Deep Multimodal Emotion Recognition: Fusion of Facial Features and Neurophysiological Signals

Farshad Safavi
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, USA
fsafavi1@umbc.edu

Vikas Reddy Venkannagari
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, USA
wh22215@umbc.ed

Ramana Kumar Vinjamuri
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, USA
rvinjam1@umbc.edu

2024 IEEE 20th International Conference on Body Sensor Networks (BSN) | 979-8-3315-3014-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/BSN63547.2024.10780554

Abstract—Multimodal emotion recognition through the fusion of facial and neurophysiological features plays an important role in various applications, such as advertising, the automotive industry, wearable devices, and human-computer interactions. Fusing human facial expressions and neurophysiological signals traditionally requires domain-specific knowledge and complex preprocessing steps. However, with the advent of deep learning, we can fully leverage the end-to-end capabilities of these techniques for the intermediate integration of facial and neurophysiological signals in emotion recognition systems. As a result, we introduce a novel end-to-end deep network that leverages transformers to learn rich feature representations of neurophysiological signals, integrated with a transformer-inspired technique for facial expression recognition and emotion classification. By integrating transformers and deep neural networks, our approach successfully captures complex temporal and spatial patterns in the data. This combination allows for more robust analysis, enhancing the system's overall performance in recognizing and classifying emotions accurately. We validated our approach through experiments on the well-known DEAP dataset, achieving performance comparable to the state-of-the-art, with accuracy rates of 97.64% for valence and 97.78% for arousal.

Keywords— Multimodal Emotion Recognition, Affective Computing, Deep learning, Emotion Detection, Transformer

I. Introduction

Emotion recognition, which enables machines to respond to human emotional states, has become a crucial component of human-computer interaction. Using emotion recognition systems, applications can be designed to enhance user experience across various fields such as education, healthcare, task monitoring, and the autonomous driving industry. Emotion recognition can be explained by the dimensional theory, which uses two dimensions: valence and arousal. Valence measures pleasure (pleasant to unpleasant), while arousal assesses energy levels (calm to energized). Emotions are represented as combinations of these two dimensions. In this research, we apply this theory to measure emotion. Emotions play a key role in effective communication in social contexts, consequently facilitating smooth human-computer or human-robot interactions [1]. However, most computer systems currently exhibit a significant deficiency in empathy and emotional intelligence, limiting their ability to interact authentically and effectively with humans. To address this issue, we are developing a multimodal emotion recognition system. Our work investigates two modalities of data for the emotion recognition task, including neurophysiological signals and vision. We used a deep neural network architecture to extract features from each modality (bio-sensing or vision). These features are then fused to produce more accurate results. This system assesses the emotional responses of users. It accurately recognizes their expressions by classifying them using dimensional representation, which can predict the full spectrum of emotions on the DEAP dataset [2].

Facial expression recognition has become a popular technique with the advent of deep learning, as it allows for the recognition of direct indicators of human emotions that are easily captured by cameras. Convolutional Neural Networks (CNNs) and their variations are highly effective in processing visual data for emotion recognition. For facial image sequences or videos, advanced techniques such as Recurrent Convolutional Neural Network (RCNN) [3], and a combination of CNN and Long Short-Term Memory (LSTM) are proficient in extracting spatial and temporal facial expression features [4]. With the rise in popularity of transformers in different applications [5], [6], new facial expression recognition methods are being developed using Mix Transformers [7].

Physiological and bio-sensing signals such as electroencephalogram (EEG), electrocardiogram (ECG), and galvanic skin response (GSR) data are used for emotion recognition. More recent advanced models have explored transformers and graph-based methods. Hybrid models combining CNNs with sparse autoencoders and deep neural networks have demonstrated high accuracy [8]. Additionally, models incorporating attention mechanisms and regional feature extraction through graph convolutional networks have shown promising results [9]. These studies underscore the potential of sophisticated neural network architectures in significantly advancing the field of emotion recognition from EEG data.

