EEG Feature Selection via Global Redundancy Minimization for Emotion Recognition

Xueyuan Xu, *Student Member, IEEE*, Tianyuan Jia, Qing Li, Fulin Wei, Long Ye, and Xia Wu, *Member, IEEE*

Abstract—A common drawback of EEG-based emotion recognition is that volume conduction effects of the human head introduce interchannel dependence and result in highly correlated information among most EEG features. These highly correlated EEG features cannot provide extra useful information, and they actually reduce the performance of emotion recognition. However, the existing feature selection methods, commonly used to remove redundant EEG features for emotion recognition, ignore the correlation between the EEG features or utilize a greedy strategy to evaluate the interdependence, which leads to the algorithms retaining the correlated and redundant features with similar feature scores in the EEG feature subset. To solve this problem, we propose a novel EEG feature selection method for emotion recognition, termed global redundancy minimization in orthogonal regression (GRMOR). GRMOR can effectively evaluate the dependence among all EEG features from a global view and then select a discriminative and nonredundant EEG feature subset for emotion recognition. To verify the performance of GRMOR, we utilized three EEG emotional data sets (DEAP, SEED, and HDED) with different numbers of channels (32, 62, and 128). The experimental results demonstrate that GRMOR is a promising tool for redundant feature removal and informative feature selection from highly correlated EEG features.

Index Terms—EEG, feature selection, emotion recognition, global redundancy minimization, orthogonal regression

1 Introduction

Due to its merits of high temporal resolution, portability, cost-effectiveness and nontrauma, the electroencephalogram (EEG) has received increasing attention in affective computing [1], [2], [3], [4], [5]. To represent the affective state, a wide variety of EEG features are extracted from EEG recordings in the time domain [6], [7], [8], [9], frequency domain [3], [10], [11], and time-frequency domain [4], [12], [13].

In recent years, the number of EEG sensors has continued to increase, and these sensors can provide rich spatial information for emotion recognition tasks [14], [15]. Multichannel EEG signals can be recorded by the abundant EEG electrodes, and then high-dimensional EEG feature vectors are increasingly generated. However, due to the small sample size of EEG-based emotion recognition data [16], [17], high-dimensional EEG features can easily lead to overfitting problems with the classifiers and poor emotion recognition performance.

To remove the redundant features from the original EEG features and prevent the overfitting of emotion recognition classifiers, a variety of feature selection methods have been implemented, such as ReliefF [18], information gain [19], maximum relevance minimum redundancy (mRMR) [20], conditional mutual information maximization (CMIM) [21], feature selection with orthogonal regression (FSOR) [22], etc. Depending on how they are integrated with emotion classification models, the feature selection methods commonly used for EEG-based emotion recognition can be divided into three categories: filter-based, wrapper-based, and embedded-based methods [23], [24]. The filter-based methods rank the EEG features by computing the correlation between the EEG features and corresponding emotional labels; however, this process ignores the learning model, and the selected EEG feature subset may not be optimal [25]. To address the issues with the filter-based methods, two-step wrapper-based methods were proposed [26]. Wrapper methods first select an EEG feature subset by random search or sequential search and then employ a learning or predictive model to evaluate the emotion recognition performance of the EEG feature subset [27], [28]. Nevertheless, one major shortcoming of the wrapper-based approaches is that their computational cost is much higher than that of each of the filter approaches[26].

Recently, as an alternative strategy for solving the problems of the filter-based methods, Xu *et al*. proposed a novel embedded-based method to select a small EEG feature subset that could provide a compact representation of the high-dimensional EEG features [22]. The FSOR method incorporates the EEG feature selection problem and the emotion classification problem into an objective function. Unlike other EEG feature selection methods, for nonlinear and nonstationary EEG signals, the method utilizes orthogonal regression to retain more discriminative information in the orthogonal subspace [22], [24], [29], [30]. Additionally, the

- *Xueyuan Xu, Tianyuan Jia, Qing Li, Fulin Wei, and Xia Wu are with the School of Artificial Intelligence, Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing Normal University, Beijing 100875, China. E-mail: {xuxueyuan, 201921210018, liqing_lq, weifulin}@mail.bnu.edu.cn, wuxia@bnu.edu.cn.*
- *Long Ye is with the State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China. E-mail: yelong@cuc.edu.cn.*

Manuscript received 30 Sept. 2020; revised 15 Mar. 2021; accepted 21 Mar. 2021. Date of publication 24 Mar. 2021; date of current version 28 Feb. 2023. (Corresponding author: Xia Wu.)
Recommended for acceptance by D. Zhang.
Digital Object Identifier no. 10.1109/TAFFC.2021.3068496

method employs a feature weight vector to accurately rank the importance of EEG features in the emotion recognition task [22], [24]. The experimental results on a 19-channel EEG data set demonstrated that FSOR performs better than ReliefF and information gain in EEG feature selection for emotion recognition [22].

However, due to the volume conduction effects in the human head [31], [32], the measurement at each EEG sensor is the linear combination of several underlying source signals [33], [34], [35], which introduces interchannel dependence and redundancy [36], [37], [38]. In other words, the EEG signals recorded at the adjacent electrodes in the topographical map may contain plenty of repeated or similar information [39], [40], which results in highly correlated EEG features extracted from such electrodes [41], [42]. As the number of electrodes increases, the issue becomes more serious because the electrodes on the scalp are more densely distributed [43], [44], [45], [46].

