Cognitive Neurodynamics (2023) 17:671–680
https://doi.org/10.1007/s11571-022-09851-w

RESEARCH ARTICLE

A multi-stage dynamical fusion network for multimodal emotion recognition

Sihan Chen[2] · Jiajia Tang[1] · Li Zhu[1] · Wanzeng Kong[1]

Received: 18 October 2021 / Revised: 21 June 2022 / Accepted: 6 July 2022 / Published online: 31 July 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract

In recent years, emotion recognition using physiological signals has become a popular research topic. Physiological signal can reflect the real emotional state for individual which is widely applied to emotion recognition. Multimodal signals provide more discriminative information compared with single modal which arose the interest of related researchers. However, current studies on multimodal emotion recognition normally adopt one-stage fusion method which results in the overlook of cross-modal interaction. To solve this problem, we proposed a multi-stage multimodal dynamical fusion network (MSMDFN). Through the MSMDFN, the joint representation based on cross-modal correlation is obtained. Initially, the latent and essential interactions among various features extracted independently from multiple modalities are explored based on specific manner. Subsequently, the multi-stage fusion network is designed to split the fusion procedure into multi-stages using the correlation observed before. This allows us to exploit much more fine-grained unimodal, bimodal and trimodal intercorrelations. For evaluation, the MSMDFN was verified on multimodal benchmark DEAP. The experiments indicate that our method outperforms the related one-stage multi-modal emotion recognition works.

Keywords Physiological signals · Emotion recognition · Multimodal dynamic fusion · Multi-stage fusion

Introduction

Emotions play an important role in our daily life and can influence human decisions. The accuracy and efficiency of human-computer interaction can be improved on emotion recognition. Emotion recognition using physiological signals is an upcoming research area in brain-computer interface (BCI). Physiological signals can express subjects' emotional states more objectively than other data such as facial expressions and texts (AlZoubi et al. 2012; Chen et al. 2015; Shu et al. 2018). Due to the reliability and objectivity, emotion recognition based on physiological signals was to be performed in a wider way. Gender-specific affective responses have been shown base on neurophysiological signalsGoshvarpour and Goshvarpour (2019). Electroencephalography brain connectivity patterns of different emotions can also be used in disorder researchMehdizadehfar et al. (2020). Signals were used to investigate the role of basal ganglia network in generating bipolar oscillationsBalasubramani and Chakravarthy (2020).

Researchers in the field of BCI have been studying EEG for many years and proposed five specific features of

This work was supported by National Key R &D Program of China for Intergovernmental International Science and Technology Innovation Cooperation Project (2017YFE0116800), National Natural Science Foundation of China (U20B2074, U1909202), Science and Technology Program of Zhejiang Province (2021C03003), and supported by Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province (2020E10010)

✉ Wanzeng Kong
kongwanzeng@hdu.edu.cn

Sihan Chen
csh_up@163.com

Jiajia Tang
hdutangjiajia@163.com

Li Zhu
zhulibrain@gmail.com

[1] The College of Computer Science, Hangzhou Dianzi University, Hangzhou, China

[2] HDU-ITMO Joint Institute, Hangzhou Dianzi University, Hangzhou, China

analyzing subjects' emotional states. They are power spectral density (PSD), differential entropy (DE), differential asymmetry (DASM), rational asymmetry (RASM) and differential causality (DCAU) (Thammasan et al. 2016; Shi et al. 2013; Davidson and Fox 1982; Zheng et al. 2017; Zheng and Lu 2015). Inspired from the results in image processing and natural language processing, recent researches have also applied various deep learning models to recognize emotion state via physiological signals (Hinton et al. 2012). Due to the advantage of recurrent neural network (RNN) that can extract specific information from sequences, emotion-related feature in temporal domain is extracted from EEG by RNN (Kim and Jo 2018). Considering both complex dependencies between adjacent signals and sequantial information within chain-like data, CNN and RNN have also be combined to model a cascade and parallel network for learning the combined spatial-temporal information of raw EEG streams (Zhang et al. 2018). Due to the different information in multichannel EEG signals, GNN (graph neural network) is used to perform EEG emotion recognition. Dynamical graph convolutional neural networks (DGCNN) is proposed to dynamically learn the intrinsic relationship between different EEG channels, which is benefit for more discriminative EEG feature extraction (Song et al. 2018).

Emotion is a complex individual performance that is comprised of a lot of internal physiological activities. In other words, emotion-changing leads to different effects on each modalities. Raw signals among different modalities always describe different aspect emotions and multimodal data contain much more complementary information. Multimodal data is more effective for model robust emotion recognition structure. Physiological signals include electroencephalography (EEG), electrooculography (EOG), electromyography (EMG), temperature (T), galvanic skin response (GSR), respiration (RSP), etc. To improve the stability and reliability of emotion recognition, multimodal physiological signal has been used to extract rich representations. High-level coordinated representation based on EEG and eye movement data is extracted by maximizing the canonical correlation via jointly learned parameters of two view' non linear transformations (Qiu et al. 2018). Multimodal residual LSTM network (MMResLSTM) is proposed to investigate the dependency among multiple modalities and high-level temporal-feature, which contains both the spatial shortcut paths provided by the residual network and temporal shortcut paths provided by LSTM for efficiently learning emotion-related high-level features (Ma et al. 2019). Huang proposed an Ensemble Convolutional Neural Network (ECNN) model, which is used to automatically mine the correlation between multi-channel EEG signals and peripheral physiological signals in order to improve the emotion recognition accuracy (Huang et al. 2019). However, the current fusion frameworks based on multimodal signals are mostly one-stage fusion which are unable to explicitly model the interaction between model-pair. In contrast, we decomposed multimodal fusion as a multi-stage procedure by which cross-modal interactions are explored and intra-modal messages are still reserved.