While single-modality approaches offer distinct advantages, integrating multiple modalities can yield more comprehensive and salient features. For instance, relying solely on facial expression recognition may be problematic due to the potential for deceptive expressions. However, when combined with neurophysiological signals, which are not subject to voluntary control, the reliability of the results is significantly enhanced. Multimodal affective computing has gained attention for its potential to improve emotion recognition accuracy by leveraging the strengths of different modalities. However, most of these methods use audio and video to recognize emotion, while bio-sensing and vision have received less attention. Our study addresses this gap by employing a deep multimodal fusion approach that utilizes both neurophysiological signals and vision modalities for

This research was supported by the National Science Foundation (CAREER Award HCC-2053498).

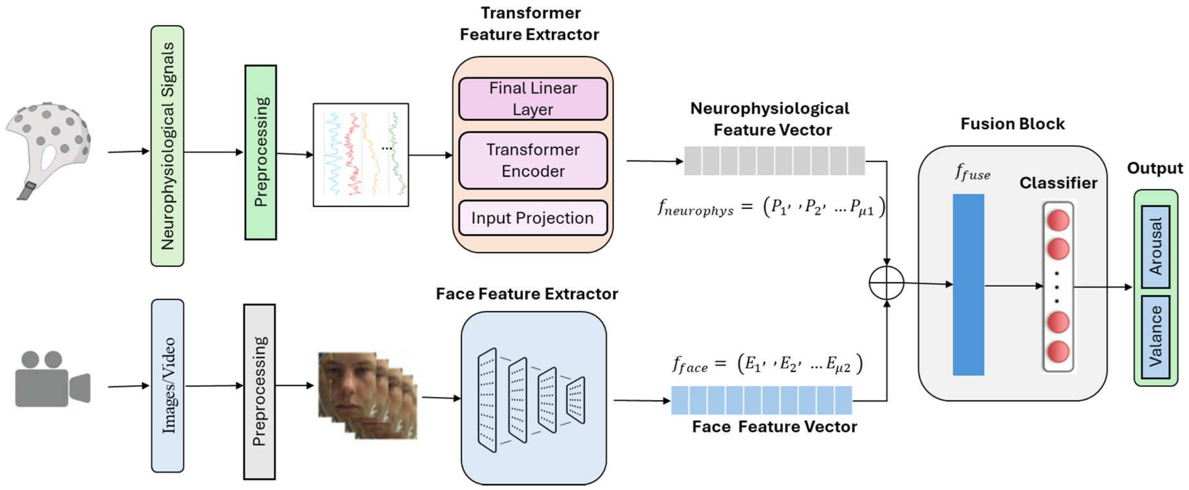

Fig. 1. The multimodal architecture integrates neurophysiological and facial features for emotion prediction, utilizing transformers for neurophysiological data and a facial expression extraction network. Feature vectors $f_{neurophys}$ and f_{face} are passed into a fusion block to produce f_{fuse}, determining the emotional dimensions.

emotion recognition. We utilize transformer, known for its effectiveness in predicting time-series data, to process neurophysiological data and capture fine-grained features. Additionally, we introduce a novel facial feature extraction method that leverages a transformer-inspired technique for patch extraction.

Our main contribution is proposing a novel end-to-end deep network for multi-modal emotion recognition using both facial and neurophysiological signals. Our deep multi-modal network employs Transformers to learn a rich feature representation of neurophysiological signals, jointly with unique facial expression recognition using a transformer-inspired technique for emotion classification. By combining both methods, we enhance the accuracy and reliability of emotion recognition system. We performed experiments on the well-known DEAP dataset, achieving performance comparable to the state-of-the-art, with 97.64% and 97.78% (valence/arousal).

