Multiscale Transformer-Based for Multimodal Affective States Estimation from Physiological Signals

Ngoc Tu Vu, Van Thong Huynh, Hyung-Jeong Yang, and Soo-Hyung Kim(✉)

Department of AI Convergence, Chonnam National University,
Gwangju, South Korea
{tu369,vthuynh,hjyang,shkim}@jnu.ac.kr

Abstract. In recent times, the estimation of affective states from physiological data has garnered considerable attention within the research community owing to its wide-ranging applicability in daily life scenarios. The advancement of wearable technology has facilitated the collection of physiological signals, thereby highlighting the necessity for a resilient system capable of effectively discerning and interpreting user states. This work introduces an innovative methodology aimed at addressing the Valence-Arousal estimation, through the utilization of physiological signals. Our proposed model presents an efficient multi-scale transformer-based architecture for fusing signals from multiple modern sensors to tackle Emotion Recognition task. Our approach involves applying a multi-modal technique combined with scaling data to establish the relationship between internal body signals and human emotions. Additionally, we utilize Transformer and Gaussian transformation techniques to improve signal encoding effectiveness and overall performance. Our proposed model demonstrates compelling performance on the CASE dataset, achieving an impressive Root Mean Squared Error (RMSE) of 1.45.

Keywords: Affective states analysis · Physiological signals · Multimodal · Mental health

1 Introduction

Recognizing emotions is a fundamental aspect of human communication, and the ability to accurately detect emotional states has significant impacts on a range of applications, from healthcare to human-computer interaction. Emotions are often reflected in physiological signals [26], facial [8], and speech [24]. Recently, the use of physiological signals for affective computing has gained considerable attention due to its potential to provide objective measures of emotional states in real time [21].

Recently, there has been a growing interest in developing machine learning algorithms for affective computing using physiological signals [2,4,5,22,28].

H. Lu et al. (Eds.): ACPR 2023, LNCS 14408, pp. 113–122, 2023.
https://doi.org/10.1007/978-3-031-47665-5_10

These algorithms can be used to classify emotional states, predict changes in emotional states over time, or identify the specific features of physiological signals that are most informative for detecting emotional states. There has also been interested in developing wearable sensors that can capture physiological signals in real-world settings, such as in the workplace or in social situations [20].

The use of end-to-end deep learning architectures for physiological signals has the potential to simplify the development and deployment of an emotion recognition system [21]. By eliminating the need for preprocessing steps, these architectures can reduce the complexity and time required for system development, as well as improve the scalability and accuracy of the system. In addition, end-to-end architectures can enable the development of systems that can process multiple physiological signals simultaneously, such as heart rate, respiration, and electrodermal activity, providing more comprehensive and accurate measures of emotional states.

Despite the potential benefits of end-to-end deep learning architectures for affective computing, there are still challenges that need to be addressed. One challenge is to develop architectures that can handle noisy and non-stationary physiological signals, which can be affected by movement artifacts, signal drift, and other sources of noise. Another challenge is to ensure that the learned features are interpretable and meaningful, which can help improve the transparency and explainability of the system.

In this paper, we propose an end-to-end multi-scale architecture for continuous emotion regression with physiological signals. We evaluate the performance of the proposed architecture using CASE dataset [25], which contains data collected from experiments carried out in a laboratory setting.

2 Related Works

2.1 Continuous Emotion Recognition from Multimodal Physiological Signal

The utilization of physiological signals has been widely acknowledged as one of the most reliable data forms for affective science and affective computing. Although individuals are capable of manipulating their physical signals such as facial expressions or speech, consciously controlling their internal state is a daunting task. Therefore, analysis of signals from the human body represents a reliable and robust approach to fully recognizing and comprehending an individual's emotional state [1,26]. This reliability factor is especially crucial in medical applications, such as mental health treatment or mental illness diagnosis.

Recognizing affect from physiological data remains a significant challenge, not only during the data acquisition process but also in terms of emotion assessment. Laboratory-based research dominates the field of affective science due to the control it affords over experimental variables. Researchers can carefully select and prepare emotional stimuli, and employ various sensor devices to trace and record a subject's emotional state with minimal unexpected event, interference [21]. However, most of these studies rely on discrete indirect methods, such as quizzes, surveys, or discrete emotion categories for emotion assessment, which overlook the time-varying nature of human emotional experience. Sharma et al. [25] introduced Joystick-based emotion reporting interface (JERI) to overcome a limitation in emotion assessment. JERI enables the simultaneous annotation of valence and arousal, allowing for moment-to-moment emotion assessment. The Continuously Annotated Signals of Emotion (CASE) dataset, acquired using JERI, provides additional information for researchers to identify the timing of emotional triggers.

