Feature Fusion Based on Mutual-Cross-Attention Mechanism for EEG Emotion Recognition

Yimin Zhao[0009−0007−9615−0386] and Jin Gu*(✉)[0000−0003−1147−6170]

School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu, China
gujin@swjtu.edu.cn

Abstract. An objective and accurate emotion diagnostic reference is vital to psychologists, especially when dealing with patients who are difficult to communicate with for pathological reasons. Nevertheless, current systems based on Electroencephalography (EEG) data utilized for sentiment discrimination have some problems, including excessive model complexity, mediocre accuracy, and limited interpretability. Consequently, we propose a novel and effective feature fusion mechanism named Mutual-Cross-Attention (MCA). Combining with a specially customized 3D Convolutional Neural Network (3D-CNN), this purely mathematical mechanism adeptly discovers the complementary relationship between time-domain and frequency-domain features in EEG data. Furthermore, the new designed Channel-PSD-DE 3D feature also contributes to the high performance. The proposed method eventually achieves **99.49% (valence)** and **99.30% (arousal)** accuracy on DEAP dataset. Our code and data is open-sourced at https://github.com/ztony0712/MCA.

Keywords: Emotion Recognition · Attention Feature Fusion · 3D-CNN · EEG Feature.

1 Introduction

Autism and depression are serious psychological problem, potentially leading to detrimental outcomes. A recent study indicated that dysarthria, mood disorders, rumination, literal understanding of problems or communication difficulties make their assessment difficult [3]. Therefore, it is essential for psychological therapist to examine more reliable indicators such as EEG data from patients. By promptly integrating the emotion judgements derived from these signals into the diagnostic process, psychologists are better equipped to formulate tailored treatment strategies for their patients.

In recent years, the academic community has achieved some advances in emotion recognition through various methods [2,10,16]. Initially, the focus was on singular traditional EEG features, such as Differential Entropy (DE) [17] and Power Spectral Density (PSD) [1]. Subsequently, approaches involving feature

* Corresponding author.

fusion and deep learning were adopted to enhance recognition accuracy. With the application of these new technologies, the performance of the classifier has been improved, but there are still some problems. Currently, the mainstream fusion methods implemented different learnable models to extract feature mappings, and then concatenated them directly [8,7,4,13]. Some projects [16,2] appended extra neural networks to further process the feature mappings. These strategies, which train the networks to autonomously concentrate on significant aspects of the signal, escalates the burden of model training and diminishes the efficiency of the model's outputs. Considering that sentiment classification systems require instantaneous output in practical applications, current increasingly complicated neural networks are not beneficial. In addition, this is not conducive to the interpretability of the task, potentially resulting in moral hazards.

Furthermore, the latest study [6] indicated an emerging trend of utilizing 3D data inputs for models. The review identified two predominant structures of Channel-Time-Frame [10] and Channel-Topology-Time [15]. However, the final results were unsatisfactory. As shown in Table 4, the accuracy of 2D-Topology-DE structure of Yang et al. [18] is only 90.24%, which could be the SOTA 3D input feature structure of the other projects that use 3D-CNN network models. Our analysis suggests that the limited spatial information provided by the channel topology map may contribute to this situation.

In that case, to achieve an instant well-performing emotion justification system based on EEG analysis, this project introduces a novel solution that has been experimentally validated as the new state-of-the-art (SOTA) method. It encompasses two primary contributions:

Mutual-Cross-Attention Mechanism. Inspired by the self-attention mechanism proposed by Vaswani et al. [14], we introduce a purely mathematical method named Mutual-Cross-Attention (MCA) for it applies Attention Mechanism from each directions of two features. In the field of EEG emotion analysis, we are the first to propose a pure mathematical fusion method, coupled with customized 3D-CNN, to accomplish the task of feature fusion.

New 3D feature presentation. By analyzing existing projects, it is found that spectral information might be more prominent than spacial information (presented by channel topology). Hence, we develop a unique Channel-Frequency-Time 3D feature structure. This innovative feature presents spectral and temporal information simultaneously.