A. Modeling Facial Emotion and Bio-sensing Signals

To formalize our approach, we define the input data and model structure as follows. Assume that for the $\boldsymbol{i_{th}}$ set of facial emotion video frames, we have $\boldsymbol{I_i = \{I_t | 1, \dots, n_i\}}$, where $\boldsymbol{I_t}$ denotes the $\boldsymbol{t_{th}}$ image. Additionally, we have neurophysiological signals, $\boldsymbol{N_i = \{e_t \,| 1, \dots, m_i\}}$ where $\boldsymbol{e_t}$ signifies the $\boldsymbol{t_{th}}$ data point. The lengths of the video and neurophysiological signals are represented by $\boldsymbol{n_i}$ and $\boldsymbol{m_i}$, respectively. For the $\boldsymbol{i_{th}}$ instance, the ground-truth annotation $\boldsymbol{y_i}$ indicates either a valence or arousal value. We train a comprehensive model $\boldsymbol{G_\theta}$ using tuples $\boldsymbol{\{(I_i, N_i), y_i : i \in [0, T]\}}$, where $\boldsymbol{y_i} \in \boldsymbol{[0, 1]}$, and T is the total number of instances in the dataset. Here, $\boldsymbol{y_i}$ corresponds to low valence/arousal, and high valence/arousal. During the prediction phase, given a test video and neurophysiological signal pair $\boldsymbol{(I_j, N_j)}$ as input, $\boldsymbol{G_\theta}$ estimates $\hat{y}_j$, which approximates the ground-truth annotation $\boldsymbol{y_j}$ as follows: $\hat{y}_j = \boldsymbol{G_\theta((I_j, N_j);\ \Theta)}$.

II. Multimodal Architecture

Our multimodal architecture is inspired by the DeepVaNet model [4], utilizing both neurophysiological and facial features to predict emotional states. As depicted in Figure 1, the architecture of our model includes a transformer-based neurophysiological feature extractor, a facial feature extractor, and a fusion block. In this study, we utilize transformers for neurophysiological signals, which have demonstrated their significance in predicting time-series data. Additionally, we employ a unique neural network that combines convolutional layers for initial patch extraction and embedding, followed by an LSTM for sequence processing to extract facial expressions. Preprocessing steps involved downsampling to 128 Hz, removing electrooculography (EOG) artifacts, segmentation, and baseline removal by subtracting each subject's initial 3 seconds of resting-state data. The EEG data was then divided into one-second intervals. For face preprocessing, 5 frames per second were extracted, and image cropping to a size of 64 x 64 pixels was performed based on facial landmarks.

A. Neurophysiological Feature Extractor

The transformer model is designed for neurophysiological feature extraction, leveraging the capabilities of transformer architectures to handle temporal dynamics in neurophysiological signals. By projecting the input into a higher-dimensional space, employing multi-head attention, and utilizing residual connections with normalization, our model has a robust feature extraction suitable for downstream classification tasks. The data flow in the transformer-based architecture for processing neurophysiological signals begins with the input data, which has a shape of (40, 128), representing 40 neurophysiological signals, including the integration of thirty-two EEG (electroencephalogram) and eight physiological signals (EOG: electrooculogram, GSR: galvanic skin response, BVP: blood volume pulse, RSP: respiration, EMG: electromyogram, SKT: skin temperature, and pulse wave), all sampled at a rate of 128 Hz. This data can be related to NLP, with 128 words and 40-word embeddings. We apply this analogy to bio-sensing data to capture the advantages of transformers in understanding complex temporal patterns and contextual relationships. In our implementation, the neurophysiological feature extractor processes the data through an encoding process where the tensor is permuted along the sequence length dimension (128, 40) as per our analogy. As depicted in Figure 1, the tensor is then sent to the input projection layer in transformer feature extractor. This projected tensor is then passed through the transformer encoder, which consists of layers of attention and feedforward neural networks. The attention mechanism allows the model to capture long-range dependencies within the sequence. Finally, we flatten the output of the transformer block, and a linear layer is mapped to get the desired number of feature vectors, resulting in an output of neurophysiological feature scores. This final output represents the encoded features of the input neurophysiological data. This feature vector, represented as a 1×μ1 vector, is denoted as: $f_{neurophys} = (P_1, P_2, \dots, P_{\mu 1})$.