For the abovementioned feature selection methods, the feature rankings of highly correlated features tend to be almost equal because they are considered equally important for the classification task [26], [47]. As a result, many EEG features with top scores in the selected feature subset are often dependent on and redundant with each other [48]. Since redundant EEG features cannot provide additional useful information for emotion recognition, one of them should be rejected [30]. Nevertheless, popular feature selection methods mostly choose EEG feature subsets regardless of the correlation within them [47]. The mRMR [20] and CMIM methods [49] can minimize the redundant information in the selected feature subset to a certain extent. However, they utilize a greedy strategy to evaluate the feature dependence; thus, global redundancy minimization cannot be achieved, and the selected EEG feature subset is not an optimal output [47], [48].

To solve this problem, we propose a novel EEG feature selection method for emotion recognition, named global redundancy minimization in orthogonal regression (GRMOR). We introduce a global redundancy evaluation term to compute the sum of the redundancies of an EEG feature with other features. The proposed method can effectively recognize the redundant EEG features from a global view and then obtain a discriminative and nonredundant EEG feature subset for emotion recognition.

Moreover, the main contributions of our work are as follows:

- A novel EEG feature selection method for emotion recognition is proposed. The GRMOR method introduces global redundancy information into the orthogonal regression. The proposed method implements orthogonal regression to keep more statistical and structural information of EEG data in the orthogonal subspace. Additionally, from a global view of redundant information among the EEG features, our method can effectively evaluate the dependence among the EEG features and select a discriminative and nonredundant EEG feature subset for emotion recognition.
- Two popular low-density EEG data sets (DEAP and SEED) are first utilized to verify the effectiveness of GRMOR. Additionally, we collected a 128-channel EEG data set to further examine the performance of GRMOR on the high-density EEG. The experimental results on the three EEG data sets verify that the proposed GRMOR method has superior performance to those of nine popular feature selection methods.
- We further analyze the electrode locations of selected features of GRMOR. To all the three data sets, the CP5/50th electrode in the left temporal region can be observed, which indicates that this electrode is a critical electrode for emotion recognition.

The remainder of the paper is organized as follows. Section 2 gives the notations and definitions of norms used in this paper, and briefly reviews related research works. In Section 3, the proposed EEG feature selection framework for emotion recognition is introduced in detail. Then, an optimization strategy of the proposed objective function is illustrated in Section 4. The experimental data and settings are introduced in Section 5. We discuss the experimental results in Section 6. Section 7 concludes this paper.

2 Notations and Related Works

2.1 Notations and Definitions

The definitions of the norms and notations used throughout the paper are summarized here. All vectors and matrices are denoted by lowercase boldface letters ($\mathbf{a}$, $\mathbf{b}$ ...) and capital letters (A, B ...), respectively. The transpose of a vector or matrix is denoted by an uppercase superscript T. The operator $\circ$ is the Hadamard product. The trace of any matrix S is denoted by $tr(S)$. The Frobenius norm of any matrix S is denoted by

$$\|S\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij}^2} = \sqrt{tr(S^T S)}. \tag{1}$$

$X = [\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_d]^T \in \mathbb{R}^{\mathrm{d}\times n}$ is the emotional feature data extracted from EEG recordings and $\mathbf{x}_d \in \mathbb{R}^{1\times n}$. $Y \in \mathbb{R}^{k\times n}$ is the corresponding emotional label matrix. d, n, and k represent the numbers of features, samples, and classes, respectively. $\mathbf{1}_n = (1, 1 \ldots, 1)^T \in \mathbb{R}^{n\times 1}$ is defined as a column vector of all ones. I_n is defined as an $n \times n$ identity matrix. Notations are summarized in Table 1.

2.2 Related Works

According to the relationship with classification models, the EEG feature selection approaches can be classified into three types: filter-based, wrapper-based, and embedded-based methods [25], [26], [50].

Filter-based approaches evaluate the importance of EEG features in emotion recognition tasks according to certain criteria and the discriminative EEG features with high rankings are then selected, including, but not limited to, ReliefF [18], information gain [19], Cohen's effect size f^2 [51], one-way analysis of variance (ANOVA) [52], [53], mRMR [20], [54], CMIM [21], etc. Nevertheless, due to the neglect of learning models, the above filter-based approaches may abandon several useful features that are discriminative when combined with other features [26], [50].

Wrapper-based methods 'wrap' features into candidate EEG feature subsets by random search or sequential search, and then evaluate the performance of these candidate EEG feature subsets by a learning or predictive model. Some

TABLE 1
Notations

Notation	Definition
d	The number of features
n	The number of samples
k	The number of classes
λ	The balance parameter
$\mathbf{b} \in \mathbb{R}^{k\times 1}$	A bias vector
$\boldsymbol{\theta}$	A feature score vector
$\mathbf{x}_d \in \mathbb{R}^{1\times n}$	The dth EEG feature
$\mathbf{1}_n = (1, 1, \ldots, 1)^T$	A row vector of all ones
$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d]^T$	The EEG feature data matrix
$Y \in \mathbb{R}^{k\times n}$	The emotional label matrix
$I_n \in \mathbb{R}^{n\times n}$	An $n \times n$ identity matrix
$W \in \mathbb{R}^{d\times k}$	An orthogonal matrix
$\Theta \in \mathbb{R}^{d\times d}$	A diagonal matrix
$A \in \mathbb{R}^{d\times d}$	An EEG feature redundancy matrix
$\|.\|_F$	The Frobenius norm of a matrix
vec	The vectorization of a matrix
$tr(.)$	The trace of a square matrix

well-known wrapper-based methods have been proposed or utilized on EEG-based emotion recognition, such as evolutionary computation (EC) algorithms [42] and ReliefF based genetic algorithm (REGA)[55]. However, due to the iterative process of feature subset searching, the computational complexities of the wrapper-based approaches are usually high [50].

