Digital Signal Processing 117 (2021) 103153

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Emotion recognition based on fusion of long short-term memory networks and SVMs

Tian Chen [a,b,c], Hongfang Yin [a,b,c], Xiaohui Yuan [e,*], Yu Gu [a,b,c], Fuji Ren [a,b,c,d], Xiao Sun [a,b,c]

[a] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, 230009, China
[b] Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology, China
[c] Intelligent Interconnected Systems Laboratory of Anhui Province, Hefei University of Technology, China
[d] Faculty of Engineering, The University of Tokushima, Tokushima, Japan
[e] Department of Computer Science and Engineering, University of North Texas, 76207, USA

ARTICLE INFO

Article history:
Available online 9 July 2021

Keywords:
EEG
ECG
DS theory
Multimodal
Emotion recognition

ABSTRACT

This paper proposes a multimodal fusion emotion recognition method based on Dempster-Shafer evidence theory, which includes electroencephalogram (EEG) and electrocardiogram (ECG). For EEG, we use the SVM classifier to classify features, and for ECG, we establish the corresponding Bi-directional Long Short-Term Memory network emotion recognition structure, which is fused with EEG classification results through the evidence theory. We selected 25 video clips with five emotions (happy, relaxed, angry, sad, and disgusted), and a total of 20 subjects participated in our emotional experiment. The experimental results prove that the performance of the multi-modal fusion model proposed in this paper is superior to the single-modal emotion recognition model. In the Arousal and Valance dimensions, the average accuracy is improved by 2.64% and 2.75% compared with the EEG signal-based emotion recognition model. Compared with the emotion recognition model based on the ECG signal, the accuracy is improved by 7.37% and 8.73%.

1. Introduction

To improve the ability to cooperate or interact with others, intelligent human-machine systems with the ability to accurately understand interpersonal communication are highly demanded [1]. Emotional computing plays an important part in human-computer interaction [2] and is a key stage in sentiment analysis. Many mental illnesses are expressed as emotions, such as depression, autism, and other diseases [3,4], which motivates the recognition and understanding of human emotions.

Physiological and non-physiological signals have been investigated for emotion recognition. Methods based on physiological signals demonstrated greater reliability. In recent years, studies have been conducted with physiological signals such as electroencephalogram (EEG), electrocardiogram (ECG), electrooculogram (EOG), and electromyography (EMG), among which EEG provides key information in the recognition of emotions [5]. The success of EEG-based emotion recognition inspired many methods [6–8]. However, as different signals present various aspects of emotions, integration of EEG with other physiological signals has been explored [9]. The challenges lie in the amalgamation of inconsistent signals distorted with noise [10].

In this article, we proposed a method that fuses the results of certain channel data classification using fuzzy integration to achieve an improved emotion recognition model. A Bi-directional Long Short-Term Memory (Bi-LSTM) network is developed to build an emotion recognition model. Features including heart rate (HR) and heart rate variability (HRV) in the time domains are extracted and classified with the Bi-LSTM network. The classification outputs are fused with the outputs from the classification of EEG signals to make the final emotion recognition.

The rest of this article is organized as follows. Section 2 briefly reviews the related methods for emotion recognition. Section 3 presents the preprocessing, feature extraction, classification methods, and multimodal fusion strategy. Section 4 discusses the detailed experimental steps, experimental preparations, and our experimental results in detail. Section 5 concludes this paper with a summary.

* Corresponding author.
E-mail address: xiaohui.yuan@unt.edu (X. Yuan).

https://doi.org/10.1016/j.dsp.2021.103153
1051-2004/

2. Related work

For sentiment computing, there are many signals available for us to use, including facial expressions, speech, and gestures [11,12]. Since cerebral cortex activity is necessarily related to human emotions, it is feasible to use EEG signals for emotion classification. Compared to the external signals described above, EEG has attracted much attention due to its low cost and high reproducibility and portable implementation [13].

EEG signals capture the changes in the electric potential of a human subject. It is necessary to extract features of the preprocessed EEG signals to reflect the distinction between different emotions. Features extracted from EEG include time-series features [14], spectrum features [15], spatial synchronization features [16] and complexity measurement features [17]. Krisnandhika et al. [18] studied the utilization of relative wavelet energy as the feature extraction, and a modified radial basis function neural networks are implemented as the classifier. Pereira et al. [19] explored the effect of different emotional stimulation times on emotional recognition rate. Higher-Order Crossing (HOC) feature values are selected and SVM is used as a classifier. When the duration of the collection exceeds 60 seconds, the emotion recognition rate is better.

Since EEG signal acquisition devices often have multiple channels, to make better use of the information on each channel, Chao et al. [20] proposed a new integrated deep learning framework, which integrates parallel DBN-GC and CRF, applies to multiple channels of EEG signals, and finally obtains the final sentiment prediction result through the KNN-based decision merge layer. Zheng et al. [21] demonstrated that in the EEG-based emotion recognition process, using a subset of data from a few selected channels obtains the highest emotion recognition accuracy.

To improve the performance, feature selection has been adopted and developed for ECG signal processing. Ferdinando et al. [22] investigated supervised dimensionality reduction, linear discriminant analysis (LDA), NCA (neighborhood component analysis), and MCML (maximum fold metric learning) for emotion recognition based on ECG signals from the Mahnob-HCI database. The classification was performed using the KNN classifier and the results showed that NCA outperformed the other methods. Mert and Akan [8] adopted multivariate synchrosqueezing transform to extract time-frequency features from EEG signals and applied a neural network to make the classification. Hsu et al. [23] proposed an ECG-based emotion recognition algorithm using a class separability-based (SFSF-KBCS's) feature selection algorithm based on a sequence forward floating selection kernel and utilized generalized discriminant analysis (GDA) to efficiently select important ECG features associated with emotions and to reduce the selection function. A least-squares support vector machine (LS-SVM) is used for emotion recognition.