B. Facial Expression Feature Extractor

The facial expression feature extractor network is designed to process a sequence of facial images and extract meaningful features for fusion. This network combines the strengths of Convolutional Neural Networks (CNNs) and sequence models. The process begins with the input images, which are preprocessed and individually passed through a pretrained CNN on the AEFW dataset [10]. This CNN processes each image to produce feature maps with reduced spatial dimensions. The model then processes this sequence of feature maps to capture temporal dependencies across the sequence of images. We use a unique technique inspired by transformers in the architecture of our facial expression extractor, which employs a convolutional embedding to reduce the spatial dimensions and then handle the temporal sequence. The output is flattened and passed into an LSTM.

After processing, the output features are passed through a fully connected layer to reduce the dimensionality to the desired feature size. The final output is a feature vector capturing the essential facial features from the input image sequence. This feature vector, represented as a 1×μ2 vector, is denoted as: $f_{face} = (E_1, E_2, \dots, E_{\mu 2})$.

C. Fusion Block

The fusion block is designed to fuse facial and bio-sensing data for tasks such as emotion prediction. It incorporates two main feature extractors: a facial feature extractor and a neurophysiological extractor. The facial feature extractor reduces spatial dimensions using convolutional embedding and captures essential facial features. The neurophysiological extractor is implemented using a transformer model, which processes bio-sensing data by capturing long-range dependencies within the sequence and encoding it into a feature vector. For feature-level fusion, the face appearance feature f_{face} and the neurophysiological feature $f_{neurophys}$ are concatenated to generate a multi-modal feature vector $f_{fuse} = f_{face} \oplus f_{neurophys}$. This combined feature vector is passed through a multi-layer classifier that sequentially reduces the dimensionality and applies activation functions, culminating in a sigmoid function to produce the final output.. This fusion approach leverages the data's spatial and temporal aspects, making it highly effective for complex and dynamic tasks, particularly in emotion prediction from multi-modal inputs.

III. Experiment Setup

A. DEAP Dataset

Dataset for emotion Analysis using physiological signals dataset is a widely used multimodal dataset designed to analyze human emotional states. This dataset comprises physiological data collected from 32 participants (16 males and 16 females) while they watched 40 music videos, each selected for its potential to elicit a range of emotional responses. EEG signals were recorded using a 32-channel electrode cap conforming to the “10-20” international standard at a sampling frequency of 512 Hz. For emotion classification, the DEAP dataset provides labels for valence and arousal, enabling the categorization of emotional states into distinct classes. In addition to EEG data, video recordings of facial expressions were made for 22 of the 32 participants, providing a rich dataset for multimodal emotion analysis. The DEAP dataset offers a comprehensive resource for investigating the neural and physiological correlates of emotion, facilitating the development and evaluation of emotion recognition models across various modalities.

B. Training and Validation

We use binary cross-entropy as the loss function. We consider the training batch size, the target emotion label, and the predicted score. The loss is calculated based on these elements. We pass the test video and physiological signal during inference through our proposed network to obtain a fusion score. Based on this score, the final prediction is made: if the score is greater than 0.5, the prediction is "High"; otherwise, it is "Low". We train and test our model on each subject, a process referred to as a per-subject experiment. Our model undergoes 10-fold cross-validation, and the average testing accuracy is used to measure performance. For validation, we utilize the mean recognition accuracy of both valence and arousal. The feature size for face appearance and neurophysiological signals is 64.

IV. Results and Discussion

To evaluate the performance of our proposed model on emotion recognition, we compared it with recent techniques of single modality and multimodality in the same DEAP Dataset. Table I provides a performance evaluation of these models. To compare our model to single modality facial expression recognition, we utilized Emonet [11], a deep neural network architecture specifically designed to analyze facial affect under naturalistic conditions with high accuracy. Two distinct methods were employed to evaluate Emonet accuracy. The first method involves using a fixed threshold, where the labels are based on ratings equal to or above 5; we call it Emonet V1 in Table I. The second method, in contrast, employs the mean of the distribution as a dynamic threshold while applying a similar rating-based approach for the determination of labels; we called it Emonet V2 in Table I. Our results in Table I demonstrate the significant gap between the average accuracies of facial expression recognition (53.52% and 53.72% obtained from versions V1 and V2 of Emonet) and our multimodal fusion model, which achieved an average accuracy of 97.71%. This result indicates a significant improvement when using multimodal fusion compared to single-modal facial expression recognition.

