Context-Aware Deep Emotion Recognition with Physiological Signals

Anonymous
Anonymous Institution
Anonymous Email

Anonymous
Anonymous Institution
Anonymous Email

Anonymous
Anonymous Institution
Anonymous Email

***Abstract*—Emotion recognition holds multifaceted relevance across various fields, necessitating researchers to balance precision and applicability in developing emotion recognition systems. While existing systems of emotion recognition with physiological signals demonstrate advantages in both criteria, a gap exists in correlating physiological signals with shifts in emotional states instead of emotional states themselves. To leverage this concern, this paper proposes a deep learning pipeline that includes previous time stages' emotions for the prediction of future time stages' emotions. The proposed pipeline integrates both discrete emotional states and continuous valence-arousal data, and it achieves an overall accuracy of 0.66 across 12 emotion classes in the Continuously Annotated Signals of Emotion (CASE) dataset. The results show that it exhibits a comparable accuracy level to existing methodologies while advancing in a more detailed spectrum of emotion recognition.**

I. Introduction

The pursuit of constructing emotion recognition systems has incurred substantial attention due to their applications in various domains, including but not limited to healthcare [1], human-computer interaction [2], entertainment [3], and mental health monitoring [4]. Researchers have thrived to recognize emotion as being discrete emotional states [5] or continuous valence and arousal data [6]. Recognition is done on various choices of signals, including facial expression [5], brain image [7], speech [6], [18], and physiological signals [8].

Since emotion recognition is usually based on analyzing human-related information, their accuracy and availability are both key consideration factors. Among the different approaches for emotion recognition, physiological signals analysis presents several distinct advantages for both factors. Physiological signals are generated by the human body continuously and involuntarily, not relying on explicit actions or controlled conditions. In addition, physiological signals can be captured non-invasively and passively, allowing for natural and unobtrusive data collection. This availability facilitates real-time monitoring of emotional states in diverse settings without necessitating active user involvement, thereby enabling the capture of spontaneous and genuine emotional responses [8]. Furthermore, physiological signals often offer a high degree of accuracy and reliability in reflecting emotional arousal and valence due to their direct connection with the autonomic nervous system's responses [9].

For different purposes of research, emotion recognition aims to recognize either discrete emotional states [5] or continuous valence and arousal data [6], [9]. Discrete emotion recognition is centered on categorizing specific emotional states and provides a structured framework for understanding and labeling emotions. It also allows for precise classification and targeted responses. On the other hand, continuous emotion assessment, focusing on dimensions of valence and arousal, evaluates emotion as a spectrum, and is capable of offering a more comprehensive representation of emotion experiences. In this research, the two approaches are combined, leveraging both of their advantages. Specifically, the emotion prediction of valence and arousal is classified into 12 distinct emotional states, as demonstrated in Figure .1.

Emotion recognition using Neural Networks has exhibited notable practicality, having undergone extensive exploration into the efficacy of diverse deep learning models. Previous studies have predominantly employed Deep Convolutional Neural Networks (DCNN) for emotion recognition utilizing physiological data [14]. Subsequent inquiries into Long-Short Term Memory (LSTM) models and the transformer model have showcased proficiency in handling both discrete and continuous emotion recognition scenarios [10], [11], [15]. Inspired by these works, this study introduces a deep learning algorithm that integrates elements from both LSTM and the transformer model. The outcomes are compared against established benchmarks proposed in earlier literature.

Existing research on emotion recognition investigated how physiological signals are capable of predicting emotion on independent time stages [8]. However, these approaches often overlook how physiological signal correlates with the dynamics of emotion between time stages. To address the correlation, this paper utilizes a custom mechanism for prediction, analyzing the shifts in emotional states by utilizing previous emotional information on current emotion stage prediction. The prediction of emotion in this work integrates the two approaches of continuous and discrete. The proposed algorithm of this work demonstrated performance comparable to previous work on the Continuously Annotated Signals of Emotion (CASE) dataset [12].