In addition, it is claimed that a single physiological signal is relatively difficult to precisely reflect human emotional changes. Therefore, recently, there has been much research focusing on detecting human emotion through multimodal physiological signals. There are many types of physiological signal used in these studies. While some studies record heart-related signals such as electrocardiographic (ECG) [7,17,18], blood volume pulse (BVP) [15,33], others use electrical activity of the brain (Electroencephalogram/EEG) [13,14] or muscle electrical reaction (Electromyogram/EMG) [19,23]. Furthermore, some even employ skin temperature (SKT) [19], skin sweat glands (EDA) [13,23], the depth and rate of breathing (respiratory/RSP) [23].

2.2 Transformer-Based Method for in Multimodal Emotion Recognition from Physiological Signal

Similar to other emotion recognition problems that involve physical signals, affective computing in physiological data has witnessed extensive adoption of machine learning techniques, particularly deep learning methodologies. Dominguez et al. [4] employed various conventional machine learning techniques, including Gaussian naive Bayes, k-Nearest Neighbors, and support vector machines, for valence-arousal estimation. However, these approaches are heavily dependent on the quality of handcrafted feature selection and feature extraction processes. To overcome this challenge, other studies [5,22,28] proposed the use of Deep Learning techniques for an end-to-end approach, where the model learns to extract features automatically without the need for pre-designed feature descriptors.

With the advancement of deep learning, various state-of-the-art techniques have been used to analyze physiological signals. Santamaria et al. [22] used convolutional neural networks (CNN) with 1D convolution layers for emotion detection, while Harper et al. [5] combined CNNs with frequently used recurring neural networks (RNN) for emotion recognition from ECG signals. Since their introduction in 2016, Transformers [27] have emerged as preferred models in the field of deep learning. Their robust performance in natural language processing, a type of data that shares some characteristics similar to time-series data, has demonstrated the potential of Transformers when applied to time-series signals. As a result, recent research in the time series domain has utilized Transformers as the core module in their model architecture [9,10,30]. For physiological signals, some studies have proposed using Transformers and their variants to detect emotions [28,29,31,32]. In the works of Vazquez et al. [28,29], they focused on applying pre-trained Transformers for multimodal signal processing. However, this is still a very basic application of Transformer modules. Wu et al. [31] and Yang et al. [32] proposed using more advanced techniques of Transformer-based models, which are self-supervised and Convolution-augmented transformers for single- and multimodal signal processing. Although these studies have demonstrated the effectiveness of transformers for physiological signals, they often feed the model with fixed original size signals, which may lead to the loss of global feature information. To address this issue, we propose a new multi-scale transformer-based architecture for multimodal emotional recognition.

3 Proposed Approach

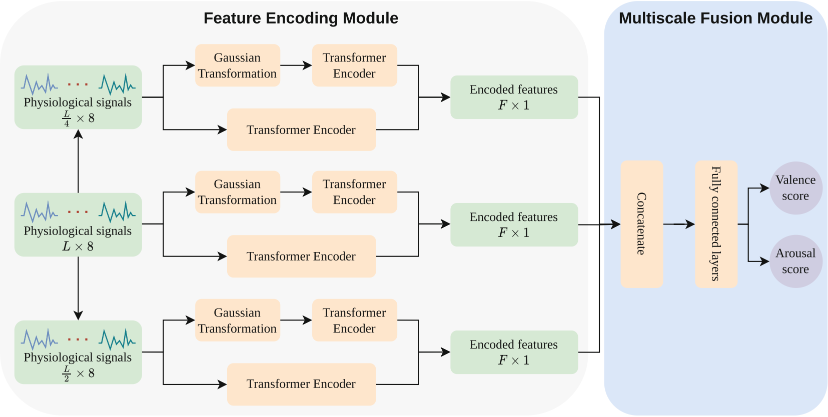

Fig. 1. An overview of our proposed architecture.