This paper absorbs a strong inspiration from ARGF to decomposed the one-stage fusion problem into multiple stages (Mai et al. 2020). However, ARGF model unimodel, bimodal and trimodal interactions independently before fusion procedure that makes fusion architecture too complicated. According to the characteristic of physiological signal, we simplified ARGF to decrease the burden on multimodal fusion and avoid model over-fitting. MSMDFN disintegrate fusion problem into multiple stages, each of them focused on the specialized fusion of selected two modalities. Multimodal interaction are obtained build upon the representation learned from previous stage. In other words, MSMDFN decompose the multi-stage fusion as a recurrent system.

In this article, we proposed a multi-stage multimodal dynamic fusion network (MSMDFN) to sequentially model the joint representation based on cross-modal correlation. First of all, the high-level local feature within each modalities are extracted separately. The multimodal dynamic fusion procedure is split into multiple stages according to the cross-modality dynamic correlation coefficient. The MSMDFN decomposed the one-stage fusion into multiple recursive stages, which allows each stage to concentrate on more specialized and fine-grained inter-correlations. The ablation study which attend to the effect of various modalities demonstrates that raw signals include multiple modalities are more effective for emotion recognition task. Moreover, the comparative experiment illustrates the effectiveness of the proposed multi-stage fusion mechanism. We evaluate MSMDFN on multimodal benchmark DEAP and achieves state-of-the-art performance.

Related work

In recent years, emotion recognition using physiological signal has become a popular research topic which objectively reflect the real emotional state of individuals. Most of the current studies focus on EEG and have achieved relatively impressive results. Vernon proposed EEGNet (Lawhern et al. 2018), a compact convolutional neural network model that consists of deepwise convolution and separable convolution (Chollet 2017). The experimental results confirm that EEGNet can achieve excellent experimental results on EEG datasets of different paradigms.

However, EEG-based emotion recognition overlooks the complementarity and consistency among modalities, leading to the great drop of performance. Multimodal emotion analysis combines multiple physiological signal, which improves the stability and reliability compared to single EEG signal. Chen selected significant multimodal features respectively by two comparative feature selection methods (Fisher Criterion Score and Davies-Bouldin index) and used hidden Markov models (HMMs) to performed emotion recognition (Chen et al. 2015). Liu extracted high level representation features by the Bimodal Deep AutoEncoder, that contain complementary information of EEG and eye features (Liu et al. 2016). Tang introduced Bimodal-LSTM model to take temporal information into account for emotion recognition with multimodal signals (Tang et al. 2017). Most of the physiological signals-based methods use well-designed classifiers with hand-crafted features to recognize human emotions. Inspired by the breakthroughs in the image domain using deep convolutional neural network, Lin present an approach to perform emotional states classification by end-to-end learning of deep convolutional neural network (CNN) (Lin et al. 2017). Shabnam investigated time-varying functional connectivity derived from the Jackknife Correlation method to recognize emotionsGhahari et al. (2020). Qiu adopt Deep Canonical Correlation Analysis (DCCA) for high-level coordinated representation to make feature extraction from EEG and eye movement data (Qiu et al. 2018). Ma proposed a multimodal residual LSTM (MMResLSTM) network to learn the correlation between the EEG and other physiological signals, which contains both the spatial shortcut paths provided by the residual network and temporal shortcut paths provided by LSTM for efficiently learning emotion-related high-level features (Ma et al. 2019). Yilmaz propose a multimodal fusion method between electroencephalography (EEG) and electrooculography (EOG) signals for emotion recognition (Yilmaz and Kose 2021). Before the feature extraction stage, different angle-amplitude transformations are applied to EEG-EOG signals that take arbitrary time domain signals and convert them two-dimensional images named as Angle-Amplitude Graph (AAG). Then, image-based features are obtained by using a scale invariant feature transform method. Liao used convolutional neural network to learn the spatial representations of multi-channel EEG signals and the Long Short-term Memory network to learn the temporal representations of peripheral physiological signals (Liao et al. 2020). Then two representations are combined for emotion recognition and classification. However, above methods all adopted one-stage fusion strategy that ignore the detailed interaction of every modality pair, resulting in un-detailed fusion in subset modalities and inadequate cross-modal knowledge.

Methodology

In this part, multi-stage multimodal dynamic fusion network (MSMDFN) is described to perform fusion to modal cross-modal feature that is used for emotion recognition. Give three modalities EEG(e), EOG(o), EMG(m), the signal from each modality is represented as $x_e \in \mathbb{R}^{C_e \times Fs}$, $x_o \in \mathbb{R}^{C_o \times Fs}$ and $x_m \in \mathbb{R}^{C_m \times Fs}$, where C_e, C_o and C_m refers to the number of channels of three modalities, Fs is represented as the number of data points. As shown in Fig. 1, the modal is mainly composed of two blocks, extracting block and fusion block. In extracting block, EEG extractor is mainly composed of convolutional neural network (CNN). Additionally, EOG extractor is composed of wavelet transform and CNN as well as EMG extractor. The feature of three modalities are pre-extracted separately, which denotes as f_e, f_o and f_m. Then, three features are projected into the space where features have same length and Pearson correlation coefficients are derived for determining the dynamic order of multi-stage fusion. After multi-stage fusion, an emotion feature which contained trimodal information is obtained.