II. Method Overview

In this study, the analysis of physiological data collection involves Electrocardiography (ECG), Blood Volume Pulse (BVP), Galvanic Skin Response (GSR), and Respiration. ECG measures the electrical activity of an individual's heart over

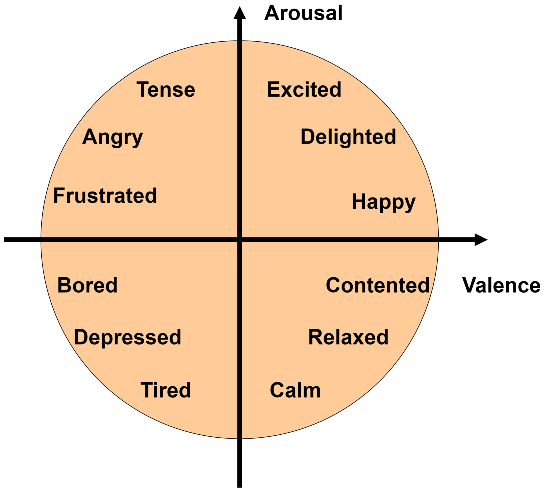

Fig. 1. The emotional model proposed by Robert Plutchik [13]. In his theory, emotion is determined by valence and arousal, and different combinations of valence and arousal would result in 12 different types of emotion.

time. BVP, similarly, measures changes in blood volume within blood vessels. GSR measures the skin's electrical conductance, and respiration data reflects breathing patterns. In addition, the analysis of emotion data exhibits two stages. It is first done on valence and arousal values, and then classified as discrete emotional states according to the widely applied Robert Plutchik's emotional cycle shown in Figure. 1 [13].

A. Data Processing

Data collection and data processing are done with one measurement channel and one rating channel. On the measurement channel, the participant would put on sensors, whose timely voltage signal would be recorded. On the rating channel, participants are asked to provide valence and arousal ratings while viewing or listening to certain stimuli, which are then aligned with the physiological signals measured at the corresponding times [12].

Between different stimuli videos, there exists a time gap where no stimuli video is played and participants are asked to wait. Gaps are defined based on the end of time of each stimulus. To maintain consistency of emotion recognition, the ratings during time gaps are removed and not considered, enabling analysis of consciousness levels among participants to be limited to the stimulation processes. Following the collection of all continuous data, they undergo processing based on time segmentation. The time duration per 10-second segment and 5-second segment is arbitrarily selected, enabling a comparative analysis. For each proficient time stage's valence and arousal data, where no less than 10 seconds of preceding stimuli have occurred, the corresponding sequence of physiological signals, valence data and arousal data of the time-lapse before it are paired with the time stage's valence and arousal value, forming one single data point.

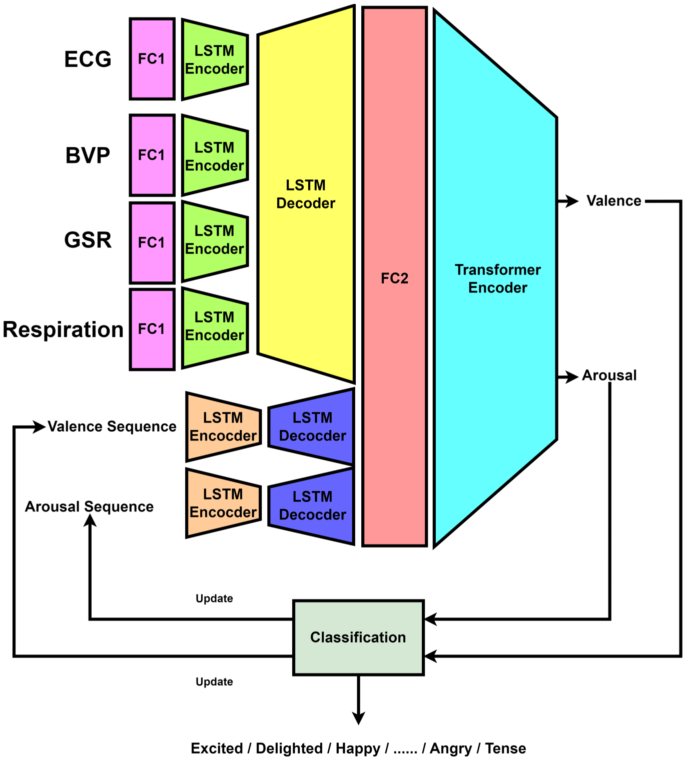

Fig. 2. Neural network architecture. LSTM refers to the Long-Short Term Memory layer, and FC refers to the Fully Connected layer. Prediction results of valence and arousal would be classified into a discrete emotional state according to the emotion cycle in Figure .1, and then used to update the sequence of prior emotional data for the next time stage's prediction.