3.1 Problem Definition

The emotion recognition in multimodal physiological signal problem takes as input 8 physiological signals, namely ECG, BVP, EMG_CORU, EMG_TRAP, EMG_ZYGO, GRS, RSP and SKT, extracted from human subjects during emotion-inducing stimuli. This is denoted as the 8 sequence with L length. In the affective computing field, the objective of the emotion recognition problem varies depending on the indicated emotional models. In the scope of this study, following the use of the SAM (Self-Assessment Manikin) [3] model of the CASE dataset, the problem objective is the estimated Valence-Arousal (V-A) value. The V-A score consists of two continuous floating-point numbers ranging from 0.5 to 9.5. A value of 0.5 denotes the most negative valence or the lowest arousal, 5 indicates the neutral valence or arousal, and 9.5 indicates the most positive valence or the highest arousal.

3.2 Methodology

We constructed a new multiscale architecture for the estimation of valence arousal from 8 physiological signals. Our architecture consists of two core modules: Feature encoding module and multiscale fusion module. The process involves feeding raw physiological data into a feature encoding module, designed to extract vital information across varying global and local scales. Subsequently, the multi-scale features are fused and utilized for the estimation of Valence-Arousal scores. The overall architecture is shown in Fig. 1.

Feature Encoding. To enable the feature encoding module to extract global features for the estimator and eliminate noise and interference information from the input, we employ 1-Dimensional average pooling to scale the 8 input signals into three different lengths: L, $L/2$, and $L/4$. This process helps to improve the model's ability to extract useful information and eliminate unwanted noise and interference.

Then, we simultaneously apply two types of feature encoders, which are the Gaussian transform [16] and the transformer encoder [27]. The transformer encoder block is used as multi-headed self-attention as its core mechanism. Given an input sequential signal $S \in R^{L \times C}$, where L represents the length of the signal sequence and $C = 8$ is the number of channels (signal modalities), we apply a positional encoding and embedding layer to convert the raw input into a sequence of tokens. Subsequently, the tokens are fed into transformer layers consisting of multi-headed self-attention (MSA) [27], layer normalization (LN), and multilayer perceptron (MLP) blocks. Each element is formalized in the following equations:

$$y^i = MSA(LN(x^i)) + x^i \tag{1}$$

$$x^{l+i} = MLP(LN(y^i)) + y^i \tag{2}$$

Here, i represents the index of the token, and x^i denotes the generated feature's token. It is worth noting that since the multi-headed self-attention mechanism allows multiple sequences to be processed in parallel, all 8 signal channels are fed into the Transformer Encoder at once.

The Gaussian transform [16] is traditionally employed to kernelize linear models by nonlinearly transforming input features via a single layer and subsequently training a linear model on top of the transformed features. However, in the context of deep learning architectures, random features can also be leveraged, given their ability to perform dimensionality reduction or approximate certain functions via random projections. As a non-parametric technique, this transformation maps input data to a more compressed representation that excludes noise information while still enabling computationally efficient processing. Such a technique may serve as a valuable supplement to Transformer Encoder architectures, compensating for any missing information.

Multi-scale Fusion. From the features extracted from the feature encoder module on different scales, we fuse them using the concatenation operation. The concatenated features are then fed through a series of fully connected layers (FCN) for the estimation of the 2 valence and arousal scores. The Rectified Linear Unit (ReLU) activation function is chosen for its ability to introduce non-linearity into the model, thus contributing to the accuracy of the score estimation. The effectiveness of this approach lies in its ability to efficiently estimate the desired scores while maintaining a simple and straightforward architecture.

4 Experimental and Results

4.1 Dataset

CASE dataset [25] contains data from several physiological sensors and continuous annotations of emotion. These data were acquired from 30 subjects while they watched several video-stimuli and simultaneously reported their emotional experience using JERI. The devices used include sensors for electrocardiography (ECG), blood volume pulse (BVP), galvanic skin response (GSR), respiration (RSP), skin temperature (SKT) and electromyography (EMG). These sensors return 8 types of physiological signals: ECG, BVP, EMG_CORU, EMG_TRAP, EMG_ZYGO, GRS, RSP and SKT. Emotional stimuli consisted of 11 videos, ranging in duration from 120 to 197 s. The annotation and physiological data were collected at a sampling rate of 20 Hz and 1000 Hz, respectively. The initial range of valence arousal scores was established at [−26225, 26225].