Embedded-based methods, which were proposed to overcome the shortages of filter-based or wrapper-based feature selection methods [56], incorporate the process of feature selection into the process of training models and directly evaluate the importance of each EEG feature in emotion classification tasks during the optimization of classification models. Least square regression is a widely-used statistical model for embedded-based feature selection methods [57], [58], [59].

The traditional least square regression is to minimize the following objective function:

$$\min_{W,b} \left\| X^T W + \mathbf{1}_n \mathbf{b}^T - Y \right\|_F^2. \tag{2}$$

The aim of least square regression is to learn a projection matrix W and a bias vector $\mathbf{b}$ to estimate the label matrix Y. To perform feature selection, a row sparse constraint is introduced to the projection matrix W in [59]. The problem can be reformulated as

$$\min_{W,b} \left\| X^T W + \mathbf{1}_n \mathbf{b}^T - Y \right\|_F^2 s.t. \|W\|_{2,0} = k. \tag{3}$$

A sparse selection subspace W can be obtained and the pivotal features could be selected with only k non-zero rows in the projection matrix W.

However, the l_2-norm based loss function is sensitive to outliers. To solve the problem, a $l_{2,1}$-norm based least square regression method was proposed in [57], called robust feature selection (RFS). The framework of RFS is represented as

$$\min_{W} \left\| X^T W - Y \right\|_{2,1} + \gamma \|W\|_{2,1}. \tag{4}$$

The RFS method is robust to the outliers by employing joint $l_{2,1}$-norm minimization on both loss function and regularization.

However, the least square regression neglects the local structure information (dependent or independent) of data due to the simplicity of the model [60], [61]. Because the least square regression assumes that the variables in the measurement error are all dependent, it would not obtain an effective result when input variables contain independent variables [62]. In order to overcome the drawback, orthogonal constraint ($W^T W = I_k$) was introduced to the project matrix W in the least square regression model. Instead of minimizing the horizontal distance in the least square regression, the aim of orthogonal regression is to minimize the perpendicular distance from the data points to the regression line [63]. Through the above different distance calculation approach, the orthogonal regression method could consider both dependent and independent variables in the measurement error [64]. By utilizing orthogonal constraint, orthogonal regression can be formulated as

$$\min_{W^T W = I_k, \mathbf{b}} \left\| X^T W + \mathbf{1}_n \mathbf{b}^T - Y \right\|_F^2. \tag{5}$$

To sum up, compared with least square regression, orthogonal regression could retain more local statistical and structure information of data. In [30], a $l_{2,1}$-norm regularized orthogonal regression model was proposed to select relevant features. Its formulation is presented as follows:

$$\min_{W^T W = I_k, \mathbf{b}, \gamma} \left\| X^T (\gamma W) + \mathbf{1}_n \mathbf{b}^T - Y \right\|_F^2, \tag{6}$$

where γ is introduced to the orthogonal regression to address the defect concerning lack of scale change.

To accurately evaluate the importance of features in classification tasks and preserve the statistical structure of the data, we proposed FSOR method which introduces feature weighting matrix into orthogonal regression model [24]. FSOR is to minimize the following objective problem:

$$\begin{aligned} &\min_{W, \mathbf{b}, \theta} \left\| W^T \Theta X + \mathbf{b} \mathbf{1}_n^T - Y \right\|_F^2 \\ &\text{s.t. } W^T W = I_k, \boldsymbol{\theta}^T \mathbf{1}_d = 1, \boldsymbol{\theta} \geq 0, \end{aligned} \tag{7}$$

where $\Theta \in \mathbb{R}^{d\times d}$ with $\boldsymbol{\theta}^T \mathbf{1}_d = 1$ ($\boldsymbol{\theta} \geq 0$) is a diagonal matrix.

Nevertheless, for the above least square regression and orthogonal regression based feature selection methods, the feature scores of highly correlated EEG features tend to be almost equal and are retained in the selected EEG feature subset because they are considered equally important in emotion classification [26], [47].

3 The Proposed Framework

We aim to build a novel EEG emotional feature selection model in which the importance of each EEG feature in emotion recognition can be accurately scored and the redundant information between pairwise EEG features can be evaluated from a global view.

Unlike many other feature selection methods commonly used for emotion recognition, GRMOR utilizes orthogonal regression to keep more statistical and structural information

in the subspace and the global redundancy matrix to analyze the correlation between pairwise EEG features. Additionally, the weight of each EEG feature can be learned through the balance between its importance in emotion recognition and its redundancy with other EEG features.

3.1 The Definition of the Global Redundancy Matrix

A global redundancy matrix is introduced to evaluate the correlation between pairwise EEG features. The elements of the global redundancy matrix A can be calculated by

$$A_{i,j} = \left(B_{i,j}\right)^2 = \left(\frac{\mathbf{f}_i^T \mathbf{f}_j}{\|\mathbf{f}_i\| \|\mathbf{f}_j\|}\right)^2, \tag{8}$$

where $\mathbf{f}_i \in \mathbb{R}^{n\times 1}$ and $\mathbf{f}_j \in \mathbb{R}^{n\times 1}$ are column vectors containing the ith and jth centralized EEG features of $\mathbf{x}_i$ and $\mathbf{x}_j$ $(i, j = 1, 2 \ldots, d)$, respectively. The above centralized EEG features can be computed by the following equations:

$$\begin{cases} \mathbf{f}_i = H\mathbf{x}_i^T \\ \mathbf{f}_j = H\mathbf{x}_j^T \end{cases}, \tag{9}$$

where $H = I_{\mathrm{n}} - \frac{1}{n}\mathbf{1}_n \mathbf{1}_n{}^T$ [65] is the centering matrix.