2 Methods

To evaluate the proposed MCA, we designed a complete experimental pipeline with five steps: Data Acquisition, Pre-process, Feature Extraction, Feature Fusion, and Classification. In terms of the feature fusing procedure, the complementarity between multiple features and the ability of the fusion mechanism to

find important information are both crucial. Finding the optimal combination is challenging. It is widely recognized that DE and PSD complement each other [9]. Hence, these two features are selected for further feature fusion research. The accuracy results, based on the Circumplex Model of Affect that concentrates on arousal and valence, are compared with other SOTA methods to demonstrate the validity of our methodology

2.1 Data acquisition

The DEAP dataset from Queen Mary University was selected for our experimental setup. In the study, 32 individuals were monitored using electroencephalogram (EEG) and peripheral physiological signals as they viewed 40 one-minute music video clips. Thus, the chosen data is cross-individual and cross-session. Participants rated each video on a scale of 1 to 9 in terms of arousal, valence, likeability, dominance, and familiarity [5]. The data acquisition equipment has 32 channels and work with 512 Hz sampling frequency.

2.2 Pre-process

The dataset is pre-processed guided by the well-known steps of Steve Luck [9]. It is cleaned by filtering wave and excluding noise components. Firstly, Notch Filter is implemented to eliminate 50 Hz signal, commonly associated with interference from AC power sources. Additionally, considering measurement tool inaccuracies and environmental interferences, a 4-45 Hz band-pass filter is set. Following this, Independent Component Analysis (ICA) is applied to the filtered EEG to cancel noise elements like Electrooculogram, Electrocardiogram, Electromyography. The final step involved downsampling the original 512 Hz data to 128 Hz. These operations improve data quality, reduce data volume, and accelerate computation speed.

2.3 Feature extraction

The DE and PSD extractions are adopted across five distinct frequency bands to enhance feature representation and prevent information from influencing each other. The categories include θ (4-7 Hz), α (8-10 Hz), slow α (8-13 Hz), β (14-29 Hz), and γ (30-45 Hz).

DE extraction. There are several methods to calculate DE. If the signal fits the Gaussian distribution, which is performed as $N(\mu, \epsilon^2)$. The mathematical formulation is equal to the following one:

$$\begin{aligned} h(X) = &- \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\epsilon^2}} e^{-\frac{(x-\mu)^2}{2\epsilon^2}} \log\left(\frac{1}{\sqrt{2\pi\epsilon^2}} e^{-\frac{(x-\mu)^2}{2\epsilon^2}}\right) dx \\ =&\frac{1}{2} log\left(2\pi e\epsilon^2\right) \end{aligned} \tag{1}$$

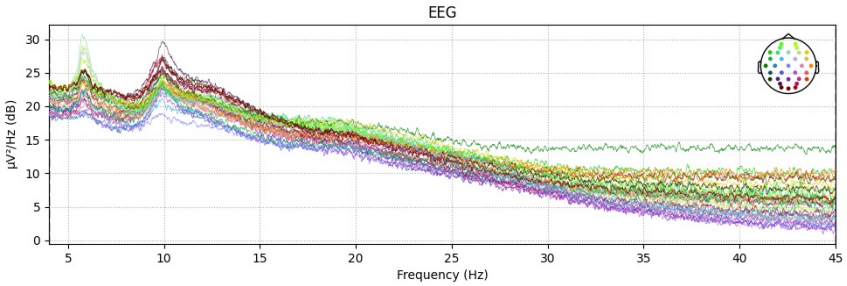

Fig. 1. PSD diagram of subject 01.

where ϵ is the standard deviation of $f(x)$. It has been proven that the EEG data filtered by a 4-45 or similar band-pass filter fits a Gaussian distribution every 2 Hz [11]. The formulation is applied every 2 seconds of data. Then these segments of DEs are collected and constructed as a DE trial array. Finally, the band-pass filter is applied to consider the DE feature separately according to that 5 different frequency bands mentioned in Sect. 2.3.