B. Prediction Pipeline

Upon the training phase, the neural network aims to predict the valence and arousal levels for each time stage. The four physiological data streams undergo initial processing through a fully connected layer before being fed into a three-layer Long Short-Term Memory (LSTM) encoder. Subsequently, their outputs are concatenated and fed into a three-layer LSTM decoder. Simultaneously, ground truth values of valence and arousal data over the 10-second duration go through a one-layer LSTM encoder-decoder network. After that, the outputs from the physiological data and the valence-arousal statistics are concatenated and passed through a fully connected layer and by a LeakyReLU to allow for the additional complexity, a measure that had been proven to be effective in previous works [6]. The output would be fed into a five-layer transformer encoder. The final output of this stage comprises two values, which are the predicted current valence and arousal levels. The loss is calculated by the absolute difference of the ratio between predicted and ground truth valence and arousal values. The detailed network architecture can be found in Figure. **??**.

Upon the prediction phase, the predicted valence and arousal values of the previous time stage would be used to update the valence and arousal sequence input of the network. As referred to in Figure. **??**, the update step is done by removing the oldest value and adding itself into the sequence, forming

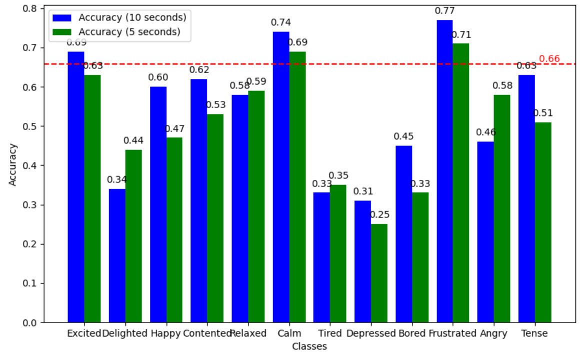

Fig. 3. Result of the prediction pipeline, evaluated on 40,591 data points extracted from the CASE dataset with Leave-One-Out Cross Validation (LOOCV). The blue bar represents the accuracy of predicting each emotion class on the 10-second time-lapse, and the green bar represents the prediction on the 5-second time-lapse. The overall accuracy across all prediction classes is 0.66 for 10 10-second time-lapse, as marked by the dashed line. The overall accuracy of 5 seconds time-lapse is 0.61.

a new sequence of valence and arousal that enables prediction of the valence and arousal value of the new time stage.

The ultimate output of each time stage's valence and arousal would be arbitrarily taken ratio and mapped to the emotion cycle proposed in Figure.1, where the time stage's emotion status would be classified into a discrete state emotion. Specifically, the ratio between the predicted valence value and arousal value would be matched to the emotion class that demonstrates the closest absolute distance with the ratio of the predicted valence value and arousal value.

III. Preliminary Evaluation

In the current stage of research, experimental environments have not been fully established. To evaluate the proposed method, a custom dataset is employed and performance is compared with various baseline models running on the same dataset. The training and testing employ Leave-One-Out Cross Validation (LOOCV), where each of the 30 participants' data in the dataset is used as the testing set and the rest of the 29 participants as the training set. The results of the validations are then averaged.

A. Dataset

The preliminary evaluations of the proposed approach entail an examination on the Continuously Annotated Signals of Emotion (CASE) dataset [12]. The CASE dataset includes 30 volunteers' self-evaluated valence and arousal data during the viewing of 8 selected video stimuli via a Joystick-based Emotion Reporting Interface (JERI). The videos were meticulously chosen to elicit a range of emotions, including amusement (high valence, high arousal), relaxation (high valence, low arousal), fear/anger (low valence, high arousal), and boredom (low valence, low arousal). The dataset includes 8 categories of physiological data. This study employs 4 of them (ECG, BVP, GSR, Respiration) for analysis.

