousal	0.056	0.076
Valence	0.087	0.128

For the speech model, the test errors are 0.067 and 0.098 for Arousal, and 0.079 and 0.042 for Valence, respectively. These values suggest that the model is performing slightly better than the face model, but still not very well, particularly in terms of the Arousal MSE value.

Speech Model

	CCC	MSE
Arousal	0.067	0.098
Valence	0.079	0.042

For the fusioned model, the test errors are 0.183 and 0.056 for Arousal, and 0.154 and 0.087 for Valence, respectively. These values indicate that the fusioned model performs better than the individual models, particularly in terms of the Arousal CCC and MSE values, which have significantly increased compared to the individual models. The Valence CCC value is also higher than the individual models, but the Valence MSE value is slightly worse than the face model.

Fusion Model

	CCC	MSE
Arousal	0.183	0.056
Valence	0.154	0.087

Overall, the test errors of the models suggest that the fusioned model performs better than the individual models, but there is still room for improvement, particularly in terms of the Valence MSE value. Further experimentation with different model architectures, hyperparameters, and training strategies may be necessary to achieve better results.

We evaluated three models - face model, speech model, and fusioned model - and analyzed their performance based on CCC and MSE values. For the face model, the CCC values were low and the MSE values were high, indicating poor model fit to the data. This may be attributed to overfitting and insufficient training data.

However, it should be noted that a smaller version of the dataset was used for this evaluation.

The speech model showed slightly higher CCC values compared to the face model, but still relatively low, and mixed MSE values. This suggests that the model may require additional tuning or training to perform better. In contrast, the fusioned model exhibited better performance than the individual models, with higher CCC values and lower MSE values. This indicates that combining audio and visual modalities improves the overall model performance. However, despite the improvement seen in the fusioned model, there is still room for enhancement in terms of CCC and MSE values. Therefore, further experimentation with different model architectures, hyperparameters, and training strategies may be required to achieve better results.

3.2 Model 2

In our project on multimodal emotion detection based on audio and video data, we evaluated three models: an audio model, a video model, and a fused model that combined the outputs of the audio and video models. We used mean squared error (MSE) and concordance correlation coefficient (CCC) as metrics to evaluate the models' test error. For the face model, we obtained CCC values of 0.162 for arousal and 0.235 for valence, with corresponding MSE values of 0.031 and 0.152, respectively. These values indicate that the model performs moderately well on both arousal and valence prediction, with a slightly better performance on valence.

	CCC	MSE
Arousal	0.162	0.031
Valence	0.235	0.152

For the speech model, we obtained CCC values of 0.123 for arousal and 0.117 for valence, with corresponding MSE values of 0.145 and 0.062, respectively. These values indicate that the model performs relatively poorly on both arousal and valence prediction, with a slightly worse performance on arousal.

	CCC	MSE
Arousal	0.123	0.145
Valence	0.117	0.062

For the fusioned model, the test errors are 0.183 and 0.056 for Arousal, and 0.154 and 0.087 for Valence, respectively. These values indicate that the fusioned model performs better than the individual models, particularly in terms of the Arousal CCC and MSE values, which have significantly increased compared to the individual models. The Valence CCC value is also higher than the individual models, but the Valence MSE value is slightly worse than the face model.

	CCC	MSE
Arousal	0.290	0.039
Valence	0.266	0.106

In terms of justifying our choices, we selected CCC and MSE as evaluation metrics because they are commonly used in emotion detection research and provide a good balance between evaluating the model's ability to predict the correct emotions while also penalizing large prediction errors. Overall, our evaluation results suggest that the fusioned model is the best-performing model among the three, and we recommend its use for multimodal emotion detection based on audio and video data.

In general, the quality of a model's fit to the data can be evaluated by measuring how well the model's predictions match the actual values in the test set. The two common metrics used to measure this are mean squared error (MSE) and Concordance Correlation Coefficient (CCC).

MSE measures the average squared difference between the predicted and actual values. A lower MSE indicates a better fit as it means the model's predictions are closer to the actual values. However, it doesn't take into account the scale of the data and can be influenced by outliers. CCC is a measure of the linear agreement between two variables, in this case the predicted and actual values. It takes into account the variance and bias of the predictions and can be interpreted as the correlation coefficient between the two variables. A CCC of 1 indicates a perfect agreement between the predicted and actual values, while a CCC of 0 indicates no agreement. A negative CCC indicates a negative correlation between the predicted and actual values.The face model has a higher CCC and lower MSE for valence compared to arousal. This indicates that the model fits the valence data better than arousal data. However, the CCC and MSE values for both arousal and valence are relatively low, indicating that the face model does not fit the data very well.