Obviously, $A = B \circ B$. Hence, we can modify Eq. (8) and obtain the following form:

$$B = DF^T FD = (FD)^T FD, \tag{10}$$

where $F = [\mathbf{f}_1, \mathbf{f}_2 \ldots, \mathbf{f}_d]$ is a centralized EEG feature matrix. D is a diagonal matrix, and the elements of D can be obtained by $D_{i,i} = \frac{1}{\|\mathbf{f}_i\|}$ $(i = 1, 2, \ldots, d)$. The matrix B has been proven to be a positive semidefinite matrix [48].

3.2 Problem Formulation

To overcome the shortage of embedded feature selection methods, a novel supervised EEG feature selection method for emotion recognition is proposed. The GRMOR method introduces global redundancy information into the orthogonal regression to accurately evaluate the redundant information between the highly correlated EEG features. GRMOR is to minimize the following objective problem:

$$\begin{aligned} &\min_{W,\mathbf{b},\theta} \left\|W^T \Theta X + \mathbf{b}\mathbf{1}_n^T - Y\right\|_F^2 + \lambda \boldsymbol{\theta}^T A \boldsymbol{\theta} \\ &\text{s.t. } W^T W = I_k, \boldsymbol{\theta}^T \mathbf{1}_d = 1, \boldsymbol{\theta} \geq 0, \end{aligned} \tag{11}$$

where $W \in \mathbb{R}^{d\times k}$ with the orthogonal constraint $W^T W = I_k$ is the projection matrix, and $\Theta \in \mathbb{R}^{d\times d}$ with $\boldsymbol{\theta}^T \mathbf{1}_d = 1$ $(\boldsymbol{\theta} \geq 0)$ is a diagonal matrix. It should be noted that Θ is a diagonal matrix with feature score vector $\boldsymbol{\theta}$ in the diagonal. $\mathbf{b} \in \mathbb{R}^{k\times 1}$ is a bias vector, and $A \in \mathbb{R}^{d\times d}$ is the feature redundancy matrix. The elements of the matrix A are computed by the cosine similarity-based correlations between pairwise EEG features.

Since the matrix B is positive semidefinite and $A = B \circ B$, the global redundancy matrix A is nonnegative and positive semidefinite. The detailed proof can be found in [48], [66]. Since the global redundancy matrix A is positive semidefinite, the objective function of GRMOR is a convex problem.

The optimization problem involves two components: the error-minimizing part and the global redundancy-minimizing part. In Eq. (11), the term $\boldsymbol{\theta}^T A \boldsymbol{\theta}$ is utilized to evaluate the global feature redundancy. λ $(\lambda > 0)$ is a tuning parameter used to balance the error-minimizing term and the global redundancy-minimizing term. By optimizing the objective function of GRMOR, the square error and the global redundancy are minimized simultaneously, and then the vector $\boldsymbol{\theta}$ is obtained by extraction from the diagonal matrix Θ. The elements of θ can be utilized to rank the weights of all the EEG features.

3.3 Why the Proposed GRMOR Framework Works

To illustrate the mechanism by which the GRMOR method can minimize the redundancy from a global perspective, the global redundancy minimizing term $\boldsymbol{\theta}^T A \boldsymbol{\theta}$ in Eq. (11) can be transformed into $\sum_{i,j=1}^{d} A_{i,j}\boldsymbol{\theta}_i \boldsymbol{\theta}_j$. The GRMOR method first employs an error-minimizing term based on orthogonal regression to score the weights of EEG features $\mathbf{x}_i$ and $\mathbf{x}_j$. A large value of $A_{i,j}$ represents that the EEG features $\mathbf{x}_i$ and $\mathbf{x}_j$ are highly correlated. When $\boldsymbol{\theta}_i > \boldsymbol{\theta}_j$, $\boldsymbol{\theta}_j$ will be small to minimize the global redundancy term. In other words, if the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ are redundant with each other, the value of $\boldsymbol{\theta}_j$ decreases and that of $\boldsymbol{\theta}_i$ is retained in the process of iterative optimization of the objective function.

4 Optimization

In this section, based on the GPI and ALM methods, we exploit an efficient, alternative optimizing algorithm to solve the objective function of GRMOR.

4.1 Generalized Power Iteration

It is extremely difficult to acquire the closed-form solution for orthogonal least squares regression. To address this issue, orthogonal regression is usually transformed into the quadratic problem on the Stiefel manifold (QPSM) [67]

$$\min_{W^T W = I_k} \mathrm{Tr}\left(W^T JW - 2W^T M\right), \tag{12}$$

where $M \in \mathbb{R}^{d\times k}$ and the symmetric matrix $J \in \mathbb{R}^{d\times d}$.

To address the QPSM, in this paper, we employ the GPI method proposed by Nie *et al.* in [68] to update the orthogonal matrix W. The GPI algorithm is summarized in Algorithm 1.