B. Baselines

The efficacy of the proposed methodology is also compared against several established baselines. The baselines encompass diverse approaches:

- DCNN [14]: Four-layer 1D convolutional neural network and utilizes a three-layer fully connected network for subsequent classification.
- Attn-BiLSTM [15]: Multilayer bidirectional LSTM utilized to capture temporal information from multimodal signals, followed by fully connected layers of classifier.
- MulT [10]: Transformer-based multimodal fusion method applied to video, audio, and text. Includes initial processing of unimodal data through a temporal convolutional network to acquire low-level features, followed by transformers employing cross-modal attention and self-attention mechanisms for fusion.
- Transformer-SS [11]: Self-supervised learning framework designed to learn generalized representations from an extensive pool of unlabeled samples. Combines temporal convolution, transformers and multimodal data fusion.
- XGBoost [16], [17]: Scalable tree boosting system with sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning, empowered by specifically designed cache access patterns and data compression.

C. Results

The result of the proposed method's accuracy is shown in Figure.3. The dataset contains a total of 40,591 datapoints after the data pocessing step described in previous sections. Across the 12 classes of emotion in the CASE dataset, the proposed method reached an overall accuracy of 0.61 for a 5 seconds lapse, and 0.66 for a 10 seconds lapse. The improvement justifies the assumption that previous emotions states are effective in performing emotion recognition.

The proposed method demonstrates comparable performance as previous methods, as shown in Table I. Baselines include systems that classify on only valence or arousal (single criteria), or both valence and arousal (combined criteria). Baseline methods categorize emotions into low and medium, or low, medium, and high value categories. However, this work's proposed method extends the classification to delineate 12 distinct emotional classes, defined by both valence and arousal dimensions, offering a more detailed and comprehensive classification framework.

TABLE I
Model Comparison

Model	Classes	Criteria	Accuracy(%)
DCNN	3	Single	56
Attn-BiLSTM	3	Single	58
MulT	3	Single	63
Transformer-SS	3	Single	66
XGBoost	4	Combined	66
Ours	**12**	**Combined**	**66**

IV. Future Work

Moving forward, the trajectory of this research unfolds into several goals including the construction of a comprehensive dataset, exploring varied temporal windows, and conducting in-depth analyses of individual physiological signals.

A custom dataset would be created to align with the capabilities of the proposed method. While existing datasets often encompass stimuli categorized into combinations of 4 or 9 classes based on valence and arousal levels, the proposed methodology's capacity for detailed classification prompts the construction of a new dataset featuring 12 distinct emotion classes as demonstrated in Figure.1. The constructed dataset would be used to provide more detailed investigation into the system's ability of predicting each class of emotion.

Moreover, the time span of 5 seconds and 10 seconds preceding the current stage was considered for prediction in the present work. Future investigations will evaluate the impact of altering this time span, probing into shorter or longer duration to discern how varying temporal scopes influence prediction accuracy. This exploration seeks to discover the optimal temporal windows for predicting emotional states.

Additionally, while the current methodology incorporates the entirety of 4 physiological signals (ECG, BVP, GSR, Respiration), forthcoming research will investigate individual signals' performance on the prediction process, which would discern the distinctive contributions and predictive efficacy of each physiological signal. Understanding the relative strengths and weaknesses of individual signals will facilitate the refinement of the current deep learning model, potentially uncovering signal-specific nuances that can augment the accuracy.

V. Conclusion

This paper introduces a transformer-based deep learning pipeline for emotion recognition, which explores the intricate correlation between physiological signals and emotional changes. The custom methodology demonstrates a classification accuracy of 0.66 on 12 classes of emotions, comparable to previous works while expanding the classification spectrum. The outcomes underscore the potential of this methodology in capturing the nuances of emotional states. Looking ahead, future endeavors aim to construct a dataset tailored to the classification capabilities of the proposed model, thereby enabling a more detailed exploration of each physiological signal's contribution and a comprehensive investigation into varying prediction time lapses.