Table I shows models that have solely used EEG data to predict emotions. Our model, with an accuracy of 97.64%, outperforms the EEG Graph-based model (60.18% accuracy, 59.19% arousal) [11], the CNN multispectral model (90.62% accuracy, 86.13% arousal) [12], and the 3D CNN model (96.61% accuracy, 96.43% arousal) [13]. It also surpasses the DCNN model, which has a valance accuracy of 87.84% and arousal of 87.69% [14]. This evidence confirms the superiority of multimodal fusion over single-modality

TABLE I
PERFORMANCE EVALUATION OF NETWORKS ON DEAP

EEG Models	**Valance**	**Arousal**	**Average**
EEG Graph-based	60.18	59.19	59.68
CNN multispectral	90.62	86.13	88.38
3D CNN	96.61	96.43	96.52
DCNN	87.84	87.69	87.76
Facial Models			
Emonet V1	48.16	58.87	53.52
Emonet V2	51.26	56.19	53.72
Fusion Models			
3D CNN ensemble	96.13	96.79	96.46
CNN and attention	96.63	98.18	97.40
Our Model	**97.64**	**97.78**	**97.71**

Table I evaluates EEG-based, facial expression-based, and fusion models. Integrating facial and neurophysiological features, our model achieves the highest average accuracy, surpassing all other models.

approaches. In addition, our fusion model, with an accuracy of 98.18%, outperforms the 3D CNN ensemble model (96.13% valence, 96.79% arousal) [3] and the CNN and attention model (96.63% valence, 98.18% arousal) [15]. The superiority of our proposed model can be attributed to the following reasons. First, our model uses a transformer-based architecture for EEG data to capture complex spatial and temporal patterns effectively. In addition, the multi-head attention block contributes to capturing different data dependencies, allowing for more comprehensive feature extraction. Moreover, the facial expression extractor, utilizing a unique technique inspired by transformers, obtains meaningful spatial features from images, and LSTM further refines these features to capture temporal dependencies. Finally, the integration of the transformer-based model for EEG data and the transformer-inspired facial emotion recognition, followed by the fusion of their extracted features, provides a robust framework for emotion prediction. This architecture effectively captures and integrates diverse aspects of neurophysiological and visual data, resulting in improved accuracy and reliability of emotion prediction. As presented in Table I, our proposed model exhibits superior accuracy compared to the evaluated methods, resulting in enhanced performance in emotion recognition tasks. Furthermore, incorporating both facial and EEG data demonstrates an enhancement in performance over using EEG data or facial expression data separately.

V. Conclusion

This research presents a novel multimodal emotion recognition system that leverages neurophysiological signals and vision to enhance accuracy and reliability. The network accepts face image sequences and neurophysiological signals (e.g., EEG, EOG, ECG, GSR, etc.) as input, yielding valence-arousal labels for emotion recognition. By integrating transformers and deep neural networks, our approach successfully captures the data's complex temporal and spatial patterns. Experiments on the DEAP dataset show that our model outperforms existing single-modality methods and is comparable to state-of-the-art multimodal methods. These results underscore the potential of multimodal fusion in advancing emotion recognition technology, paving the way for more empathetic and emotionally intelligent human-computer interactions and enhanced usability in wearable technology. Additionally, the limited availability of datasets containing both facial and EEG signals presents a significant challenge for research in this area, which is one of the areas we can work on to advance the field further.

VI. References

[1] F. Safavi *et al.*, “Emerging Frontiers in Human–Robot Interaction,” *J. Intell. Robot. Syst.*, vol. 110, no. 2, 2024.