PSD extraction. The Welch's method is used to extract power spectra density (PSD). The first step to acquiring the PSD value is dividing the whole signal into K batches and calculating for each of them. The mathematical presentation of the k_{th} PSD value on frequency f is [12]:

$$p_k(f) = \frac{1}{W}|F_k(f)|^2 \tag{2}$$

where W is related to the Hanning window and $F_k(f)$ is a windowed fast Fourier transform (FFT) at a specific frequency f, which is set as 128 Hz according to the analysis above. The window size is 2 seconds. Finally, the estimation of PSD with the Welch method is combined with the results from all segments:

$$P_s(f) = \frac{1}{K}\sum_{k=1}^{K} p_k(f) \tag{3}$$

To preliminarily evaluate the validity of the PSD, a corresponding diagram is plotted. Fig. 1 indicates that the local value of the PSD fluctuates in the frequency range of about 5-7 Hz and 9-11 Hz, which suggests there might be emotion presentation in these ranges. That is the reason for the Sect. 2.3 indicating two frequency bands (slow α and α) in 8-13 Hz range. Finally, the PSD's 4-45 Hz spectrum is categorically divided into five bands for further analysis.

2.4 Feature fusion

The MCA mechanism is applied across each selected frequency band to fuse DE and PSD. Initially, respectively consider DE and PSD as key and query

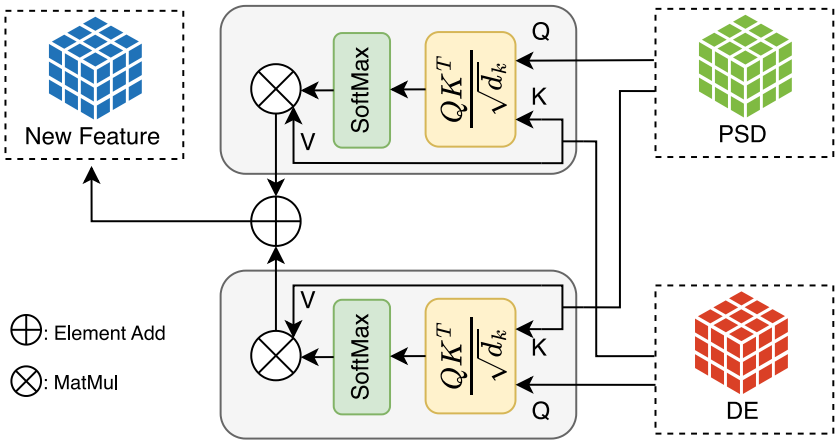

Fig. 2. Overview of mutual-cross-attention mechanism.

vector. Then, designate PSD as value and implement basic Scaled Dot-Product Attention, which is presented by:

$$\text{Atten}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

where Q, K, V respectively represent query, key, and value. And d_k is the size of the query's last dimension.

That is one direction calculation in MCA. After that, PSD is used as Q, DE as K and V. Again, the Scaled Dot-Product Attention operation is implemented. The results from both directions are added together to get the new feature. Fig. 2 illustrates the entire process, and its mathematical presentation is:

$$\text{MCA}(f_1, f_2) = \text{Atten}(f_1, f_2, f_2) + \text{Atten}(f_2, f_1, f_1) \tag{5}$$

where f_1 is the first feature (DE) and f_2 is the second feature (PSD).

2.5 Classification

After those feature extraction operations are implemented, the final single feature is denoted as $F_f \in \mathbb{R}^{32\times5\times60}$. However, it takes too long to perform classification tasks. Therefore, every F_f is split into 20 $F_s \in \mathbb{R}^{32\times5\times3}$. This operation allows the model to output a sentiment prediction every 3 seconds.