Extractor

For EEG extractor, a compact CNN framework is used to modal EEG feature contain temporal and spatial representation, which is consisting of temporal, depthwise and separable convolution. First of all, the input samples x_e are passed into a 2D convolutional layer containing N temporal kernels of size $\left[1, \frac{Fs}{2}\right]$. The convolution operation is executed along the temporal domain and the stride is set to 1. N temporal feature maps in the size of $[C_e, Fs]$ are obtained, which contain specific time-representation on each channel. Then, feature maps are fed into the depthwise convolution to learn spatial feature of each temporal map independently. The depthwise convolutional layer contains $N \times D$ spatial kernels of size $[C_e, 1]$. D denotes the number of spatial kernels specialized for different temporal map. Additionally, the generated $N \times D$ feature contained temporal and spatial feature are down sampled to size $\left[1, \frac{Fs}{4}\right]$ via average pooling layer. Separable convolution consists of depthwise layer and point-wise convolution, which can be used to obtain the interaction among temporal feature maps. Feature maps obtained from depthwise convolution are subsequently fed into depthwise layer with kernels of size $\left[1, \frac{Fs}{8}\right]$ and $N \times D$ point-wise kernels. Batch normalization, activation ELU and dropout layer are employed after each convolution block to avoid model over-fitting. Finally, $f_e \in \mathbb{R}^{1 \times \frac{N \times D \times Fs}{32}}$ are obtained after down-sampled operation and flatten layer. Separable convolution not only significantly reduces the number of parameters compared

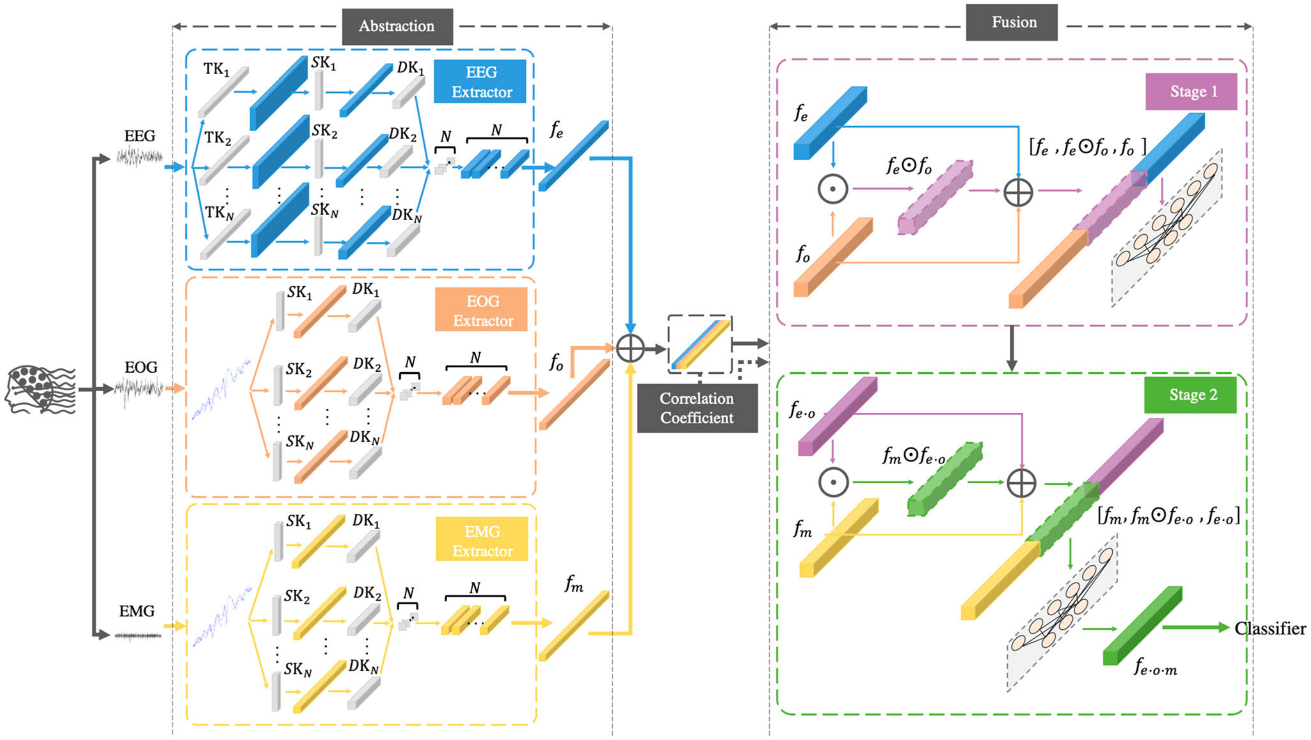


Fig. 1 An implementation for multi-stage multimodal dynamic fusion network (MSMDFN). The left part is extracting block and right one is fusion block. TK, SK, DK are represented as temporal kernels, spatial kernels and depthwise kernels. For illustration purpose, D was set as 1. Additionally, coefficient between f_e and f_o is maximum so the first fusion stage selects and creates an intermediate representation $f_{e \cdot o}$. The second fusion stage select f_m to be fused with intermediate vector $f_{e \cdot o}$ obtained in the previous stage

to normal convolution, but also explicitly decouples the relationships within and between the feature maps.

Before obtaining abstract EOG feature, raw physiological signals x_o are pre-processed via wavelet transform to get high-resolution information in time-frequency domain with size $\mathbb{R}^{S \times C_o \times Fs}$, where S refers to scale. Subsequently, new-obtained signals are fed into CNN framework to extract feature contained both temporal and inter-channel information. There are mainly two modules. Firstly, convolution layer is performed with M 2D-kernals of size $[C_o, 1]$to observe spatial information followed by downsample operation. After that, M feature maps in the size of $\left[1, \frac{Fs}{4}\right]$ are obtained. Next, the feature maps are fed into the second one, which is composed of 2D-kernels of size $\left[1, \frac{Fs}{8}\right]$ and M point-wise kernels. Above two modules are both followed by batch normalization, ELU activation and dropout layer for avoiding overfit. Finally, EOG feature is obtained after an average pool layer and a flatten layer, $f_o \in \mathbb{R}^{1 \times \frac{M \times Fs}{32}}$. EMG extractor is similar to EOG one.