[2] S. Koelstra *et al.*, “DEAP: A Database for Emotion Analysis ;Using Physiological Signals,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, 2012, doi: 10.1109/T-AFFC.2011.15.

[3] E. S. Salama, R. A. El-Khoribi, M. E. Shoman, and M. A. Wahby Shalaby, “A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition,” *Egypt. Informatics J.*, vol. 22, no. 2, pp. 167–176, 2021, doi: https://doi.org/10.1016/j.eij.2020.07.005.

[4] Y. Zhang, M. Z. Hossain, and S. Rahman, “DeepVANet: A Deep End-to-End Network for Multi-modal Emotion Recognition,” in *Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30 – September 3, 2021, Proceedings, Part III*, 2021, pp. 227–237. doi: 10.1007/978-3-030-85613-7_16.

[5] F. Safavi and M. Rahnemoonfar, “Comparative Study of Real-Time Semantic Segmentation Networks in Aerial Images During Flooding Events,” *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 15–31, 2023, doi: 10.1109/JSTARS.2022.3219724.

[6] F. Safavi, T. Chowdhury, and M. Rahnemoonfar, “Comparative Study Between Real-Time and Non-Real-Time Segmentation Models on Flooding Events,” in *2021 IEEE International Conference on Big Data (Big Data)*, 2021, pp. 4199–4207. doi: 10.1109/BigData52589.2021.9671314.

[7] F. Safavi, K. Patel, and R. K. Vinjamuri, “Towards Efficient Deep Learning Models for Facial Expression Recognition using Transformers,” in *2023 IEEE 19th International Conference on Body Sensor Networks (BSN)*, 2023, pp. 1–4. doi: 10.1109/BSN58485.2023.10331041.

[8] J. Liu *et al.*, “EEG-Based Emotion Classification Using a Deep Neural Network and Sparse Autoencoder,” *Front. Syst. Neurosci.*, vol. 14, 2020, doi: 10.3389/fnsys.2020.00043.

[9] Z. Fan, F. Chen, X. Xia, and Y. Liu, “EEG Emotion Classification Based on Graph Convolutional Network,” *Appl. Sci.*, vol. 14, no. 2, 2024, doi: 10.3390/app14020726.

[10] J. Kossaifi, G. Tzimiropoulos, S. Todorovic, and M. Pantic, “AFEW-VA database for valence and arousal estimation in-the-wild,” *Image Vis. Comput.*, vol. 65, pp. 23–36, 2017, doi: https://doi.org/10.1016/j.imavis.2017.02.001.

[11] A. Toisoul, J. Kossaifi, A. Bulat, G. Tzimiropoulos, and M. Pantic, “Estimation of continuous valence and arousal levels from faces in naturalistic conditions,” *Nat. Mach. Intell.*, 2021, [Online]. Available: https://www.nature.com/articles/s42256-020-00280-0

[12] M. A. Ozdemir, M. Degirmenci, E. Izci, and A. Akan, “EEG-based emotion recognition with deep convolutional neural networks,” *Biomed. Eng. / Biomed. Tech.*, vol. 66, no. 1, pp. 43–57, 2021, doi: doi:10.1515/bmt-2019-0306.

[13] Y. Zhao, J. Yang, J. Lin, D. Yu, and X. Cao, “A 3D Convolutional Neural Network for Emotion Recognition based on EEG Signals,” in *Proceedings of the International Joint Conference on Neural Networks*, 2020, pp. 1–6. doi: 10.1109/IJCNN48605.2020.9207420.

[14] Y. An, N. Xu, and Z. Qu, “Leveraging spatial-temporal convolutional features for EEG-based emotion recognition,” *Biomed. Signal Process. Control*, vol. 69, p. 102743, 2021, doi: https://doi.org/10.1016/j.bspc.2021.102743.

[15] S. Wang, J. Qu, Y. Zhang, and Y. Zhang, “Multimodal Emotion Recognition From EEG Signals and Facial Expressions,” *IEEE Access*, vol. 11, pp. 33061–33068, 2023, doi: 10.1109/ACCESS.2023.3263670.