Algorithm 1. Generalized Power Iteration (GPI) Method

Input: the symmetric matrix $J \in \mathbb{R}^{d\times d}$ and matrix $M \in \mathbb{R}^{d\times k}$.
Output: the matrix $W \in \mathbb{R}^{d\times k}$.
1: Initialize the random W and the parameter α such that $\widetilde{J} = \alpha I_d - J \in \mathbb{R}^{d\times d}$ is a positive definite matrix.
2: **repeat**
3: Update Z by $Z = 2\widetilde{J}W + 2M$;
4: Calculate $USV^T = Z$ via the compact singular value decomposition method on Z;
5: Update W by $W = UV^T$;
6: **until** convergence

4.2 General Augmented Lagrangian Multiplier

The ALM method is proposed to solve the constraint optimization problem in the following form:

$$\min_{\varphi(X)=0} f(X). \tag{13}$$

The full process of the ALM method is summarized in Algorithm 2.

Algorithm 2. The Augmented Lagrangian Multiplier (ALM) Method

Input: $1 < p < 2, u > 0, \boldsymbol{\delta}$.
Output: X
1: **repeat**
2: Update X by $\min_X f(X) + \frac{\mu}{2}\left\|\varphi(X) + \frac{1}{\mu}\boldsymbol{\delta}\right\|_F^2$;
3: Update $\boldsymbol{\delta}$ by $\boldsymbol{\delta} = \boldsymbol{\delta} + \mu\varphi(X)$;
4: Update μ by $\mu = p\mu$;
5: **until** convergence

4.3 Optimization Algorithm

We first take the derivative of the objective function in Eq. (11) with respect to b and then set the derivative to zero.

$$\frac{\partial\left(\left\|W^T\Theta X + \mathbf{b}\mathbf{1}_n^T - Y\right\|_F^2 + \lambda\boldsymbol{\theta}^T A\boldsymbol{\theta}\right)}{\partial\mathbf{b}} = 0. \tag{14}$$

We obtain

$$\mathbf{b} = \frac{1}{n}\left(Y\mathbf{1}_n - W^T\Theta X\mathbf{1}_n\right). \tag{15}$$

Substituting the derived result from Eq. (15) into Eq. (11), we can rewrite Eq. (11) as

$$\begin{aligned}&\min_{W,\boldsymbol{\theta}}\left\|W^T\Theta XH - YH\right\|_F^2 + \lambda\boldsymbol{\theta}^T A\boldsymbol{\theta}\\ &\text{s.t. } W^TW = I_k, \boldsymbol{\theta}^T\mathbf{1}_d = 1, \boldsymbol{\theta} \geq 0.\end{aligned} \tag{16}$$

Then, we apply an alternative optimization approach to find a solution to the proposed problem in Eq. (16). The optimization of GRMOR is further divided into the following two subproblems.

4.3.1 Optimization of W by Fixing Θ

When Θ is fixed, the objective function of GRMOR can be reformulated as

$$\min_{W^TW = I_k} \mathrm{Tr}\left(W^TJW - 2W^TM\right). \tag{17}$$

Here, $J = \Theta XHX^T\Theta^T$ and $M = \Theta XHY^T$. Obviously, as in Section 4.1, Eq. (17) is a QPSM problem.

In this paper, the optimization problem in Eq. (17) is approached by using the GPI method mentioned in Section 4.1. The solution of W can be obtained until the stopping criteria in Algorithm 1 are satisfied.

4.3.2 Optimization of Θ by Fixing W

When W is fixed, the objective function of GRMOR becomes

$$\begin{aligned}&\min_{W,\boldsymbol{\theta}}\left[\mathrm{Tr}\left(\Theta XHX^T\Theta WW^T\right) - \mathrm{Tr}\left(2\Theta XHY^TW^T\right)\right.\\ &\quad\left.+\lambda\boldsymbol{\theta}^TA\boldsymbol{\theta}\right]\\ &\quad\text{s.t. } W^TW = I_k, \boldsymbol{\theta}^T\mathbf{1}_d = 1, \boldsymbol{\theta} \geq 0.\end{aligned} \tag{18}$$

Lemma 1. *For a diagonal matrix S,* $\mathrm{Tr}(\mathrm{SCSG}) = \mathbf{s}^T\left(C^T \circ G\right)\mathbf{s}$

Proof.

$$\begin{aligned}\mathrm{Tr}(\mathrm{SCSG}) &= \mathbf{s}^T\mathrm{diag}(\mathrm{CSG})\\ &= \mathbf{s}^T\mathrm{vec}\left\{\mathbf{c}_i^TS\mathbf{g}_i\right\}\\ &= \mathbf{s}^T\mathrm{vec}\left\{\left(\mathbf{c}_i \circ \mathbf{g}_i\right)^T\mathbf{s}\right\}\\ &= \mathbf{s}^T\left(C^T \circ G\right)^T\mathbf{s}\\ &= \mathbf{s}^T\left(C^T \circ G\right)\mathbf{s}.\end{aligned}$$

It should be noted that the symbol vec is the vectorization step of a matrix. □

Based on Lemma 1, the optimization problem in Eq. (18) is equivalent to the following problem:

$$\begin{aligned}&\min_{W,\boldsymbol{\theta}}\left[\boldsymbol{\theta}^T\left[\left(XHX^T\right)^T \circ \left(WW^T\right) + \lambda A\right]\boldsymbol{\theta} - \boldsymbol{\theta}^Ts\right]\\ &\text{s.t. } W^TW = I_k, \boldsymbol{\theta}^T\mathbf{1}_d = 1, \boldsymbol{\theta} \geq 0.\end{aligned} \tag{19}$$

Eq. (19) can be further reformulated as

$$\min_{\boldsymbol{\theta}^T\mathbf{1}_d = 1, \boldsymbol{\theta} \geq 0} \boldsymbol{\theta}^TQ\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{s}, \tag{20}$$

where

$$\begin{cases} Q = \left(XH^TX^T\right) \circ \left(WW^T\right) + \lambda A\\ \mathbf{s} = diag\left(2XHY^TW^T\right)\end{cases}.$$

Therefore, the optimization problem in Eq. (18) is converted to Eq. (20). To tackle the constrained optimization problem in Eq. (20), we utilize the ALM method and further decompose Eq. (20) into the following subproblems.