Multi-stage fusion

Three features are projected into the space applying a linear transformation to make features have same length that is for later multi-stage fusion:

$$\bar{f}_e = tanh(W_e f_e + b_e) \tag{1}$$

$$\bar{f}_o = tanh(W_o f_o + b_o) \tag{2}$$

$$\bar{f}_m = tanh(W_m f_m + b_m) \tag{3}$$

$W_e \in \mathbb{R}^{L \times \frac{N \times D \times Fs}{32}}$, $W_o \in \mathbb{R}^{L \times \frac{M \times Fs}{32}}$ and $W_m \in \mathbb{R}^{L \times \frac{M \times Fs}{32}}$ are massive weight matrixes, L is the new vector length, and b_e, b_o and b_m are the bias vectors. The length of linear transformation output $\bar{f}_e$, $\bar{f}_o$ and $\bar{f}_m$ are L.

Subsequently, Pearson correlation are derived for determining the dynamic order of multi-stage fusion. MSMDFN focuses on the relative values of the correlation coefficient between the modalities. The Pearson correlation coefficient formula is shown below, where $E(\cdot)$ denotes expectation:

$$\rho_{XY} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{4}$$

For each sample, fusion order is dynamically determined by the value of cross-modal correlation. After calculating the correlation coefficients between every two modalities, the order of fusion is determined based on the value of the correlation coefficients and then multi-stage fusion is performed.

Firstly, the module selects a subset of multimodal feature from $\left[\overline{f}_e, \overline{f}_o, \overline{f}_m\right]$ base on cross-modality correlation that will be used for current stage of fusion. As shown in Fig. 2, we assume the correlation between $\overline{f}_e$ and $\overline{f}_o$ is relatively high, so the first fusion stage selects them. Afterwards, $\overline{f}_{e\cdot o}$ is obtained that is observed fine-grained information within each feature and exploited inter-modal interaction. $\odot$ is element-wise production, $W \in R^{(L\times(L\times3))}$. Subsequently, $\overline{f}_{e\cdot o}$ and $\overline{f}_m$ are fed into the second fusion stage:

$$\overline{f}_{e\cdot o} = tanh\left(W_1\left[\overline{f}_e, \overline{f}_e \odot \overline{f}_o, \overline{f}_o\right] + b_1\right) \tag{5}$$

$$\overline{f}_{e\cdot o\cdot m} = tanh\left(W_2\left[\overline{f}_m, \overline{f}_m \odot \overline{f}_{e\cdot o}, \overline{f}_{e\cdot o}\right] + b_2\right) \tag{6}$$

As a final step, the multimodal representation obtained from the second stage is fed into a fully connected layer for classification. $W_{class} \in \mathbb{R}^{class\times L}$, class is number of candidate answers. p_{class} represents the probability over the candidate answers:

$$p_{class} = Softmax\left(W_{class}\overline{f}_{e\cdot o\cdot m} + b_{class}\right) \tag{7}$$

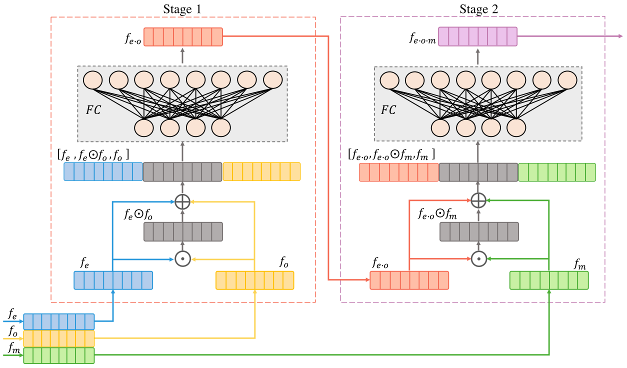


Fig. 2 The multi-stage fusion in MSMDFN. Multi-Stage fusion begins with the extracted features. For illustration purpose, coefficient between f_e and f_o is maximum so the first fusion stage selects and creates an intermediate representation $f_{e\cdot o}$. The second fusion stage select f_m to be fused with intermediate vector $f_{e\cdot o}$ obtained in the previous stage

Experiments

Dataset

To evaluate the performance of MSMDFN, the multimodal emotion dataset DEAP (Koelstra et al. 2011) is adopted. DEAP records EEG, EOG, EMG and other peripheral physiological signal from 32 participants while watching 40 one-minutes long music videos. The 32-channel EEG signal was collected using an EEG cap designed according to the 10–20 international lead standard. After viewing a video, each subject rated the current clip on a scale of 1–9 in terms of validity, arousal, dominance, and liking. The dataset used in the following experiments is a pre-processed version that has down sampled the signal to 128 Hz.

In our experiment, raw data include 36 channels(32 EEG-channels, 2 EOG-channels and 2 EMG-channels) were selected to validate MSMDFN. As shown in Fig. 3, The two dimensions (Valence, Arousal) were selected for binary-classification and four-classification. The binary-classification was divided into high valence (HV)/low valence (LV) or high arousal (HA)/low arousal (LA) using 5 as the threshold. The four-classification is based on both Valence and Arousal dimensions. Four categories HAHV, HALV, LAHV and LALV are set using 5 as the threshold in two dimensions independently. A ten-fold cross-validation method was used for the experiments.