Emotion has a close relationship with physiological and psychological changes, and signals from different modalities can reflect different information about emotion. Hence, the fusion of signals from multiple modalities is beneficial to make full use of all the information to get a more stable and higher recognition accuracy emotion recognition model. Asibul Islam et al. [24] carried out a multimodal fusion of EEG signals and facial expressions. The final experimental results indicate that the accuracy after multi-modal fusion is higher than that of the individual modal. Yea-Hoon et al. [25] combined EEG signals with GSR signals, and used CNN to fuse EEG spectrograms with GSR features, and finally completed the process of multimodal emotion recognition. Katsigiannis et al. [26] used a portable device to acquire EEG and ECG signals during emotion elicitation by audiovisual stimuli. PSD features were extracted from the EEG signals, where were fused with the HRV and HR features of the ECG signal at the feature level. The classification was achieved with an SVM. The experimental results demonstrated that the recognition accuracy obtained is better than that of unimodal EEG and ECG in the Arousal dimension. Zhao et al. [27] performed a multimodal fusion of EEG with other physiological signals, including EOG and EMG. The EEG signal was used as auxiliary information during training. The other physiological signals were used to create a new emotion recognition space using DCCA (discriminative canonical correlation analysis). In the testing stage, only other physiological signals used for testing are spatially projected and classified. The results show that combining EEG and ECG improves the emotion recognition rate. However, the recognition rate does not meet the real-life application needs. Therefore, improving the multimodal fusion emotion recognition rate is still an open challenge. This paper proposes a multimodal fusion emotion recognition model based on Dempster–Shafer (DS) evidence theory.

3. Proposed method

For emotion recognition, one of the most important steps is feature extraction, in which we extract its time domain, frequency domain, or time-frequency domain features from the signal. Fig. 1 illustrates an overall view of the proposed method. For EEG data, we extracted features in five frequency bands and all bands. The five frequency bands are Delta(1-3 Hz), Theta(4-7 Hz), Alpha(8-13 Hz), Beta(14-30 Hz), Gamma(31-43 Hz), and classified them using SVM. For the ECG data, we extracted the features of HR and HRV related parameters and classified them by Bi-LSTM network. The classification results of EEG signals and ECG signals are fused using the DS theory.

3.1. Feature extraction

For the EEG signal, we extract four kinds of features, the Lempel-Ziv complexity, the wavelet detail factor, the degree of cointegration relationship, and the approximate entropy after Empirical Mode Decomposition (EMD) [28]. For the Lempel-Ziv complexity, it is first necessary to binarize the raw signal to 0 and 1 according to the threshold T (usually using the median) to obtain the sequence $P = s_1, s_2, ..., s_n$, traverse the sequence P, and increase the complexity counter $c(n)$ by one unit when a new subsequence appears, i.e., $c(n)$ is the number of new patterns in the sequence P and n is the length of the sequence.

The specific algorithm is the following. We construct the sequences S, Q and SQ, where SQ is the concatenation of S and Q, $SQ\pi$ is the sequence obtained by deleting the last character of the SQ sequence, and $v(SQ\pi)$ denotes the set of all distinct subsequences in $SQ\pi$. Initially, let $c(n) = 1$, $S = s_1$, $Q = s_2$, $SQ = s_1, s_2$, $SQ\pi = s_1$, and in general, $S = s_1, s_2, ..., s_r$, $Q = s_{r+1}$, $SQ = s_1, s_2, ..., s_r, s_{r+1}$, and $SQ\pi = s_1, s_2, ..., s_r$. If Q belongs to $v(SQ\pi)$, then Q is a subsequence of $SQ\pi$, S remains unchanged, Q is updated to $Q = s_{r+1}, s_{r+2}$, and continues to determine whether Q belongs to $SQ\pi$ until Q does not belong to $SQ\pi$. If Q does not belong to $SQ\pi$, $c(n)$ is increased by one unit, S is updated to $S = s_1, s_2, ..., s_{r+i}$, and Q is updated to $Q = s_{r+i+1}$. The above process is repeated till the last character in Q. The resulting complexity counter $c(n)$ is time-dependent and needs to be normalized. An upper bound of $c(n)$ is given by [29],

$$c(n) < \frac{n}{(1-\varepsilon_n)\log_\alpha(n)} \tag{1}$$

where α represents the number of possible symbols in the sequence P. Thus $\alpha = 2$ and $\varepsilon_n\,(n \to \infty)$ converges to 0, i.e.,

$$\lim_{n\to\infty} c(n) = b(n) = \frac{n}{\log_2(n)}. \tag{2}$$

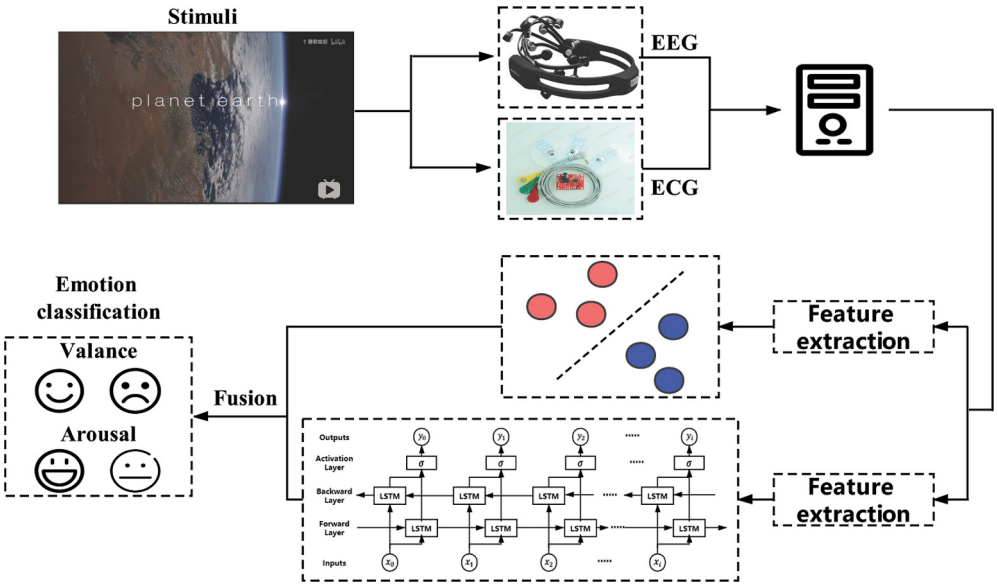


Fig. 1. An overall view of the proposed method for emotion recognition.

The Lempel-Ziv complexity after normalization is

$$LZC = \frac{c(n)}{b(n)} = \frac{c(n) \times \log_2 n}{n}. \tag{3}$$

LZC is the normalized Lempel-Ziv complexity, which represents the growth rate of the new pattern of the sequence. The wavelet detail coefficients are the mean values of the wavelet coefficients computed after triple decomposition using db5 wavelets under each channel. An optimal channel is selected from the other channels using the ReLiefF algorithm [30]. The amount of cointegration relationship between that a channel and its optimal channel is the degree of cointegration relationship. The amount of cointegration relationship is the mean value of the number of cointegration relationships to the two channels by EG test [31] method in the sample time. A finite number of Intrinsic Mode Functions (IMFs) are obtained after EMD decomposition. Since the variance contribution is an indicator for evaluating IMFs, and the first four IMFs reach more than 95% of the sum of the variance of all IMFs, the first four IMFs are selected and their approximate entropy is calculated to obtain the feature.