A special 3D-CNN structure is proposed to process the feature F_s . As shown in Fig. 3, the network begins with a 3D convolutional layer, configured with one input channel and 32 output channels, utilizing a 3x3x3 kernel. This layer is followed by another 3D convolutional layer, which maintains the same number of output channels and kernel size. Using two consecutive convolutional layers with the same number of channels enhances the network. The trick deepens the

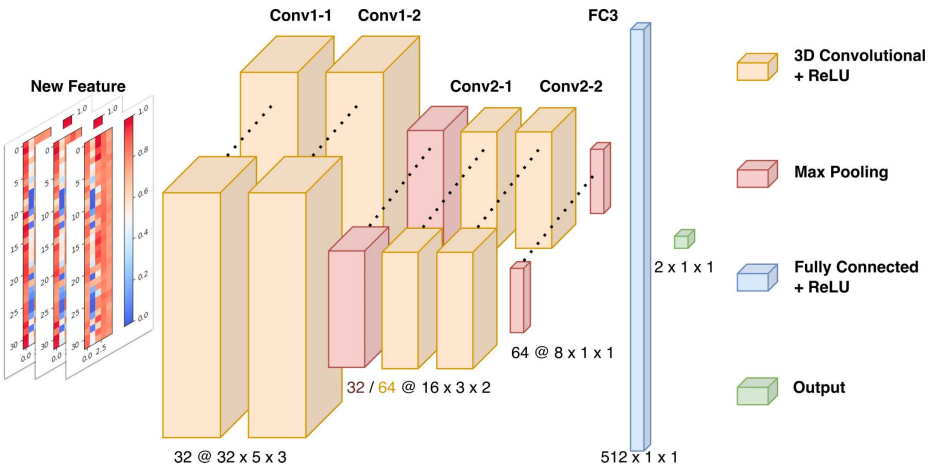

Fig. 3. Structure of new designed 3D-CNN.

network's capacity to extract features without immediately reducing the spatial dimensions of the input data. After the initial convolutional stage, a 2x2x2 kernel Max Pooling layer is employed with a (0, 1, 1) asymmetric padding.

Subsequently, the network extends into another set of convolutional layers, where the number of output channels is doubled to 64. Following this, the second Max Pooling layer further downsamples the feature maps. Finally, the network transitions to a fully connected layer, which classifies the extracted features into two categories.

2.6 Experiments setup

During the training phase, the hyperparameters are tuned to optimize model performance. Firstly, a test set randomly containing 10 percent of total dataset is split out. Then, the optimal hyperparameters are listed: Adam optimizer, 32 batch size, 12 epochs, 0.0001 weight decay, 0.0002 learning rate, and scheduler that reduces the learning rate by 0.2 every 10 epochs. For hardware, we used two NVIDIA GeForce RTX 3090 GPUs. The GPU driver version is 525.147.05 and the CUDA version is 12.0. The CPU is a AMD Ryzen Threadripper 3960X 24-Core Processor. However, a single RTX 3060 GPU is also enough.

3 Results and discussions

The performance of the proposed model is demonstrated through various evaluation metrics as detailed in Table 1. It is clear that all indices exceed 99%. Additionally, this section includes not only ablation experiments but also comparisons with other SOTA results. All these experiments and comparisons are conducted using the DEAP dataset. Ultimately, the proposed methodology is proven to be effective.

Table 1. Valence and arousal evaluation metrics of MCA-3D-CNN.

Category	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Valence	99.49	99.60	99.54	99.57
Arousal	99.30	99.45	99.31	99.38

3.1 Ablation experiments

The experiment primarily examines the impact of both singular and fused features on the results. For comparison with our proposed MCA method, a baseline configuration labeled "DE+PSD" is established, which is based on the summation of 3D-DE and 3D-PSD. According to the Table 2, the accuracy results for single DE and single PSD are almost the same to those of "DE+PSD". However, the valence at 99.49% and arousal at 99.30% achieved by the proposed method are significantly higher than those of "DE+PSD". This proves that the proposed method has advantages in the complementary integration of DE and PSD information.

Table 2. Ablation experiments results. "DE+PSD" represents the element-wise summation between 3D-DE and 3D-PSD.