Preprocessing and training setup

There are totally 36-channel physiology signals involve three modalities(EEG, EOG and EMG). In the data

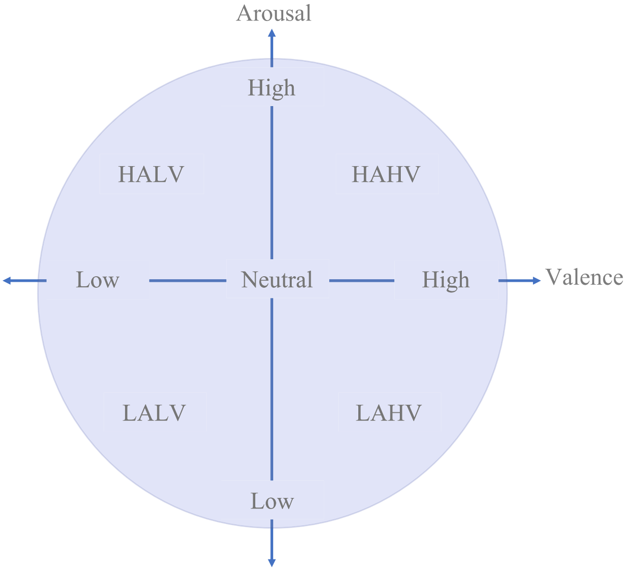


Fig. 3 The four-classification Arousal-Valence: LALV, HALV, LAHV and HAHV

acquisition experiments, each video clip was acquired for 63 seconds. The first 3 seconds were the resting-state signal before viewing the clip, and the rest 60 seconds were the task-state signal. Each 63 second-long trial was cut into 1 second-long samples. In this way, the signal of each subject can be transformed to the size of $[40, 36, 63, 128]$. In order to reduce the interference of external factors such as experimental environment, the resting-state mean matrix was obtained by averaging the first 3 seconds signal, and the resting-state mean matrix was subtracted from the task-state signal in the 60 seconds. The final data can be obtained in $[2400, 36, 128]$ format. The model was trained with a learning rate of 0.001, batchsize is 32 and patience is 20. For avoiding over-fitting, the dropout is set to 0.25. Parameters of MSMDFN are set as follows: D=2, N=8, and M=16. The wavelet transform used the Python wavelet analysis library Pywavelets (Lee et al. 2019), and the wavelet type used for the experiments was 'cgau8'. To verify the effectiveness of the fusion method, the experiments were finalized using a fully connected layer and a Softmax activation function layer for emotion classification. The loss function used for model training was the cross-entropy loss function and the optimizer was Adam.

Result and discussions

Comparison with dxisting models

Firstly, we compared the proposed method with multi-modal physiological signal fusion methods that have been proposed in recent years. Table. 1 shows the performance of various models for different classification tasks. For a more detailed comparison, we pointed out the modalities used in each method. The bottom row of table 1 shows that the framework we proposed has better performance in the classification task on multimodal physiological signals. On multimodal signal DEAP dataset, MSMDFN exceeded the existing best method MM-ResLSTM and CNN + LSTM. In valence dimension, MSMDFN exceeded the previous best model MM-ResLSTM by a margin of 6.55%. Additionally, MSMDFN improved 5.71% compared with best model CNN+LSTM in Arousal dimension. It is interesting to see that performance of some existing manners is significantly different under various tasks. The results of MSDMDFN obviously reveal that the large discrepencies under diverse tasks have been shrunk.

Improvement of wavelet transformers

Due to the low signal-to-noise ratio of the original EOG and EMG, a round of noise reduction can be performed on the signal through wavelet transform. Table. 2 shows the improvement of wavelet transformers in EMG and EOG extractors according to the comparison with CNN framework contains time-frequency maps which is specially used for EEG-based analysis. The experimental results show that the wavelet transformer can improve the signal-to-noise ratio of the data so that the subsequent feature extraction model can extract more effective emotional information. Additionally, the performance of wavelet transformer is more evident in arousal binary-classification.

Performance on multiple modalities

In order to explore the effectiveness of multimodal physiological signal, the features obtained by the MSMDFN method were firstly compared with the single model EEG,

Table 1 Comparison of performance for emotion classificaiton accuracy on DEAP using various models. Valence and Arousal are binary-classifications, V-A(Valence-Arousal) is four-classification

Method	Signal	DEAP		
		Valence	Arousal	V-A
HMMChen et al. (2015)	All	83.98.	85.63	–
KNN+RFChen et al. (2016)	All	–	–	70.04
BDAELiu et al. (2016)	EEG, EOG	85.2	80.5	–
Bimodal-LSTMTang et al. (2017)	All	83.82	83.23	–
CNNLin et al. (2017)	All	85.5	87.3	–
DCCAQiu et al. (2018)	EEG, EOG	85.62	84.33	85.51
MM-ResLSTMMa et al. (2019)	All	92.3	92.87	–
AAGYilmaz and Kose (2021)	EEG, EOG	90.31	91.53	–
CNN + LSTMLiao et al. (2020)	All	91.95	93.06	–
NASLi et al. (2021a)	EEG	97.74	97.94	–
ASTG-LSTMLi et al. (2021b)	EEG	98.71	98.71	**98.28**
MSMDFN (ours)	EEG, EOG, EMG	**98.85**	**98.77**	98.14

Table 2 Performance of models with/without wavelet transformer in EOG and EMG extracotr

Valence	Without	With
Precision	96.79 ± 1.95	98.29 ± 0.81
Sensitivity	96.66 ± 1.91	98.23 ± 0.83
Accuracy	97.06 ± 1.37	98.85 ± 0.70
Specificity	95.68 ± 3.59	97.97 ± 1.00
Arousal	Without	With
Precision	89.53 ± 1.73	98.28 ± 0.96
Sensitivity	90.30 ± 1.46	98.13 ± 0.95
Accuracy	95.09 ± 4.64	98.77 ± 0.84
Specificity	83.94 ± 3.01	97.70 ± 1.21

Table 3 Performance of different modalities under Valence (binary-classification), Arousal (binary-classification) and Valence-Arousal (V-A, four-classification) emotion recognition tasks

Modalities	Valence	Arousal	V-A
EEG	96.39 ± 3.48	94.88 ± 3.26	91.01 ± 4.74
EEG+EOG	98.15 ± 1.17	98.11 ± 1.10	97.05 ± 1.62
EEG+EMG	98.33 ± 1.00	98.33 ± 1.16	97.47 ± 1.31
EOG+EMG	97.62 ± 0.92	97.65 ± 1.15	96.15 ± 1.57
EEG+EOG+EMG	98.85 ± 0.70	98.77 ± 0.84	98.14 ± 1.06

and then compared with the feature obtained by bimodal fusion.