ECG signals are typically characterized by extracting specific parameters of heart rate (HR) and heart rate variability (HRV) in the time and frequency domains. Before extracting the features, the signal was divided with an overlap-free window of length 15s, and a total of 4200 samples were acquired. A complete ECG signal mainly consists of P-wave, Q-wave, R-wave, S-wave, and T-wave. Therefore, we can extract these five waveforms from the preprocessed ECG signal. Fig. 2 shows that the R-wave is at its highest peak, and when the R-wave is extracted, the other waveforms are correspondingly easier to extract. We use the Pan-Tompkins QRS detection algorithm [32] to detect the R wave and then detect the other four peaks separately.

In our method, we calculate the statistical characteristics of P-wave, Q-wave, R-wave, S-wave, and T-wave, including the mean, median, standard deviation, maximum, minimum, and the difference between maximum and minimum values, as well as the PQ spacing, QR spacing, RS spacing, and ST spacing. The HRV-related parameters are then extracted, and in the time domain, the number of all normal heartbeat intervals (NN), the mean of all normal heartbeat intervals (NNAVG), the mean of R-wave intervals (MeanRR), the standard deviation of all normal heartbeat intervals (SDNN), the number of all pairs of vector normal heartbeat intervals that differ by more than 50 ms (NN50), NN50 divided by the total number of intervals between normal heartbeats for all pairs (pNN50), the difference between consecutive RR intervals (RMSSD), the quotient of SDNN divided by RMSSD (SDRM). Table 1 shows the features extracted from the EEG and ECG signals.

3.2. Emotion classification

For the EEG signals, there are four types of features used for classification, as the input to the SVM classifier. The collected EEG signals have a total of 14 channels that are positioned in various locations on the surface of the head. Considering the influence of different brain regions on emotion recognition, the classification results obtained by each channel are fused based on Takagi-Sugeno fuzzy integration. The classification result obtained for each classifier can be expressed as (a, b), where $(1, 0)$ represents the first class and $(0, 1)$ represents the second class. For all channels in the classification result, a composes the sequence $\vec{x} = \{ x_1, x_2, \ldots, x_n \}$, b composes the sequence $\vec{y} = \{ y_1, y_2, \ldots, y_n \}$, and n is the total number of channels. The fuzzy integral of the two sequences $\vec{x}$ and $\vec{y}$ is calculated as follows, taking the sequence $\vec{x}$ as an example,

$$\int f(\vec{x}) d\mu = \bigcup_{i=1}^{n} (f(x_i) \cap \mu(A_i)) \tag{4}$$

where $f(x_i)$ represents the value of a and $\mu(A_i)$ represents the joint fuzzy measure value of the corresponding channel. The fuzzy integral of the sequence $\vec{y}$ is calculated in the same way, comparing the fuzzy integral values of the two sequences, and if the former is large, the identification results in the first category, otherwise, the identification is of the second category.

In our previous work [28], the optimal channel selection experiment was also carried out, considering all channel combination cases, using an SVM classifier with an RBF kernel function and a penalty factor of 1. In each dimension, the channel combination with the highest average recognition rate was found separately. Finally, for the Valance dimension, the optimal channel combination was T7, T8, FC5, FC6, F3, and F4. For the Arousal dimension, the optimal channel combination was T7, T8, FC5, FC6, F7, and F8.

Long short-term memory (LSTM) network is a special type of Recurrent Neural Networks (RNNs) that combines short-term memory with long-term memory through gate control, which addresses the gradient disappearance problem [33]. Fig. 3 shows the

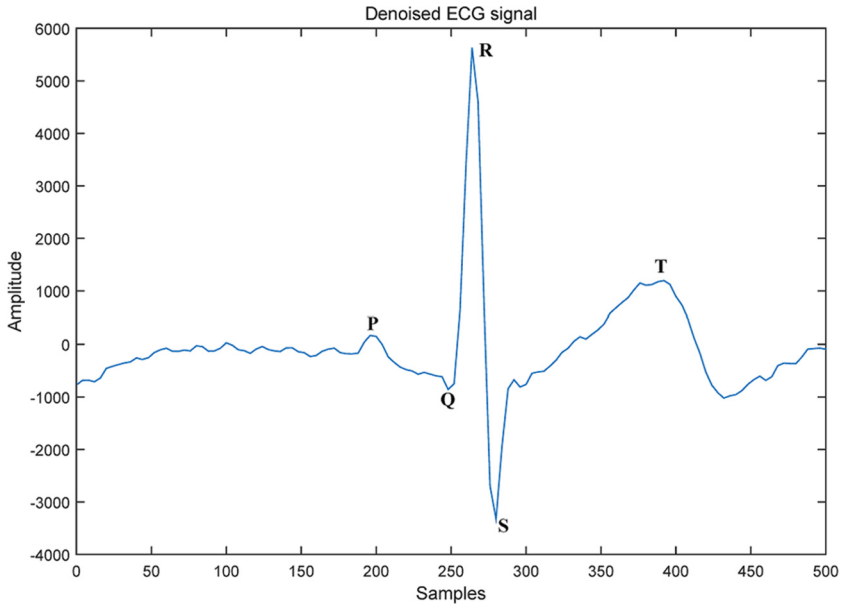


Fig. 2. ECG signal.

Table 1
Features Extracted from ECG and EEG Signals.

Modality	Extracted features	
EEG	Lempel-Ziv complexity	
	wavelet detail coefficients	
	degree of co-integration relationship	
	approximate entropy after EMD	
ECG	PQRST features	mean, median, standard deviation, min, max
		PQ,QR,RS,ST distance
	HRV features	NN, NNAVG, MeanRR, SDNN, NN50, pNN50, RMSSD, SDRM

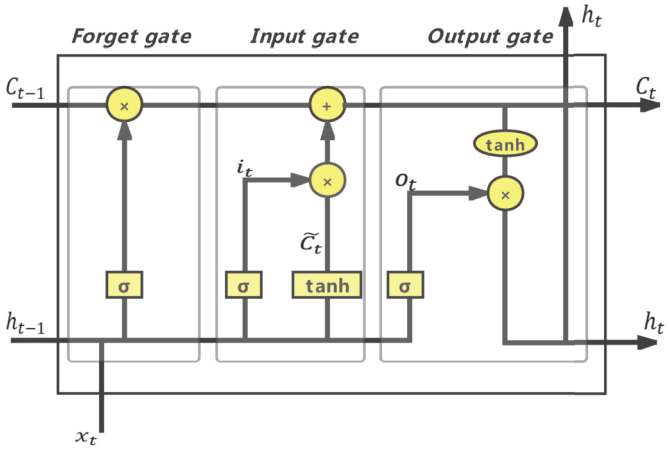


Fig. 3. The internal structure of the LSTM network.

internal structure of the LSTM network. LSTM has a long-term memory state C_{t-1} to the previous short-term memory h_{t-1} and new information x_t at the current moment to calculate the current short-term memory h_t, and adds an internal memory neuron $\widetilde{C}_t$ and three gates to control the passage of information: the forget gate f_t, the input gate i_t, and the output gate o_t. The forget gate determines how much information in C_{t-1} is forgotten. The input gate determines how much of the information in $\widetilde{C}_t$ is updated to the memory cell. The output gate is used to control how much of h_t depends on the current long-time memory cell C_t.