We can rewrite Eq. (20) in a new form

$$\min_{\boldsymbol{\theta}^T\mathbf{1}_d = 1, \mathbf{v} \geq 0, \mathbf{v} = \boldsymbol{\theta}} \boldsymbol{\theta}^TQ\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{s}. \tag{21}$$

The augmented Lagrangian function in Eq. (20) is written as follows:

$$\begin{aligned}L(\boldsymbol{\theta}, \mathbf{v}, \mu, \boldsymbol{\delta}_1, \delta_2) &= \boldsymbol{\theta}^TQ\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{s} + \frac{\mu}{2}\left\|\boldsymbol{\theta} - \mathbf{v} + \frac{1}{\mu}\boldsymbol{\delta}_1\right\|_F^2\\ &+ \frac{\mu}{2}\left(\boldsymbol{\theta}^T\mathbf{1}_d - 1 + \frac{1}{\mu}\delta_2\right)^2 \text{ s.t. } \mathbf{v} \geq 0.\end{aligned} \tag{22}$$

Here, the parameters $\mathbf{v}$ and $\boldsymbol{\delta}_1$ are both column vectors, and μ is the Lagrangian multiplier. When $\mathbf{v}$ is fixed, the abovementioned equation is equivalent to

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\boldsymbol{\theta}^TE\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{f}, \tag{23}$$

where

$$\begin{cases} E = 2Q + \mu I_d + \mu\mathbf{1}_d\mathbf{1}_d^T\\ \mathbf{f} = \mu\mathbf{v} + \mu\mathbf{1}_d - \delta_2\mathbf{1}_d - \boldsymbol{\delta}_1 + \mathbf{s}\end{cases}. \tag{24}$$

Accordingly, the optimal solution of θ can be found by $\hat{\theta} = E^{-1}\mathbf{f}$.

When θ is fixed, Eq. (22) is transformed into

$$\min_{\mathbf{v}\geq 0}\left\|\mathbf{v}-\left(\boldsymbol{\theta}+\frac{1}{\mu}\boldsymbol{\delta}_1\right)\right\|^2. \tag{25}$$

By solving Eq. (25), we can obtain that the optimal solution $\mathbf{v}$ is defined as

$$\hat{\mathbf{v}}=\mathrm{pos}\left(\hat{\boldsymbol{\theta}}+\frac{1}{\mu}\boldsymbol{\delta}_1\right), \tag{26}$$

where the function $pos(t)$ converts each negative element in a vector t to 0. With all the subproblems solved, the detailed procedure for solving Eq. (23) is summarized in Algorithm 3.

Algorithm 3. Algorithm to Solve Eq. (23)

Input: $p>1$, $\boldsymbol{\theta}_i=\frac{1}{d}(1\leq i\leq d)$, $\mathbf{v}=\boldsymbol{\theta}$, $\delta_2=0$, $u>0$, $\boldsymbol{\delta}_1=(0,0,\ldots,0)^T\in\mathbb{R}^{d\times 1}$.
Output: $\widehat{\theta}$.
1: **repeat**
2: Update E by $E=2Q+\mu I_d+\mu\mathbf{1}_d\mathbf{1}_d^T$;
3: Update $\mathbf{f}$ by $\mathbf{f}=\mu\mathbf{v}+\mu\mathbf{1}_d-\delta_2\mathbf{1}_d-\boldsymbol{\delta}_1+\mathbf{s}$;
4: Update $\widehat{\boldsymbol{\theta}}$ by $\widehat{\boldsymbol{\theta}}=E^{-1}\mathbf{f}$;
5: Update $\mathbf{v}$ by $\hat{\mathbf{v}}=\mathrm{pos}(\hat{\boldsymbol{\theta}}+\frac{1}{\mu}\boldsymbol{\delta}_1)$;
6: Update $\boldsymbol{\delta}_1$ by $\boldsymbol{\delta}_1=\boldsymbol{\delta}_1+\mu(\widehat{\boldsymbol{\theta}}-\mathbf{v})$;
7: Update δ_2 by $\delta_2=\delta_2+\mu\left(\theta^T\mathbf{1}_d-1\right)$;
8: Update μ by $\mu=p\mu$;
9: **until** convergence

The whole process for solving the proposed GRMOR problem in Eq. (11) is depicted as Algorithm 4. In the Algorithm 4, the regression matrix W and diagonal matrix Θ are optimized alternatively until the value of the objective function is stable. The feature score vector $\boldsymbol{\theta}$ is extracted from the eventually obtained Θ. According to their rankings in $\boldsymbol{\theta}$, the EEG emotional features are sorted. Then, the m discriminative and mutually nonredundant EEG emotional features are chosen with the top rankings in $\boldsymbol{\theta}$.