Feature	Valence(%)	Arousal(%)
Channel-Frequency-DE	89.88	88.16
Channel-PSD-Time	91.88	91.56
DE+PSD(baseline)	90.90	91.30
Proposed MCA	**99.49**	**99.30**

3.2 Compare with other SOTA

The innovation of this project focuses mainly on the design of the new feature structure and the way of fusing the new features. Hence, the comparisons with other SOTA methods in these two aspects are carried out.

Table 3. Compare with results based on other fusion methods.

Features	Fusion Method(s)	Valence(%)	Arousal(%)
ResNet, LFCC	Concat & KNN [8]	90.39	89.06
DE, PSD, Hjorth, SE	CNN & SVM [2]	75.22	80.52
PSD, temporal statistics	STFFNN [16]	85.40	86.20
Time, 2D-Topology-Time	TSFFN [13]	98.27	98.53
DE, PSD	MCA & 3D-CNN (Ours)	**99.49**	**99.30**

Feature fusion comparison. There are various other fusion methods between different features as detailed in Table 3. Gao et al. [2] integrated DE, PSD, Hjorth, and Sample Entropy (SE). Our project, however, utilizes a narrower range of features. Additionally, similar to the work [16], our project fuses both frequency domain and time domain features. This demonstrates that our proposed method is effective when working with similar features, whether they are the same or fewer in categories.

In the study by Liu et al. [8], the accuracy results are around 90%, showing commendable performance. However, their approach relies on pre-trained features, which might limit its ability to instantly output results compared to our proposed method. Sun et al. [13] developed TSFFN to fuse EEG features with high accuracy. Comprising a 3D-CNN and a transformer, the TSFFN might be too complex for efficient computation. This highlights the advantages of our proposed purely mathematical MCA to feature fusion.

Table 4. Compare with results based on other singular 3D feature presentations.

Feature	Network	Valence(%)	Arousal(%)
2D-Topology-DE [18]	3D-CNN	89.78	90.24
2D-Topology-Time [15]	3D-CNN	72.10	73.10
Channel-Time-Frame [10]	3D-CNN	87.44	88.49
Channel-Frequency-DE (Ours)	3D-CNN	89.88	88.16
Channel-PSD-Time (Ours)	3D-CNN	**91.88**	**91.56**

Feature structure comparison. For 3D feature presentations, the majority of emotion recognition projects based on EEG analysis have inclined towards using topology to expand a 1D channel into a 2D format. Subsequently, data on other dimensions are combined with the topological channel map. As indicated in Table 4, the performance of our proposed feature structure outperforms all 3D-CNN methods that employ features of the topology and the Channel-Time-Frame. This validates the rationality and effectiveness of the structure we have designed.

4 Conclusion

The proposed MCA mechanism, 3D feature Channel-PSD-DE, and customized 3D-CNN show excellent capabilities in EEG-based emotion recognition. The whole system effectively overcomes the limitations of existing systems in terms of instantaneity, accuracy, and interpretability. By integrating DE and PSD features through the MCA mechanism, the ability of mathematical fusion methods to extract meaningful information from EEG data is highlighted. The emotional discrimination system developed using this solution has great potential for practical clinical psychotherapy. In the future, we will further explore the transformer