In the forget gate, the input sequence x_t and the output h_{t-1} from the previous moment are used as inputs for the current moment, and a number between 0 and 1 is output for C_{t-1} to complete the processing of the previous information. The activation of this gate is calculated as follows.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{5}$$

where W represents the independent weight vector for each input, b is the bias vector, and σ is the logistic sigmoid function, and the following is the same.

The input gate is divided into two parts, the first part determines the value to be updated via the sigmoid layer, and the second part creates a new candidate value vector $\widetilde{C}_t$ via the tanh layer. The calculation process is as follows:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \tag{6}$$

$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C). \tag{7}$$

After passing through the forget gate and the input gate, we obtain the new cell state C_t. The results are as follows:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t. \tag{8}$$

Finally, a sigmoid layer determines which parts of the cell state are output, and the output is obtained by multiplying the tanh by the output of the sigmoid layer. The calculation process is as shown follows:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o), \tag{9}$$

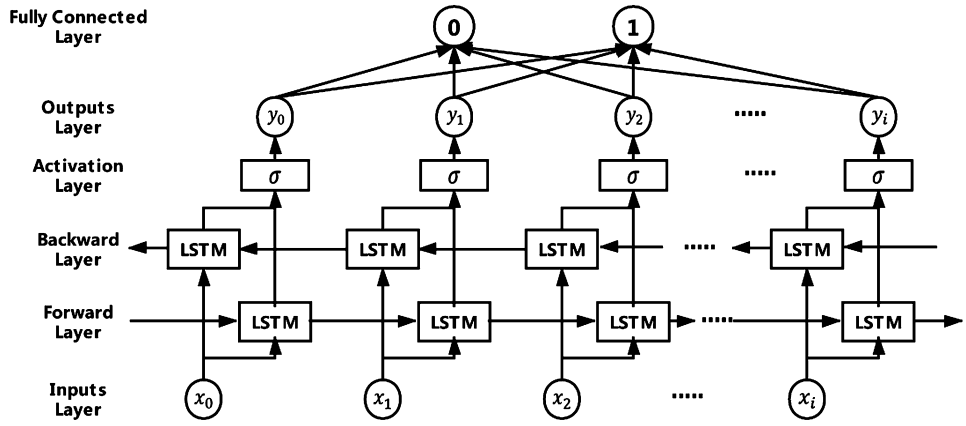


Fig. 4. Bi-directional LSTM network architecture.

$$h_t = o_t * tanh(C_t). \tag{10}$$

The LSTM network predicts the output of the next moment based on the timing information of the previous moment. Sometimes the output of the current moment is not only related to the previous state, but also the future state. Therefore, this paper uses a Bi-LSTM network with the network architecture shown in Fig. 4. A Bi-LSTM is a combination of a forward LSTM and a backward LSTM. The forward LSTM inputs the sequence in positive order and the backward LSTM inputs the sequence in reverse order, and then the LSTM output results of the corresponding input vector are connected, and the result of the connection is used as the new feature vector of the input vector, which takes into account both the last long-time memory state of the forward LSTM output and the last long-time memory state of the backward LSTM output, and the feature vector has the global information of the sequence. In this paper, the input data are time-domain features extracted from ECG data, and the LSTM layer in the Bi-LSTM network contains 10 hidden neurons. The network uses softmax as the activation function and Adam as the optimizer, with a learning rate of 0.01.

3.3. Decision fusion

To combine EEG and ECG signals in emotion recognition, we adopt a strategy of decision-level fusion based on Dempster–Shafer evidence theory. In the classification process, we classify on the Arousal and Valance dimensions respectively. There are two classes in each dimension, Category I and Category II, denoted with A and B, respectively. Given a set $X = \{X_1, X_2, ..., X_n\}$ as the set of recognition results, Θ is a recognition frame that contains all possible cases, including the empty set. Combining the n classifications results in a total of 2^n subsets. The Basic Probability Assignment (BPA) is a probability for each category, and function $m(\cdot)$ is the corresponding probability assignment function on the subset θ. For any subset θ of Θ, the following conditions are required:

$$\sum_{\theta \in \Theta} m(\theta) = 1, m(\emptyset) = 0, \text{ and } 0 \le m(\theta) \le 1.$$

In our fusion strategy, there are two classifiers for the EEG signal and ECG signal emotion recognition models. The probability assignment functions for the EEG signal are $m_1(A)$ and $m_1(B)$, and the probability assignment functions for the ECG signal are $m_2(A)$ and $m_2(B)$. These represent the confidence level of each classifier for the two classes. Table 2 presents the computation of the BPA function. In EEG, the BPA of each category is obtained by fuzzy integration. For example, the BPA function value of the first class is the ratio of the fuzzy integral value of the first class to the sum of the fuzzy integral values of the two classes. In ECG, when the softmax layer recognizes the emotion as the first type, the value of the first neuron represents its probability, which, in this case, is greater than the output of the second neuron.

When no conflict of evidence arises, for each BPA, the final BPA functions $m_1 \oplus m_2(A)$ and $m_1 \oplus m_2(B)$ for each category are calculated according to the Dempster synthesis rule, where $\oplus$ represents the combination of $m_1(A)$ and $m_2(A)$ or the combination of $m_1(B)$ and $m_2(B)$. The calculation process is shown in the following two equations, representing that all classifiers accumulate BPA functions that produce the same result value

$$m_1 \oplus m_2(A) = m_1(A) \cdot m_2(A), \tag{11}$$

$$m_1 \oplus m_2(B) = m_1(B) \cdot m_2(B). \tag{12}$$

When the values of $m_1 \oplus m_2(A)$ and $m_1 \oplus m_2(B)$ are obtained, the category that corresponds to the maximum value is the final classification result. The EEG-based sentiment recognition model is chosen as the final classification result when there is a conflict of evidence because the EEG-based sentiment recognition model performs better than the ECG-based sentiment recognition model for both pieces of evidence.