Algorithm 4. Global Redundancy Minimization in Orthogonal Regression(GRMOR)

Input: the EEG data matrix $X\in\mathbb{R}^{\mathrm{d}\times n}$, the multidimensional emotional label matrix $Y\in\mathbb{R}^{k\times n}$.
Output: the regression matrix $W\in\mathbb{R}^{d\times k}$, the diagonal matrix $\Theta\in\mathbb{R}^{d\times d}$.
1: Initialize $\Theta\in\mathbb{R}^{d\times d}$ satisfying $\boldsymbol{\theta}^T\mathbf{1}_d=1$, where $\boldsymbol{\theta}\geq 0$. $H=I_{\mathrm{n}}-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n{}^T$.
2: **repeat**
3: Update W via Algorithm 1.
4: Update Θ via Algorithm 3.
5: **until** convergence

5 Experimental Details

In this section, we illustrate the specific information regarding the experimental data sets, feature extraction, experimental setup and evaluation metrics.

5.1 Data Set Description

In this study, three EEG emotional data sets, including two low-density EEG data sets (SEED [69] and DEAP [70]) and a high-density EEG data set (HDED), are adopted to investigate the effectiveness of the proposed GRMOR method. The details of these data are shown in Table 2.

TABLE 2
Comparisons Among the Three EEG Data Sets

Data set	DEAP	SEED	HDED
Channel no.	32	62	128
Subject no.	32	15	16
Video no.	40	15	12
Session no.	1	3	1
Sample no.	1,280	675	192
Stimulus materials	music videos	film clips	film clips

5.1.1 Low-Density EEG

The DEAP and SEED data sets are two publicly available EEG data sets for emotion recognition. The DEAP and SEED data sets were collected via 32 and 62 EEG channels, respectively, and both used video stimuli to elicit emotions. The tiny difference in emotional stimuli between the two data sets is that the DEAP data set utilized one minute of music video and the SEED data set utilized four minutes of Chinese film clips. The experimental setups for obtaining the two data sets are described in detail in [70] and [69].

5.1.2 High-Density EEG

Most EEG-based emotion recognition studies utilized low-density EEG (less than 128 channels) data, such as the SEED and DEAP databases. The disadvantage of a low-density EEG is its poor spatial resolution. With the development of EEG devices, the concept of high-density EEGs (exceeding 128 channels) has received an increasing amount of attention in emotion recognition due to its high spatial resolution. Compared with a low-density EEG, a high-density EEG can provide richer spatial information for the emotion recognition task.

To further examine the effectiveness of the GRMOR method with high-density EEG data, we collected a high-density EEG data set (HDED) via a 128-channel EEG device. The experiment was approved by the Beijing Normal University Ethics Board, and each subject signed a consent form prior to the experiment. The 128-channel EEG signals were recorded from sixteen healthy volunteers (8 males and 8 females, mean age: 22.27) using a 128-channel HydroCel Geodesic Sensor Net and an Amps400 amplifier (EGI, Eugene, Oregon, USA) with a sampling frequency of 250 Hz. The Cz channel was adopted as an initial reference.

Twelve Chinese film clips were chosen as emotional stimulus materials for inducing three emotional reactions: negative, positive, and neutral reactions. The affective tags of these clips were determined based on subjective ratings by the other twenty participants. To avoid making subjects fatigued and to produce effective and precise results, each media stimulus lasted for a period of time between 3 to 5 minutes. The details of the twelve Chinese film clips are shown in Table 3.

In the experiment, the participants were instructed to sit relaxed in a comfortable fashion and maintain a long attention span. Each subject watched 12 film clips in turn. The 12

TABLE 3
Details of the Film Clips Used in the Experiments

Film	Video emotion tag	Clips
Goodbye Mr. Loser	positive	1
The Mermaid	positive	1
City of Rock	positive	1
The Eagle Shooting Heroes	positive	1
Aftershock	negative	2
Dearest	negative	1
I Belonged to You	negative	1
Mount Tai	neutral	3
World Heritage in China	neutral	1

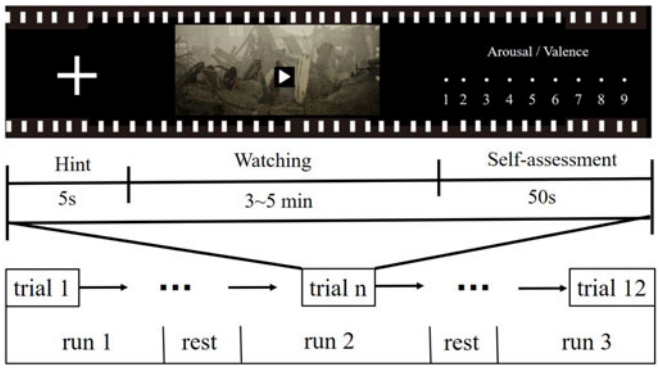


Fig. 1. Timeline for the experimental protocol.

film clips corresponded to 12 trials with EEG data, and each trial consisted of the following steps:

1) A 5 s start prompt before one video played.
2) The 3–5 minute display of one film clip.
3) A 50s period for subjective self-assessment of arousal and valence after watching a film clip.
4) A 15 s short break before the next film clip.

Fig. 1 shows the procedures of the experiment. Throughout the study, there were a total of 12 trials with EEG data for each subject. The quality of the EEG recordings was checked through visual inspection by an experienced neurophysiologist.

5.2 Feature Extraction From EEG Recordings

The EEG recordings were preprocessed by a 1–50 Hz bandpass filter. Independent component analysis was then implemented to remove physiological artifacts, such as electromyogram, electrooculogram, and electrocardiogram. We used entire trails for feature extraction without splitting a trial into several windows.