References

[1] Ayata, D., Yaslan, Y. & Kamasak, M.E. Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems. J. Med. Biol. Eng. 40, 149–157 (2020). https://doi.org/10.1007/s40846-019-00505-7

[2] M. Shamim Hossain, Ghulam Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, Information Fusion, Volume 49, 2019, Pages 69-78, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2018.09.008.

[3] G. Du, S. Long and H. Yuan. Non-Contact Emotion Recognition Combining Heart Rate and Facial Expression for Interactive Gaming Environments. IEEE Access, vol. 8, pp. 11896-11906, 2020. doi: 10.1109/ACCESS.2020.2964794.

[4] Y. Su et al., "EmotionO+: Physiological signals knowledge representation and emotion reasoning model for mental health monitoring," 2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Belfast, UK, 2014, pp. 529-535, doi: 10.1109/BIBM.2014.6999215.

[5] Jeremy N. Bailenson, Emmanuel D. Pontikakis, Iris B. Mauss, James J. Gross, Maria E. Jabon, Cendri A.C. Hutcherson, Clifford Nass, Oliver John. Real-time classification of evoked emotions using facial feature tracking and physiological responses. International Journal of Human-Computer Studies, Volume 66, Issue 5, 2008, Pages 303-317. ISSN 1071-5819. https://doi.org/10.1016/j.ijhcs.2007.10.011.

[6] Shruti Garg1, Soumyajit Behera2, K Rahul Patro2 and Ashwani Garg, Deep Neural Network for Electroencephalogram based Emotion Recognition, 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1187 012012 DOI 10.1088/1757-899X/1187/1/012012

[7] Narumoto, Jin, Yamada, Hiroki, Iidaka, Tetsuya, Sadato, Norihiro, Fukui, Kenji, Itoh, Harumi, Yonekura, Yoshiharu. Brain regions involved in verbal or non-verbal aspects of facial emotion recognition. NeuroReport 11(11):p 2571-2574, August 3, 2000.

[8] Enrique Leon, Graham Clarke, Victor Callaghan, Francisco Sepulveda, A user-independent real-time emotion recognition system for software agents in domestic environments, Engineering Applications of Artificial Intelligence, Volume 20, Issue 3, 2007, Pages 337-345, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2006.06.001.

[9] Joudeh IO, Cretu A-M, Bouchard S, Guimond S. Prediction of Continuous Emotional Measures through Physiological and Visual Data. Sensors. 2023; 23(12):5613. https://doi.org/10.3390/s23125613

[10] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

[11] Y. Wu, M. Daoudi and A. Amad, "Transformer-Based Self-Supervised Multimodal Representation Learning for Wearable Emotion Recognition," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2023.3263907.

[12] Y. Wu, M. Daoudi and A. Amad, "Transformer-Based Self-Supervised Multimodal Representation Learning for Wearable Emotion Recognition," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2023.3263907.

[13] Miller Jr, Harold L., ed. The Sage encyclopedia of theory in psychology. Sage Publications, 2016.

[14] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay and N. Arunkumar, "Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)," in IEEE Access, vol. 7, pp. 57-67, 2019, doi: 10.1109/ACCESS.2018.2883213.

[15] Chao Li, Zhongtian Bao, Linhao Li, and Ziping Zhao. 2020. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. Inf. Process. Manage. 57, 3 (May 2020). https://doi.org/10.1016/j.ipm.2019.102185

[16] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[17] M. S. Zitouni, C. Y. Park, U. Lee, L. J. Hadjileontiadis and A. Khandoker, "LSTM-Modeling of Emotion Recognition Using Peripheral Physiological Signals in Naturalistic Conversations," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 912-923, Feb. 2023, doi: 10.1109/JBHI.2022.3225330.

[18] E. Mansouri-Benssassi and J. Ye, "Speech Emotion Recognition With Early Visual Cross-modal Enhancement Using Spiking Neural Networks," 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2019, pp. 1-8, doi: 10.1109/IJCNN.2019.8852473.