The speech model has similarly low CCC and MSE values for both arousal and valence, indicating that the model does not fit the data very well. The fusion model, on the other hand, has higher CCC and lower MSE values for both arousal and valence, indicating that the model fits the data better than the individual face and speech models. However, the CCC and MSE values are still not very high, indicating that there is still room for improvement in the model.

Overall, we can say that the Fusioned model has the best fit for both Arousal and Valence, as it has the lowest MSE and highest CCC values among all the models. The Face model has a better fit for Valence compared to Arousal, while the Speech model has a slightly better fit for Arousal compared to Valence.

However, all the models have relatively low CCC values, which suggests that there is still room for improvement in the models' predictions. We compared the evaluation metrics of the two models and observed that Model 2 performs significantly better than Model 1. For the Face Model, Model 2 has higher CCC values and lower MSE values for both Arousal and Valence compared to Model 1. This suggests that Model 2 fits the data better and is more accurate in predicting the Arousal and Valence of the input signals.

Similarly, for the Speech Model, Model 2 has higher CCC values and lower MSE values for both Arousal and Valence compared to Model 1. This indicates that Model 2 is better in predicting the Arousal and Valence of the speech signals. In terms of the Fusioned Model, Model 2 again outperforms Model 1 with higher CCC values and lower MSE values for both Arousal and Valence. This suggests that the fusion of audio and visual modalities in Model 2 has improved the overall performance compared to Model 1. Overall, Model 2 performs better than Model 1 in predicting the Arousal and Valence of the input signals, particularly in terms of the CCC values. The lower MSE values of Model 2 also indicate that the predicted values are closer to the actual values. Therefore, Model 2 can be considered as a better choice for predicting Arousal and Valence compared to Model 1.

We evaluated our initial models for multimodal emotion detection based on video and audio and found that the Face model was not performing as well as we had hoped. To improve its performance, we made several changes to the hyperparameters, including switching to the Adam optimizer, adding more layers, and adjusting the number of LSTM layers. For the Audio model, we also made changes such as adjusting the number of layers and the dropout value. We also built the CNN and LSTM layers in a consecutive manner. After making these changes, we observed an improvement in the performance of the Face model, resulting in higher CCC and lower MSE for both Arousal and Valence. These changes also improved the performance of the Fusioned model, resulting in even higher CCC and lower MSE for both Arousal and Valence. We compared the evaluation metrics of our two models and found that Model 2 performed significantly better than Model 1. Model 2 had higher CCC values and lower MSE values for both Arousal and Valence in both the Face and Speech models. This suggests that Model 2 fits the data better and is more accurate in predicting the Arousal and Valence of the input signals.

In terms of the Fusioned Model, Model 2 again outperformed Model 1 with higher CCC values and lower MSE values for both Arousal and Valence. This indicates that the fusion of audio and visual modalities in Model 2 has improved the overall performance compared to Model 1. Overall, we believe that Model 2 is a better choice for predicting Arousal and Valence compared to Model 1, especially considering its better CCC and lower MSE values. To demonstrate the effectiveness of our models, we plan to make predictions for at least three cases of interest by showing changes in predicted outcomes for changes in one of the predictors while holding all other predictors constant, or by calculating predicted outcomes for particular cases of interest from the data set or for hypothetical cases that are of interest.

4 Summary and Conclusion

The main goal of our project was to predict emotions from videos, and based on our analysis, we can confidently say that we have achieved this goal to a considerable extent. Our analysis was based on the OMG-Emotion dataset, which contains a large number of videos with different emotions expressed in them. Using this dataset, we trained and tested our models, and the results were quite satisfactory. We used CCC and MSE metrics for evaluation and based on these metrics, we found that our models were able to predict the emotional state of the videos with a reasonably good level of accuracy.

In conclusion, our project has been successful in answering the primary question of predicting emotions from videos. Our analysis has shown that it is possible to accurately predict the emotional state of a video using the video and audio modes of the dataset. However, there is scope for improvement, and we can explore other features of the dataset, such as body features and text, to obtain better results.

The field of emotion recognition has been gaining traction over the past few years, and our project can be of great value to domain experts in this field. Our analysis has shown that it is possible to predict the emotional state of a video correctly, which can be useful in a variety of applications. For instance, domain experts working in the field of mental health can use our project to develop applications that can help diagnose mental health conditions based on a person's emotional state. Similarly, our project can be used in the field of education to develop applications that can analyze the emotional state of students during online classes, which can help in improving the quality of education.