The overall framework is shown in Fig. 5. For EEG, the corresponding EEG features are extracted, and each channel uses LIBSVM as a classifier. For channels involved in emotion recognition, fuzzy integration is used for multimodal fusion. For ECG, PQRST and HRV features are extracted and classified using a Bi-LSTM network. The final emotion recognition results are obtained by processing the EEG-based and ECG-based emotion recognition results using DS evidence theory.

In the fusion process, classification results are obtained based on both EEG and ECG sentiment recognition models. Since it is a two-class classification in the Valance and Arousal dimensions, the BPAs of the EEG-based and ECG-based emotion recognition models are calculated for the two classes. The outputs are fused according to the Dempster synthesis rule to obtain the final BPAs, $m_1 \oplus m_2(A)$ and $m_1 \oplus m_2(B)$.

4. Experiments and results

4.1. Data sets and settings

In our data acquisition, 20 subjects participated in our emotion experiments, including 13 men and 7 women. All participants are college students with normal vision and normal hearing and without recent psychological or physiological disorders. We selected 25 video clips to elicit five emotions: happy, relaxed, angry, sad, and disgusted. Five video clips were used for each emotion. Emotiv

Table 2
BPA calculation.

	Category I (A)	Category II (B)
EEG (m_1)	$\frac{\int f(\vec{x})d\mu}{\int f(\vec{x})d\mu + \int f(\vec{y})d\mu}$	$\frac{\int f(\vec{y})d\mu}{\int f(\vec{x})d\mu + \int f(\vec{y})d\mu}$
ECG (m_2)	Output of the first neuron	Output of the second neuron

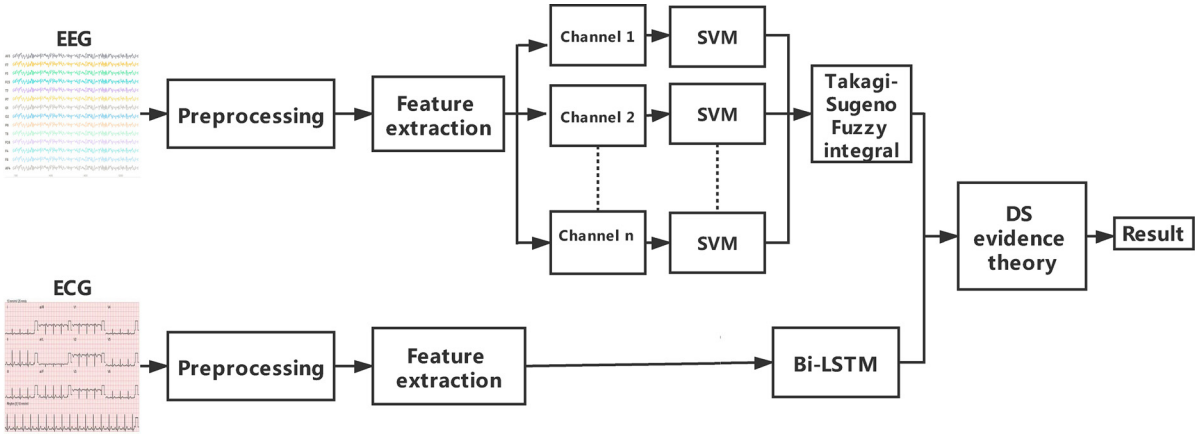


Fig. 5. The overall framework of in multimodal fusion process based on DS evidence theory.

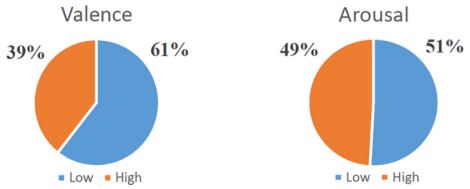


Fig. 6. Overall class distribution across all participants after conversion to a two-class rating score.

Epoc electroencephalograph was used to record the EEG signals. The sampling rate of this electroencephalograph is 128 Hz. There are a total of 14 channels and two reference electrodes, CMS and DRL, which are not used for EEG signal recording. The ECG signal was recorded by two ECG electrode patches, which are attached to the wrist pulses of the left and right hands, respectively.

Fig. 6 shows the category distribution of the dataset after the two classification schemes are divided according to the scoring scale. The datasets are slightly imbalanced in the Valance dimension with a ratio of 1.56 and mostly balanced in the Arousal dimension.

Fig. 7 shows the experimental process of each subject. There are a total of 25 video clips, 5 seconds of concentration and 5 seconds of text prompts before each video, and 45 seconds of self-assessment after watching the video, and finally take a 1-minute break. In the self-assessment phase, participants were asked to score on the Valance and Arousal dimensions. Among them, Valence (ranging from 1-9) represents the degree of happiness, that is, from a negative state to a positive state, Arousal (ranging from 1-9) represents the intensity of emotion [34].

For the label data, we defined two schemes: low/high valence (upset/happy) and low/high arousal (calm/arousal), which were subjectively assessed by each participant during the experiment using a scoring system from 0 to 9, and which served as our final labels. We used five as the midpoint when dividing label data according to their scores on the Valence and Arousal dimensions, and each dimension was divided into two categories.

To have a holistic model evaluation, accuracy, F1-scores, and AUC (Area Under roc Curve) are used in our evaluation. In our evaluation of the emotion recognition models, we divided the experimental data following the k-fold cross-validation strategy, where the percentage of each category of data contained in each fold is approximately the same as the percentage of the entire dataset. The number of folds is ten.

4.2. Accuracy analysis on bands

In the EEG-based emotion recognition process, we conducted experiments on using five bands and using all bands. Fig. 8 shows that the EEG signal is easier to distinguish after banding and the recognition rate is higher in the higher bands. As shown in Fig. 8(a), the beta band performs best in the Valance dimension, with an average accuracy of 82.24%. In the Valance dimension, the true-positive rate is 84.29% and the true-negative rate is 75.20%, which means that the beta band carries more positive emotional information than negative emotions. In addition, the beta band also obtained the best results in terms of F1-score and AUC, with values of 0.84 and 0.74, respectively. Compared to using all the bands, the average accuracy, F1-score, and AUC were improved by 0.2, 0.09, and 0.27, respectively.

Fig. 8(b) shows that the gamma band exhibits the best performance in the Arousal dimension with an average accuracy of 71.45%, an F1-score of 0.76, and an AUC of 0.68. Compared to using all bands, accuracy, F1-score, and AUC are improved by 0.25, 0.12, and 0.23, respectively. This indicates that filtering the EEG signal to a specific frequency band has a more pronounced effect on the accuracy and AUC.