We evaluate our approach with four different scenarios:

- Across-time scenario: Each sample represents a single person watching a single video, and the training and test sets are divided based on time. Specifically, the earlier parts of the video are used for training, while the later parts are reserved for testing.
- Across-subject scenario: Participants are randomly assigned to groups, and all samples from a given group belong to either the train or test set depending on the fold.
- Across-elicitor scenario: Each subject has two samples (videos) per quadrant in the arousal-valence space. For each fold, both samples related to a given quadrant are excluded, resulting in four folds, with one quadrant excluded in each fold.
- Across-version scenario: Each subject has two samples per quadrant in the arousal-valence space. In this scenario, one sample is used to train the model, and the other sample is used for testing, resulting in two folds.

4.2 Experiments Setup

Our networks were implemented using the TensorFlow framework. We trained our models using the AdamW optimizer [12] with a learning rate of 0.001 and the Cosine annealing warm restarts scheduler [11] over 10 epochs. The MSE loss function was used to optimize the network, and the evaluation stage is done with RMSE. The sequence length was set to 2048. We utilized 4 transformer layers for the transformer encoder, with each Attention module containing 4 heads. The hidden dimension of the transformer was set to 1024. All training and testing processes were conducted on a GTX 3090 GPU.

Table 1. RMSE on the test data with different scenarios.

Approach	Scenario	Arousal	Valence
Hinduja et al. [6]	Across-time	1.82	1.76
	Across-subject	1.33	1.31
	Across-elicitor	1.03	1.10
	Across-version	0.99	1.07
	Avg	1.292	1.31
Ours - without transformer on original signals	Across-time	1.550	1.612
	Across-subject	1.478	1.592
	Across-elicitor	1.600	1.653
	Across-version	1.595	1.548
	Avg	1.556	1.601
Ours	Across-time	1.503	1.639
	Across-subject	1.336	1.345
	Across-elicitor	1.509	1.514
	Across-version	1.369	1.352
	Avg	1.430	1.463

4.3 Results

Table 1 presents the results of our model in the test set in terms of the evaluation at different scenarios. Overall, the final RMSE score for the valence and arousal estimation task that we gain is 1.447. Our model showcases promising performance in comparison to the approach presented by Hinduja et al. [6]. It achieves a slightly lower score of 0.077 in Arousal and 0.153 in Valence score.

In detail, our model achieved the best performance in the across-subject scenario, with an arousal score of 1.336 and a valence score of 1.345. These results suggest that our model can effectively generalize to new subjects and accurately capture the emotion change after fully viewing the entire video-viewing process. Meanwhile, the relatively low performance in the across-elicitor scenario, with scores of 1.509 and 1.514 in arousal and valence, respectively, suggests that our model did not perform well in inferring emotional states that were not seen during training, given the specific emotional states learned previously. In the context of the across-time scenario, our results demonstrate a significantly improved performance compared to that of Hinduja et al. [6]. Specifically, our model achieves noteworthy enhancements in both Arousal and Valence scores, with a margin of 0.317 in Arousal and 0.121 in Valence. This substantial improvement opens up promising avenues for our future research.

5 Conclusion

This paper proposes a new multiscale architecture for multimodal emotional recognition from physiological signals. Our approach involves encoding the signal with the transformer encoder at multiple scales to capture both global and local features and obtain more informative representations. Our method achieved decent results on the CASE dataset.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219107). This work was also supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT).

References

1. Ahmad, Z., Khan, N.: A survey on physiological signal-based emotion recognition. Bioengineering **9**(11), 688 (2022)
2. Algarni, M., Saeed, F., Al-Hadhrami, T., Ghabban, F., Al-Sarem, M.: Deep learning-based approach for emotion recognition using electroencephalography (EEG) signals using bi-directional long short-term memory (Bi-LSTM). Sensors **22**(8), 2976 (2022)
3. Bradley, M.M., Lang, P.J.: Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry **25**(1), 49–59 (1994)
4. Domínguez-Jiménez, J.A., Campo-Landines, K.C., Martínez-Santos, J.C., Delahoz, E.J., Contreras-Ortiz, S.H.: A machine learning model for emotion recognition from physiological signals. Biomed. Sig. Process. Control **55**, 101646 (2020)
5. Harper, R., Southern, J.: A Bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. IEEE Trans. Affect. Comput. **13**(2), 985–991 (2020)
6. Hinduja, S., Bilalpur, M., Jivnani, L., Canavan, S.: Multimodal temporal modeling of emotion using physiological signals (2023)
7. Hu, L., Yang, J., Chen, M., Qian, Y., Rodrigues, J.J.: SCAI-SVSC: smart clothing for effective interaction with a sustainable vital sign collection. Fut. Gener. Comput. Syst. **86**, 329–338 (2018)
8. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affect. Comput. **13**(3), 1195–1215 (2020)
9. Li, S., et al.: Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In: Advances in Neural Information Processing Systems, vol. 32 (2019)