implementing MCA as a core, which could fuse features of larger and more complex datasets.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 62306249), the China Postdoctoral Science Foundation (No. 2023M732920), and the opening project of Neuracle Technology (Changzhou) Co., Ltd. (No. BRKOT-SWJTU-20221228G).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Duan, R.N., Wang, X.W., Lu, B.L.: Eeg-based emotion recognition in listening music by using support vector machine and linear dynamic system. In: Huang, T., Zeng, Z., Li, C., Leung, C.S. (eds.) Neural Information Processing. pp. 468–475. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2012). https://doi.org/10.1007/978-3-642-34478-7_57
2. Gao, Q., Yang, Y., Kang, Q., Tian, Z., Song, Y.: Eeg-based emotion recognition with feature fusion networks. International Journal of Machine Learning and Cybernetics **13**(2), 421–429 (Feb 2022). https://doi.org/10.1007/s13042-021-01414-5
3. Hervás, A.: Autism and depression: clinical presentation, evaluation and treatment. Medicina **83 Suppl 2**, 37–42 (Mar 2023)
4. Jia, Z., Lin, Y., Cai, X., Chen, H., Gou, H., Wang, J.: Sst-emotionnet: Spatial-spectral-temporal based attention 3d dense network for eeg emotion recognition. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2909–2917. ACM, Seattle WA USA (Oct 2020). https://doi.org/10.1145/3394171.3413724
5. Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis ;using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (Jan 2012). https://doi.org/10.1109/T-AFFC.2011.15
6. Li, X., Zhang, Y., Tiwari, P., Song, D., Hu, B., Yang, M., Zhao, Z., Kumar, N., Marttinen, P.: Eeg based emotion recognition: A tutorial and review. ACM Computing Surveys **55**(4), 1–57 (Apr 2023). https://doi.org/10.1145/3524499
7. Liu, J., Zhao, Y., Wu, H., Jiang, D.: Positional-spectral-temporal attention in 3d convolutional neural networks for eeg emotion recognition (Nov 2021). https://doi.org/10.48550/arXiv.2110.09955
8. Liu, N., Fang, Y., Li, L., Hou, L., Yang, F., Guo, Y.: Multiple feature fusion for automatic emotion recognition using eeg signals. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 896–900 (Apr 2018). https://doi.org/10.1109/ICASSP.2018.8462518
9. Luck, S.J.: An Introduction to the Event-Related Potential Technique, Second Edition. MIT Press (Jun 2014)
10. Salama, E.S., A.El-Khoribi, R., E.Shoman, M., A.Wahby, M.: Eeg-based emotion recognition using 3d convolutional neural networks. International Journal of Advanced Computer Science and Applications **9**(8) (2018). https://doi.org/10.14569/IJACSA.2018.090843

11. Shi, L.C., Jiao, Y.Y., Lu, B.L.: Differential entropy feature for eeg-based vigilance estimation. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 6627–6630 (Jul 2013). https://doi.org/10.1109/EMBC.2013.6611075
12. Solomon, O.: Psd computations using welch's method. [power spectral density (psd)]. Tech. Rep. SAND-91-1533, Sandia National Lab. (SNL-NM), Albuquerque, NM (United States) (Dec 1991). https://doi.org/10.2172/5688766
13. Sun, J., Wang, X., Zhao, K., Hao, S., Wang, T.: Multi-channel eeg emotion recognition based on parallel transformer and 3d-convolutional neural network. Mathematics **10**(17), 3131 (Jan 2022). https://doi.org/10.3390/math10173131
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
15. Wang, Y., Huang, Z., McCane, B., Neo, P.: Emotionet: A 3-d convolutional neural network for eeg-based emotion recognition. In: 2018 International Joint Conference on Neural Networks (IJCNN). pp. 1–7 (Jul 2018). https://doi.org/10.1109/IJCNN.2018.8489715
16. Wang, Z., Wang, Y., Zhang, J., Hu, C., Yin, Z., Song, Y.: Spatial-temporal feature fusion neural network for eeg-based emotion recognition. IEEE Transactions on Instrumentation and Measurement **71**, 1–12 (2022). https://doi.org/10.1109/TIM.2022.3165280
17. Yang, J., Huang, X., Wu, H., Yang, X.: Eeg-based emotion classification based on bidirectional long short-term memory network. Procedia Computer Science **174**, 491–504 (Jan 2020). https://doi.org/10.1016/j.procs.2020.06.117
18. Yang, Y., Wu, Q., Fu, Y., Chen, X.: Continuous convolutional neural network with 3d input for eeg-based emotion recognition. In: Cheng, L., Leung, A.C.S., Ozawa, S. (eds.) Neural Information Processing. pp. 433–443. Lecture Notes in Computer Science, Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-04239-4_39