4.3. Learning with Bi-LSTM

Fig. 9 shows the accuracy and loss for emotion recognition using the Bi-LSTM model based on ECG signals in the Valence dimension (A) and (C) and in the Arousal dimension (B) and (D). The x-axis is the number of epoch used in the training process. The training process converged fairly quickly in our evaluation. The loss plateaued after 50 epochs for the Arousal and continued decreasing slightly after 75 epochs for the Valence. After 100 epochs, both the accuracy and the loss rate became mostly stabilized with little fluctuation. The obtained accuracy of emotion recognition is 76.65% in the Valance dimension and 70.15% in the Arousal dimen-

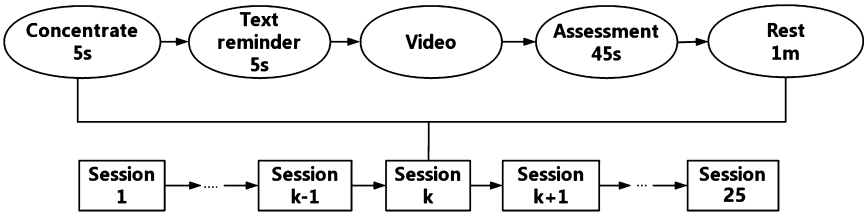


Fig. 7. Experimental flow chart.

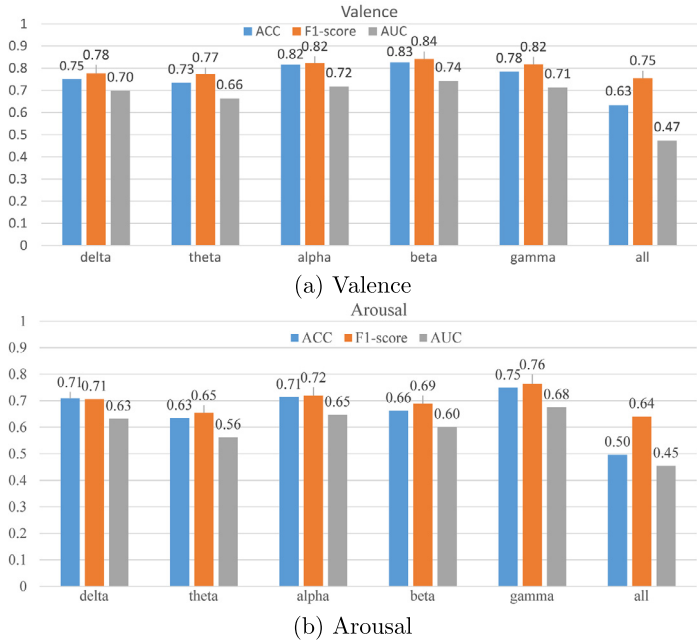


Fig. 8. Classification performance of five bands. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

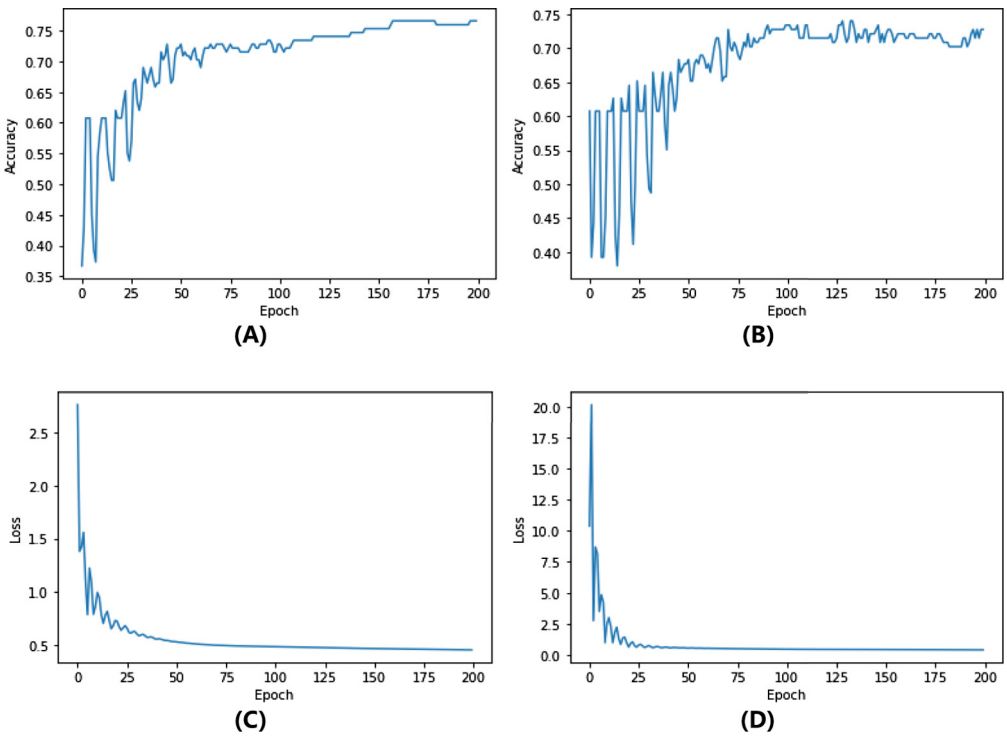


Fig. 9. Accuracy and Loss of ECG-based Bi-LSTM Models for Valence (A) and (C) and Arousal (B) and (D).

Table 3
Single-modal and multi-modal fusion emotion recognition accuracy.

Modality	Accuracy		F1-Score		AUC	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
EEG	82.63%	74.88%	0.8262	0.7640	0.7523	0.6751
ECG	76.65%	70.15%	0.7848	0.7149	0.6749	0.6775
Fusion	**85.38%**	**77.52%**	**0.8741**	**0.8409**	**0.7829**	**0.6816**

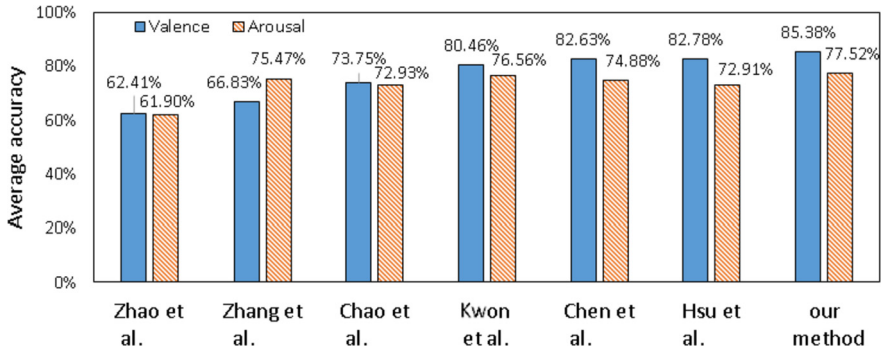


Fig. 10. Average accuracy of our proposed method and six state-of-the-art methods.

sion. Based on this training progress, we stop our training at the 200 epochs and take the model with the best performance in our experiments.