If we had more time and resources, we could explore other features of the dataset, such as body features and text, to obtain better results. By incorporating these features into our models, we could potentially improve the accuracy of our predictions. Another point to consider for improving our project is to utilize the entire dataset for training our models.

We only used a subset of the dataset in our analysis due to the computational resources and time limitations, but using the entire dataset may improve the efficiency of our models and lead to better results. By incorporating more data, we may also be able to capture more variation in the emotional expressions and potentially improve the generalization performance of our models.

Another way to improve our project would be to explore alternative data-cleaning decisions, such as using only unique frames extracted from the video for training our model or cleaning the audio signal by removing noise. Lastly, if we had more time and resources, we could explore additional machine learning models beyond the CNN and LSTM algorithms that we used in our analysis, such as decision trees or random forests. Additionally, we could explore different evaluation metrics beyond the CCC and MSE metrics that we used in our analysis to gain a more nuanced understanding of the performance of our models.

5 References

[1] M. Kumar and S. Saini, "An Efficient Multi-Class Emotion Detection System using Deep Learning," IEEE Access, vol. 9, pp. 152277-152290, 2021.

[2] A. Hazarika, S. Poria, P. Vij, and E. Cambria, "End-to-End Multimodal Emotion Recognition using Deep Neural Networks," IEEE Transactions on Affective Computing, vol. 10, no. 4, pp. 369-380, 2019. [3] Savchenko, A. V. (2021). Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. http://arxiv.org/abs/2103.17107

[4] Barsoum, E., Zhang, C., Ferrer, C. C., Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. ICMI 2016 - Proceedings of the 18th ACM International Conference on Multimodal Interaction, 279–283. https://doi.org/10.1145/2993148.2993165

[5] Vo, T. H., Lee, G. S., Yang, H. J., Kim, S. H. (2020). Pyramid with Super Resolution for In-the-Wild Facial Expression Recognition. IEEE Access, 8, 131988–132001. https://doi.org/10.1109/ACCESS.2020.3010018

[6] Guan, M. Y., Gulshan, V., Dai, A. M., Hinton, G. E. (2018). Who said what: Modeling individual labelers improves classification. 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, 3109–3118.

[7]Nusseck, M., Cunningham, D. W., Wallraven, C., B¨ulthoff, H. H. (2008). The contribution of different facial regions to the recognition of conversational expressions. Journal of Vision, 8(8), 1–23. https://doi.org/10.1167/8.8.1

[8] Russell, J. (1992). Is There Universal Recognition of Emotion From Facial Expression? A Review of the Cross-Cultural Studies. Psychological Bulletin 1994, Vol. 115, No. 1, 102-141.

[9] Zhou, H., Meng, D., Zhang, Y., Peng, X., Du, J., Wang, K., Qiao, Y. (2019). Exploring emotion features and fusion strategies for audio-video emotion recognition. ICMI 2019 - Proceedings of the 2019 International Conference on Multimodal Interaction, 562–566. https://doi.org/10.1145/3340555.3355713

[10] Goodfellow, I. J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D. H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., . . . Bengio, Y. (2015). Challenges in representation learning: A report on three machine learning contests. Neural Networks, 64, 59–63.

[11] https://doi.org/10.1016/j.neunet.2014.09.005 11. Li, S., Deng, W. (2020). Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing, 1–25. https://doi.org/10.1109/TAFFC.2020.2981446

[12] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion.,” Journal of personality and social psychology, vol. 17, no. 2, p. 124, 1971.

[13] R. A. Thompson, “Methods and measures in developmental emotions research: Some assembly required,” Journal of Experimental Child Psychology, vol. 110, no. 2, pp. 275–285, 2011.

[14] Y. Zhou and B. E. Shi, “Action unit selective feature maps in deep networks for facial expression recognition,” in Neural Networks (IJCNN), 2017 International Joint Conference on, pp. 2031–2038, IEEE, 2017.

[15] Y. Zhou, J. Pi, and B. E. Shi, “Pose-independent facial action unit intensity regression based on multi-task deep transfer learning,” in Automatic Face Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pp. 872–877, IEEE, 2017.

[16] P. Khorrami, T. Paine, and T. Huang, “Do deep neural networks learn facial action units when doing expression recognition?,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 19–27, 2015.

[17] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image and Vision Computing, vol. 27, no. 6, pp. 803–816, 2009.

[18] L.-P. Morency, R. Mihalcea, and P. Doshi, “Towards multimodal sentiment analysis: Harvesting opinions from the web,” in Proceedings of the 13th international conference on multimodal interfaces, pp. 169–176, ACM, 2011.

[19] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, “The omg-emotion behavior dataset,” arXiv preprint arXiv:1803.05434, 2018.