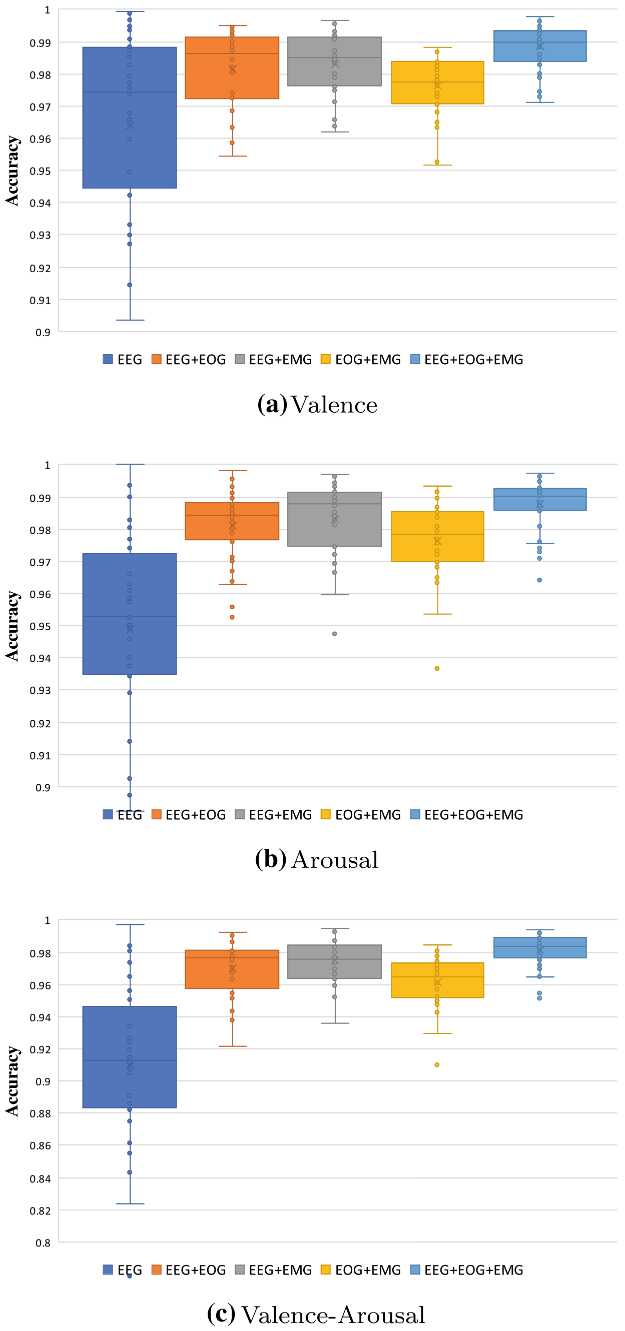
The EEG uni-modal feature extractor is the corresponding modal in MSMDFN. The performance of EEG feature and multimodal feature are shown in Table 3. As can be seen from table, the mean accuracy of unimodal EEG signals in the emotion recognition classification task was 96.39%, 94.88% and 91.01% with standard deviations of 3.48%, 3.26% and 4.74% respectively. The mean accuracy of MSMDFN was 98.85%, 98.77% and 98.14% with standard deviations of 0.7%, 0.84% and 1.06% respectively. It is known from the experimental result that the multimodal feature fused from three modalities(EEG, EOG and EMG) is significantly better than EEG unimodal in the emotion recognition task. On the mean classification accuracy of 32 subjects, the valence binary-classification task improved by 2.46%, the arousal binary-classification task improved by 3.89%, and the validity-arousal four-classification task improved by 7.13%. By comparing with the EEG unimodal emotion-based features, it can be verified that the proposed method is more effective in the emotion recognition task. After using MSMDFN, the classification results of individuals have been improved significantly. In addition, when comparing the emotion recognition ability of the proposed method in this paper for signals collected from the same subject in different tasks, it can be found that the classification performance improvement of this method is more obvious in the four-classification task than in the binary-classification task.

In order to ensure the rigor of the experiment, the bi-modal feature firstly uses the feature extraction method corresponding to the modal in MSMDFN to extract the features, and then uses the stage fusion method to fuse the features of the two modalities. Fig. 4 shows the emotion recognition results of 32 subjects with different modalities. EEG had poorer results under all three tasks, which indicates that EEG uni-modal does not yet have well generalization ability on the emotion recognition task and has poor performance capability on some subjects. From Fig. 4, it can be found that the emotion recognition effect of the features extracted by the bi-modal fusion method is better than that of the EEG uni-modal, but worse than the three-modality emotion feature. The experimental comparison between uni-modal and bi-modal features shows that the emotional information contained in the three-modality features is more sufficient. Multimodal data can make up for the shortcomings of certain subjects' EEG uni-modal data in emotion recognition tasks. Emotion recognition methods that incorporate multimodal features have a wider application range and more generalization than EEG emotion recognition. In addition, when comparing the emotion recognition ability of the method proposed in this paper on the signals collected by the same subject in different tasks, it can be found that the classification performance of this method in the four-classification task is more obvious than in the binary-classification task.

Performance on different fusion strategies

In order to validate the effectiveness of the multi-stage architecture, we also perform a comparison of one-stage fusion manner by using multimodal features extracted from the same extractors in MSMDFN. After obtaining the features of three modalities, the three one-dimensional feature vectors are concatenated into a new one-dimensional vector as a multimodal feature vector. The fused multimodal features are fed into the fully connected layer for classification task. The classification accuracy of one-stage fusion method was respectively 97.91%, 97.93% and 94.1% under Valence, Arousal and Valence-Arousal tasks. From the classification results, it can be seen that the performance of one-stage fusion is higher than the results of uni-modal, which indicates that the multimodal physiological signals contain more representations of emotion. The classification accuracy of the multimodal features obtained by MSMDFN are higher than the results of one-stage fusion method. The comparison results exhibited that the multimodal features extracted by MSMDFN observed more cross-modal interaction information. Consequently, MSMDFN is more superior on emotion recognition task.