4.4. Effect of decision fusion

Table 3 presents the results of emotion recognition based on EEG, based on ECG, and multi-modal fusion of EEG and ECG. Compared with EEG-based and ECG-based single-modal emotion recognition models, the average classification accuracy of multimodal fusion is improved. With EEG, the average accuracy for valence and arousal are 82.63% and 74.88%, respectively. The average accuracy of a model using only ECG is 76.65% and 70.15%, respectively. Our proposed method fuses the outputs from SVM and LSTM and achieved accuracy for valence and arousal at 85.38% and 77.52%, respectively. In addition, the F1-score and AUC of the fused model have also been improved, which confirms that multimodal emotion recognition using fusion strategy improves the performance of single-modal emotion recognition models.

4.5. Comparison with state-of-the-art methods

Fig. 10 compares the average accuracy of our method and six state-of-the-art methods [20,23,25,27,28,35]. The accuracy of classifying for Valence and Arousal are depicted with separate bars. The compared methods are sorted according to the accuracy of Valence in ascending order from left to right. It is clear that our proposed method outperformed the other methods by achieving the best average accuracy for both valence and arousal. In contrast to the second-best for classifying valence and arousal, the improvement is 2.6% and 0.96%, respectively. In particular, when one type of signal is used, i.e., ECG or EEG, the information is limited [23,35]. Zhao et al. [27] integrated ECG by computing the statistics (e.g., mean and median) and concatenating them with EEG signals. The performance, although satisfactory, is inferior to the others, which implies simple concatenation of multimodal signals is arguable. As shown in Table 3, the best performance of using a single modality using our method is 82.63% and 74.88% for valence and arousal, which is competitive but not better. The fusion of the decisions improved the performance.

5. Conclusion

This paper presents a method that fuses EEG-based and ECG-based emotion recognition models at the decision level through DS evidence theory that relatively improve the performance of emotion recognition. Experimental results demonstrate that the fusion of EEG and ECG signal information provides more emotional information, and multimodal fusion improves the accuracy of emotion recognition. An ECG-based emotion recognition model is proposed for decision-level fusion with an EEG-based emotion recognition model using the DS theory. For ECG, PQRST and HRV features were extracted and then classified using a bidirectional LSTM network model to establish an ECG-based emotion recognition model with an accuracy of 76.65% in the Valance dimension and 70.15% in the Arousal dimension. For the fusion of EEG-based emotion recognition model and ECG-based emotion recognition model, i.e., the final fused BPA function value is calculated using the results of both models to obtain the final classification results. The emotion recognition accuracy obtained after the multimodal fusion is 85.38% in the Valance dimension and 77.52% in the Arousal dimension, which is better than the experimental results of the EEG and ECG unimodal models before the fusion, respectively, and significantly reflects the effect of fusing the EEG and ECG modalities.

CRediT authorship contribution statement

Tian Chen: Conceptualization, Investigation, Methodology, Writing – original draft. **Hongfang Yin:** Software, Validation, Writing – original draft. **Xiaohui Yuan:** Formal analysis, Investigation, Methodology, Writing – review & editing. **Yu Gu:** Data curation, Project administration, Resources. **Fuji Ren:** Funding acquisition, Project administration, Supervision. **Xiao Sun:** Data curation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

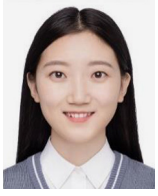
Acknowledgment

This work is supported by NSFC-Shenzhen Joint Foundation (Key Project) (Grant No. U1613217); The Key Program of the National Natural Science Foundation of China (Grant No. 61432004); The National Natural Science Foundation of China (Grant No. 61772169).