10. Li, Y., Peng, X., Zhang, J., Li, Z., Wen, M.: DCT-GAN: dilated convolutional transformer-based GAN for time series anomaly detection. IEEE Trans. Knowl. Data Eng. **35**, 3632–3644 (2021)
11. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
12. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
13. Martens, T., Niemann, M., Dick, U.: Sensor measures of affective leaning. Front. Psychol. **11**, 379 (2020)
14. Nakisa, B., Rastgoo, M.N., Rakotonirainy, A., Maire, F., Chandran, V.: Long short term memory hyperparameter optimization for a neural network based emotion recognition framework. IEEE Access **6**, 49325–49338 (2018)
15. Ragot, M., Martin, N., Em, S., Pallamin, N., Diverrez, J.-M.: Emotion recognition using physiological signals: laboratory vs. wearable sensors. In: Ahram, T., Falcão, C. (eds.) AHFE 2017. AISC, vol. 608, pp. 15–22. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-60639-2_2
16. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. In: Advances in Neural Information Processing Systems, vol. 20 (2007)
17. Rattanyu, K., Mizukawa, M.: Emotion recognition using biological signal in intelligent space. In: Jacko, J.A. (ed.) HCI 2011. LNCS, vol. 6763, pp. 586–592. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21616-9_66
18. Rattanyu, K., Ohkura, M., Mizukawa, M.: Emotion monitoring from physiological signals for service robots in the living space. In: ICCAS 2010, pp. 580–583. IEEE (2010)
19. Romeo, L., Cavallo, A., Pepa, L., Bianchi-Berthouze, N., Pontil, M.: Multiple instance learning for emotion recognition using physiological signals. IEEE Trans. Affect. Comput. **13**(1), 389–407 (2019)
20. Saganowski, S., Behnke, M., Komoszyńska, J., Kunc, D., Perz, B., Kazienko, P.: A system for collecting emotionally annotated physiological signals in daily life using wearables. In: 2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW), pp. 1–3. IEEE (2021)
21. Saganowski, S., Perz, B., Polak, A., Kazienko, P.: Emotion recognition for everyday life using physiological signals from wearables: a systematic literature review. IEEE Trans. Affect. Comput. **14**, 1876–1897 (2022)
22. Santamaria-Granados, L., Munoz-Organero, M., Ramirez-Gonzalez, G., Abdulhay, E., Arunkumar, N.: Using deep convolutional neural network for emotion detection on a physiological signals dataset (amigos). IEEE Access **7**, 57–67 (2018)
23. Schmidt, P., Dürichen, R., Reiss, A., Van Laerhoven, K., Plötz, T.: Multi-target affect detection in the wild: an exploratory study. In: Proceedings of the 2019 ACM International Symposium on Wearable Computers, pp. 211–219 (2019)
24. Schuller, B.W.: Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. Commun. ACM **61**(5), 90–99 (2018)
25. Sharma, K., Castellini, C., van den Broek, E.L., Albu-Schaeffer, A., Schwenker, F.: A dataset of continuous affect annotations and physiological signals for emotion analysis. Sci. Data **6**(1), 196 (2019)
26. Shu, L., et al.: A review of emotion recognition using physiological signals. Sensors **18**(7), 2074 (2018)
27. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)

28. Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.L.: Emotion recognition with pre-trained transformers using multimodal signals. In: 2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 1–8. IEEE (2022)
29. Vazquez-Rodriguez, J., Lefebvre, G., Cumin, J., Crowley, J.L.: Transformer-based self-supervised learning for emotion recognition. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 2605–2612. IEEE (2022)
30. Wu, N., Green, B., Ben, X., O'Banion, S.: Deep transformer models for time series forecasting: The influenza prevalence case. arXiv preprint arXiv:2001.08317 (2020)
31. Wu, Y., Daoudi, M., Amad, A.: Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Trans. Affect. Comput. (2023)
32. Yang, K., et al.: Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In: Proceedings of the 2022 International Conference on Multimedia Retrieval, pp. 562–570 (2022)
33. Zhao, B., Wang, Z., Yu, Z., Guo, B.: EmotionSense: emotion recognition based on wearable wristband. In: 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 346–355. IEEE (2018)