After compared multi-stage fusion manner with one-stage, we also investigated the influence of multi-stage fusion order on MSMDFN. For comparison, experiments were performed according to the correlation coefficient in ascending/descending order. The experiment result illustrated that it is to fuse subset modalities in descending. The performance of two fusion orders demonstrated that descending order is able to maintain the essential cross-modal interaction and inter-modal messages between two more correlated modalities. Additionally, the relatively large redundant information among them is discarded.

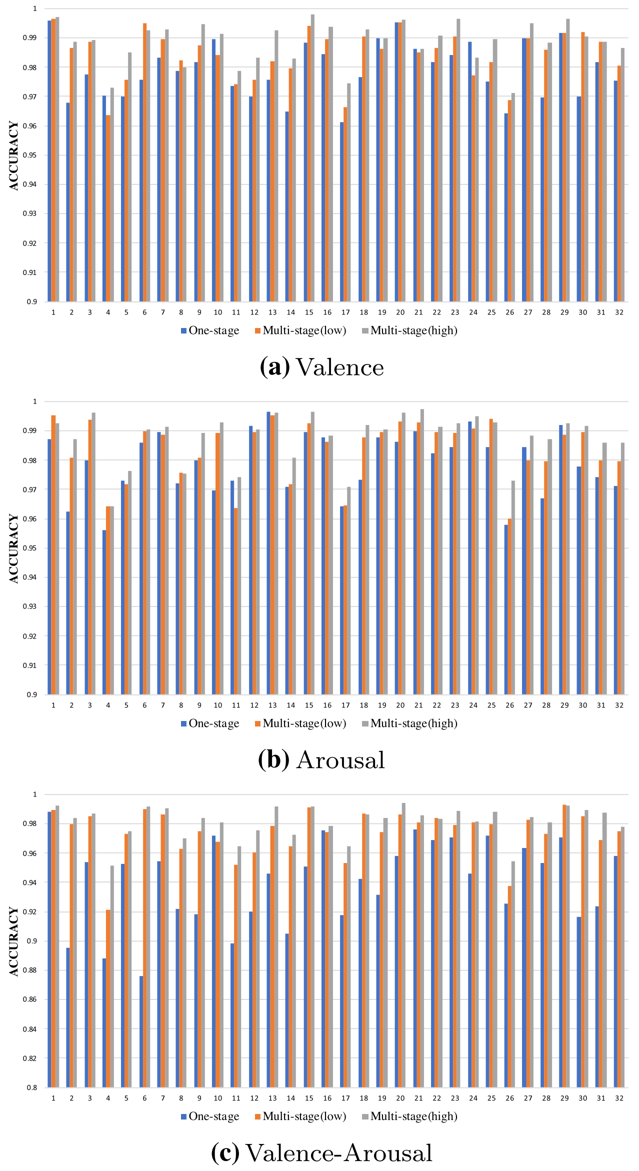


(a) Valence

(b) Arousal

(c) Valence-Arousal

Fig. 4 Classification accuracy of multiple modalities under Valence(binary-classification), Arousal(binary-classification) and Valence-Arousal(four-classification) three emotion recognition tasks



(a) Valence

(b) Arousal

(c) Valence-Arousal

Fig. 5 Classification accuracy of different fusion manners under Valence(binary-classification), Arousal(binary-classification) and Valence-Arousal(four-classification) three emotion recognition tasks. High/low means the dynamic multi-stage fusion order is ascent/descent

Table 4 Performance of different fusion methods under Valence(binary-classification), Arousal(binary-classification) and Valence-Arousal(four-classification) emotion recognition tasks. High/Low means the dynamic multi-stage fusion order is ascent/descent

	Valence	Arousal	V-A
One-stage	97.90 ± 0.93	97.93 ± 1.06	94.10 ± 2.87
Multi-stage (low)	98.45 ± 0.82	98.37 ± 1.01	97.42 ± 1.56
Multi-stage (high)	98.85 ± 0.70	98.77 ± 0.84	98.14 ± 1.06

Performance on different subjects

In addition, this paper has also investigated the differences exhibited by different multimodal fusion methods of emotion recognition on different subjects. For effective comparison, the classification results in 32 subjects under one-stage and multi-stage were used as indicators to observe the differences. Fig. 5 shows, under the same task, the emotion recognition ability performed by the different fusion method on the data of different subjects. As can be seen from the three figures, the variances in the classification performance of one-stage fusion for different subjects' physiological signal performance is relatively large. However, under the multi-stage fusion method, the emotion recognition results of multimodal signals of different subjects are more stable and have a higher accuracy rate. All the above results can indicate that the proposed method is more generalizable and shows excellent emotion recognition ability for different subjects.

Conclusion

In this paper, we investigate the effectiveness of different modalities and find that multimodal signals has better performance for emotion recognition. Then, we propose multi-stage multimodal dynamic fusion network (MSMDFN) which sequentially models the joint representation based on cross-modal correlation. MSMDFN uses three different extractors which are trained jointly to learn intra-modal features and then obtains inter-modal interactive information by multi-stage multimodal fusion which is based on bi-modal correlation. Extensive experiments based on DEAP reveal that MSMDFN has better ability in multimodal emotion recognition. MSMDFN effectively exploits much more fine-grained and comprehensive intercorrelations among multimodal signals. However, this model currently adopts element-wise product to fuse bimodal signals, which lacks more adaptable fusion method. In future work, we are interested in fusing bimodal data onto more complicated method of self-attention.