References

[1] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (2012) 42–55.
[2] S. Chu, S.S. Narayanan, C.C.J. Kuo, Environmental sound recognition using mp-based features, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 1–4.
[3] A.M. Al-kaysi, A. Al-Ani, C. Loo, T.Y. Powell, D. Martin, M. Breakspear, et al., Predicting tdcs treatment outcomes of patients with major depressive disorder using automated eeg classification, J. Affect. Disord. 208 (2017) 597–603.
[4] A.V. Bocharov, G.G. Knyazev, A.N. Savostyanov, Depression and implicit emotion processing: an eeg study, Neurophysiol. Clin./Clin. Neurophysiol. 47 (2017) 225–230.
[5] M. Soleymani, M. Pantic, T. Pun, Multimodal emotion recognition in response to videos, IEEE Trans. Affect. Comput. 3 (2) (2012) 211–223.
[6] W.L. Zheng, J.Y. Zhu, Y. Peng, B. Lu, Eeg-based emotion classification using deep belief networks, in: 2014 IEEE International Conference on Multimedia and Expo (ICME), 2014, pp. 1–6.
[7] D. Nie, X. Wang, L.C. Shi, B. Lu, Eeg-based emotion recognition during watching movies, in: 2011 5th International IEEE/EMBS Conference on Neural Engineering, 2011, pp. 667–670.
[8] A. Mert, A. Akan, Emotion recognition based on time-frequency distribution of eeg signals using multivariate synchrosqueezing transform, Digit. Signal Process. 81 (2018) 106–115.
[9] W. Wan-hui, Q. Yu-hui, L. Guang-yuan, Electrocardiography recording, feature extraction and classification for emotion recognition, in: 2009 WRI World Congress on Computer Science and Information Engineering, Vol. 4, 2009, pp. 168–172.
[10] Q. Chen, Y. Li, X. Yuan, A hybrid method for muscle artifact removal from eeg signals, J. Neurosci. Methods 353 (2021) 109104-1.
[11] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recognit. 84 (2018) 251–261.
[12] T. Chen, S. Ju, X. Yuan, M. Elhoseny, F. Ren, M. Fan, et al., Emotion recognition using empirical mode decomposition and approximation entropy, Comput. Electr. Eng. 72 (2018) 383–392.
[13] N. Birbaumer, Breaking the silence: brain-computer interfaces (bci) for communication and motor control, Psychophysiology 6 (2006) 517–532.
[14] C. Anderson, E. Stolz, S. Shamsunder, Multivariate autoregressive models for classification of spontaneous electroencephalographic signals during mental tasks, IEEE Trans. Biomed. Eng. 45 (1998) 277–286.
[15] T.P. Jung, S. Makeig, M. Stensmo, T.J. Sejnowski, Estimating alertness from the eeg power spectrum, IEEE Trans. Biomed. Eng. 44 (1) (1997) 60–69.
[16] F. Mormann, K. Lehnertz, P.R. David, C.E. Elger, Mean phase coherence as a measure for phase synchronization and its application to the eeg of epilepsy patients, Physica D (2000).
[17] J. Richman, J.R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol., Heart Circ. Physiol. 6 (2000) H2039–49.
[18] B. Krisnandhika, A. Faqih, P.D. Pumamasari, B. Kusumoputro, Emotion recognition system based on eeg signals using relative wavelet energy features and a modified radial basis function neural networks, in: 2017 International Conference on Consumer Electronics and Devices (ICCED), 2017, pp. 50–54.
[19] E.T. Pereira, H.M. Gomes, L.R. Veloso, M.R.A. Mota, Empirical evidence relating EEG signal duration to emotion classification performance, IEEE Trans. Affect. Comput. 12 (1) (2021) 154–164.
[20] H. Chao, Y. Liu, Emotion recognition from multi-channel eeg signals by exploiting the deep belief-conditional random field framework, IEEE Access 8 (2020) 33002–33012.
[21] W.L. Zheng, B. Lu, Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks, IEEE Trans. Auton. Ment. Dev. 7 (2015) 162–175.
[22] H. Ferdinando, T. Seppänen, E. Alasaarela, Enhancing emotion recognition from ecg signals using supervised dimensionality reduction, in: ICPRAM, 2017.
[23] Y.L. Hsu, J.S. Wang, W.C. Chiang, C.H. Hung, Automatic ecg-based emotion recognition in music listening, IEEE Trans. Affect. Comput. 11 (2020) 85–99.
[24] M.A. Islam, A. Hamza, M.H. Rahaman, J. Bhattacharjee, M.M. Rahman, Mind reader: a facial expression and eeg based emotion recognizer, in: 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 2018, pp. 101–107.
[25] Y.H. Kwon, S.B. Shin, S. Kim, Electroencephalography Based Fusion Two-Dimensional (2d)-Convolution Neural Networks (cnn) Model for Emotion Recognition System, Sensors, Basel, Switzerland, 2018, p. 18.
[26] S. Katsigiannis, N. Ramzan, Dreamer: a database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices, IEEE J. Biomed. Health Inform. 22 (1) (2018) 98–107.
[27] W. Zhao, Z. Zhao, C. Li, Discriminative-cca promoted by eeg signals for physiological-based emotion recognition, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), 2018, pp. 1–6.
[28] T. Chen, S. Ju, F. Ren, M. Fan, Y. Gu, Eeg emotion recognition model based on the libsvm classifier, Measurement (2020) 108047.
[29] A. Lempel, J. Ziv, On the complexity of finite sequences, IEEE Trans. Inf. Theory 22 (1) (1976) 75–81.
[30] M. Wu, Y. Wang, A feature selection algorithm of music genre classification based on relieff and sfs, in: 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), 2015.
[31] C.W.J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica 37 (3) (1969) 424–438.
[32] J. Pan, W.J. Tompkins, A real-time qrs detection algorithm, IEEE Trans. Biomed. Eng. BME-32 (3) (1985) 230–236.
[33] Y. Liu, T. Ye, Z. Zeng, Y. Zhang, G. Wang, N. Chen, et al., Generative adversarial network-enabled learning scheme for power grid vulnerability analysis, Int. J. Web Grid Serv. 17 (2) (2021) 138–151.
[34] A.B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 english lemmas, Behav. Res. Methods 45 (2013) 1191–1207.
[35] G. Zhang, M. Yu, Y. Liu, G. Zhao, D. Zhang, W. Zheng, SparseDGCNN: recognizing emotion from multichannel EEG signals, IEEE Trans. Affect. Comput. (2021) 1.

Tian Chen received the B.E., M.E., and Ph.D. in Computer Science and Technology from Hefei University of Technology (HFUT), China, in 1997, 2002, and 2011 respectively. Since 2010, she has been an Assistant Professor at the Hefei University of Technology. From 2017 to 2018, she was a Visiting Scholar at the University of North Texas, America. She is now working in Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine at the Hefei University of Technology. Her current research interests include Artificial Intelligence, Affective Computing, and Wearable Computing. She is a member of IEEE and a senior member of CCF.



Hongfang Yin was born in 1995. She received a B.E. degree in Computer Science and Technology from Qufu Normal University, China, in 2018. Since 2018, she has been a postgraduate student of Hefei University of Technology, China. Her research interests include Affective Computing and Wearable Computing.

Xiaohui Yuan received a B.S. degree in Electrical Engineering from the Hefei University of Technology, China in 1996 and a Ph.D. degree in Computer Science from the Tulane University in 2004. He is currently an Associate Professor at the University of North Texas. His research interests include computer vision, artificial intelligence, data mining, and machine learning. His research findings have been published in more than 180 peer-reviewed papers. He is the editor-in-chief of the International Journal of Smart Sensor Technologies and Applications, serves on the editorial board of several international journals, and as chairs in several international conferences. He is a recipient of the Ralph E. Powe Junior Faculty Enhancement Award in 2008.

Fuji Ren received his Ph. D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to 1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of the Engineering Academy of

Japan and the EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societies, a Fellow of IEICE, a Fellow of CAAI. He is the President of the International Advanced Information Institute, Japan.

Yu Gu received the B.E. degree from the Special Classes for the Gifted Young, University of Science and Technology of China, Hefei, China, in 2004, and the D.E. degree from the University of Science and Technology of China in 2010. In 2006, he was an Intern with Microsoft Research Asia, Beijing, China, for seven months. From 2007 to 2008, he was a Visiting Scholar with the University of Tsukuba, Tsukuba, Japan. From 2010 to 2012, he was a JSPS Research Fellow with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor and the Dean's Assistant with the School of Computer and Information, Hefei University of Technology, Hefei. His current research interests include pervasive computing and affective computing. Dr. Gu is also a member of the Association for Computing Machinery (ACM).

Xiao Sun was born in 1980. He received the M.E. degree in 2004 from the Department of Computer Sciences and Engineering at Dalian University of Technology and got his dual doctor's degree in Dalian University of Technology (2010) of China and the University of Tokushima (2009) of Japan. He is now working as a professor in Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine at the Hefei University of Technology. His research interests include Affective Computing, Natural Language Processing, Machine Learning, and Human-Machine Interaction.