Acknowledgements This work was supported by National Key R &D Program of China for Intergovernmental International Science and Technology Innovation Cooperation Project (2017YFE0116800), National Natural Science Foundation of China (U20B2074, U1909202), Science and Technology Program of Zhejiang Province (2021C03003), and supported by Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province (2020E10010).

Data Availability The datasets analysed during the current study are available in the DEAP repository, https://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

AlZoubi O, D'Mello SK, Calvo RA (2012) Detecting naturalistic expressions of nonbasic affect using physiological signals. IEEE Transact Affect Comput 3(3):298–310

Balasubramani PP, Chakravarthy VS (2020) Bipolar oscillations between positive and negative mood states in a computational model of basal ganglia. Cognitive Neurodyn 14(2):181–202

Chen J, Hu B, Moore P, Zhang X, Ma X (2015) Electroencephalogram-based emotion assessment system using ontology and data mining techniques. Appl Soft Comput 30:663–674

Chen J, Hu B, Wang Y, Dai Y, Yao Y, Zhao S (2016) A three-stage decision framework for multi-subject emotion recognition using physiological signals. In: 2016 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 470–474

Chen J, Hu B, Xu L, Moore P, Su Y (2015) Feature-level fusion of multimodal physiological signals for emotion recognition. In: 2015 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 395–399

Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp 1251–1258

Davidson RJ, Fox NA (1982) Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants. Science 218(4578):1235–1237

Ghahari S, Farahani N, Fatemizadeh E, Motie Nasrabadi A (2020) Investigating time-varying functional connectivity derived from the jackknife correlation method for distinguishing between emotions in fmri data. Cognitive Neurodyn 14(4):457–471

Goshvarpour A, Goshvarpour A (2019) Eeg spectral powers and source localization in depressing, sad, and fun music videos focusing on gender differences. Cognitive neurodyn 13(2):161–173

Hinton G, Deng L, Yu D, Dahl GE, Mohamed AR, Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN et al (2012) Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Sig Process Mag 29(6):82–97

Huang H, Hu Z, Wang W, Wu M (2019) Multimodal emotion recognition based on ensemble convolutional neural network. IEEE Access 8:3265–3271

Kim BH, Jo S (2018) Deep physiological affect network for the recognition of human emotions. IEEE Transact Affect Comput 11(2):230–243

Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2011) Deap: a database for emotion analysis; using physiological signals. IEEE transact Affect Comput 3(1):18–31

Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ (2018) Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. J Neural Eng 15(5):056013

Lee GR, Gommers R, Waselewski F, Wohlfahrt K, O'Leary A (2019) Pywavelets: a python package for wavelet analysis. J Open Source Softw 4(36):1237

Liao J, Zhong Q, Zhu Y, Cai D (2020) Multimodal physiological signal emotion recognition based on convolutional recurrent neural network. In: IOP conference series: materials science and engineering, vol 782, IOP Publishing, p 032005

Lin W, Li C, Sun S (2017) Deep convolutional neural network for emotion recognition using eeg and peripheral physiological signal. In: International conference on image and graphics, Springer, pp 385–394

Liu W, Zheng WL, Lu BL (2016) Emotion recognition using multimodal deep learning. In: International conference on neural information processing, Springer, pp 521–529

Li C, Zhang Z, Song R, Cheng J, Liu Y, Chen X (2021) Eeg-based emotion recognition via neural architecture search. IEEE Transact Affect Comput

Li X, Zheng W, Zong Y, Chang H, Lu C (2021) Attention-based spatio-temporal graphic lstm for eeg emotion recognition. In: 2021 International joint conference on neural networks (IJCNN). IEEE, pp 1–8

Mai S, Hu H, Xing S (2020) Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion. In: proceedings of the AAAI conference on artificial intelligence, vol 34, pp 164–172

Ma J, Tang H, Zheng WL, Lu BL (2019) Emotion recognition using multimodal residual lstm network. In: proceedings of the 27th ACM international conference on multimedia, pp 176–183

Mehdizadehfar V, Ghassemi F, Fallah A, Mohammad-Rezazadeh I, Pouretemad H (2020) Brain connectivity analysis in fathers of children with autism. Cognitive Neurodyn 14(6):781–793

Qiu JL, Liu W, Lu BL (2018) Multi-view emotion recognition using deep canonical correlation analysis. In: international conference on neural information processing, Springer, pp 221–231

Shi LC, Jiao YY, Lu BL (2013) Differential entropy feature for eeg-based vigilance estimation. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC), IEEE, pp 6627–6630

Shu L, Xie J, Yang M, Li Z, Li Z, Liao D, Xu X, Yang X (2018) A review of emotion recognition using physiological signals. Sensors 18(7):2074

Song T, Zheng W, Song P, Cui Z (2018) Eeg emotion recognition using dynamical graph convolutional neural networks. IEEE Transact Affect Comput 11(3):532–541

Tang H, Liu W, Zheng WL, Lu BL (2017) Multimodal emotion recognition using deep neural networks. In: international conference on neural information processing, Springer, pp 811–819

Thammasan N, Moriyama K, Fukui KI, Numao M (2016) Continuous music-emotion recognition based on electroencephalogram. IEICE Transact Inf Syst 99(4):1234–1241

Yilmaz BH, Kose C (2021) A novel signal to image transformation and feature level fusion for multimodal emotion recognition. Biomed Eng/Biomed Tech 66(4):353–362

Zhang D, Yao L, Zhang X, Wang S, Chen W, Boots R, Benatallah B (2018) Cascade and parallel convolutional recurrent neural networks on eeg-based intention recognition for brain computer interface. In: proceedings of the AAAI conference on artificial intelligence, vol 32

Zheng WL, Lu BL (2015) Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. IEEE Transact Auton Mental Dev 7(3):162–175

Zheng WL, Zhu JY, Lu BL (2017) Identifying stable patterns over time for emotion recognition from eeg. IEEE Transact Affect Comput 10(3):417–429

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.