Based on previous studies [3], [4], [15], [22], [71], 12 kinds of EEG features were extracted from the EEG recordings in the time domain, frequency domain and time-frequency domain. Time domain features, such as the non-stationary index (NSI)[8], higher-order crossing (HOC)[9], spectral entropy [72], shannon entropy[73], and C0 complexity [74], were used to identify the characteristics of the time series.

The EEG signals were divided into five frequency ranges: delta band (1-4 Hz), theta band (4-8 Hz), alpha band (8-13 Hz), beta band (13-30 Hz), and gamma band (30-50 Hz) [75].

TABLE 4
The Dimension of Each Feature Type for Three Data Sets

Feature type	DEAP	SEED	HDED
NSI	32	62	90
HOC	320	620	900
spectral entropy	32	62	90
shannon entropy	32	62	90
C0 complexity	32	62	90
AP	160	310	450
AP_β/AP_θ	32	62	90
DE	160	310	450
AHTIMF	160	310	450
IPHTIMF	160	310	450
DASM	70	135	205
RASM	70	135	205
Total	1,260	2,440	3,560

These specific bands are more prominent in certain states of mind [3]. Frequency domain features, including differential entropy (DE)[10], absolute power (AP), and the absolute power ratio of the beta band to the theta band(AP_β/AP_θ) [76], were extracted for computing frequency-domain characteristics in the five frequency bands. The AP of each EEG band was computed by the autoregressive (AR) model. Compared with other spectral estimation methods, the AR model has an advantage in the spectral estimation of nonstationary signals [77], [78].

Considering that the EEG signal is nonstationary and nonlinear, time-frequency features extracted by time-frequency analysis methods can provide additional information about dynamical changes in EEG signals for emotion recognition [4], [12], [13]. The empirical mode decomposition method [79] was implemented to decompose the original EEG signal into several IMFs, and then the two time-frequency features, including the amplitude of the Hilbert transform of each intrinsic mode function (AHTIMF) and the instantaneous phase of the Hilbert transform of each intrinsic mode function (IPHTIMF)[4], were calculated from the IMFs.

In [80], Schmidt *et al.* demonstrated that the lateralization between the left and right hemispheres of the scalp is associated with emotions. Differential asymmetry (DASM)[10] and rational asymmetry (RASM)[81] were also extracted from the symmetric electrodes in the topographical map. The definitions of the 12 EEG feature types are described in detail in [4], [9], [10], [72], [73], [74], [76]. The dimension of each feature type for three data sets is shown in Table 4. Since the 90 electrodes near the brain center (Cz electrode) in HDED can throughout the whole brain area, Only 90 electrodes were chosen to compute the features and analyze the distribution of electrodes (corresponding to the discriminative features) on the topographical map.

5.3 Experimental Setup

In the experiment, we chose nine popular feature selection methods for emotion recognition and then performed a quantitative comparison on removing redundancy and retaining discriminative information between the ten alternative feature selection methods, including ReliefF [82], CMIM [49], mRMR [83], FSOR [22], and orthogonal regression with

minimum redundancy (ORMR) [84], robust feature selection (RFS) [57], embedded supervised feature selection (ESFS) [85], robust and pragmatic multi-class feature selection (RPMFS) [59], subspace sparsity discriminant feature selection (SDFS) [86], and the GRMOR method.

The aforementioned feature selection approaches were employed to obtain EEG feature subsets, and then the selected EEG feature subsets were assigned as the input of a classifier to evaluate the performance of the subsets in the emotion recognition task. According to the survey in [3], the support vector machine (SVM) is the most commonly used classifier in almost 59 percent of EEG-based emotion recognition studies. Due to its simplicity, the SVM classifier with a linear function was implemented. We applied the popular LIBSVM toolbox to employ the multiclass SVM with a linear function. In our experiment, the dimension of arousal was chosen as the emotional evaluation standard for the DEAP dataset, and the threshold to divide the samples into two categories (high/low) according to the rated levels of arousal was set to five.

To examine the effectiveness of the different feature selection methods, 70 percent of the samples and the remaining 30 percent of the samples were randomly chosen as a training set and a test set, respectively. It should be noted that all the samples of one subject were selected either as training set or test set to evaluate the generalization abilities of all the feature selection methods. In other words, the samples of one subject will not exist in the training set and test set at the same time. We obtained 50 independent realizations, and the average emotion recognition rate was calculated to avoid possible bias.

5.4 Evaluation Metrics

We utilized two evaluation metrics to evaluate the emotion recognition and redundancy reduction capabilities of the feature selection methods.

1) Classification Accuracy: The first evaluation measure is defined as

$$Accuracy = \frac{r}{n} \times 100\%, \tag{27}$$

where r and n are the correctly classified EEG sample number and the total EEG sample number, respectively.

2) Redundancy: The redundancy among the features can be calculated by

$$\text{Redundancy}(G) = \frac{1}{m(m-1)} \sum_{\mathbf{f}_i, \mathbf{f}_j \in G, i \neq j} A_{i,j}, \tag{28}$$

where G is the obtained EEG feature subset, m denotes the number of selected EEG features in the EEG feature subset G, and $A_{i,j}$ is the correlation between the variables $\mathbf{x}_i$ and $\mathbf{x}_j$. The correlation is evaluated in terms of squared cosine similarity.

6 Experimental Results and Discussion

In this section, we present extensive experiments to show the effectiveness of the proposed GRMOR algorithm on the three EEG data sets.

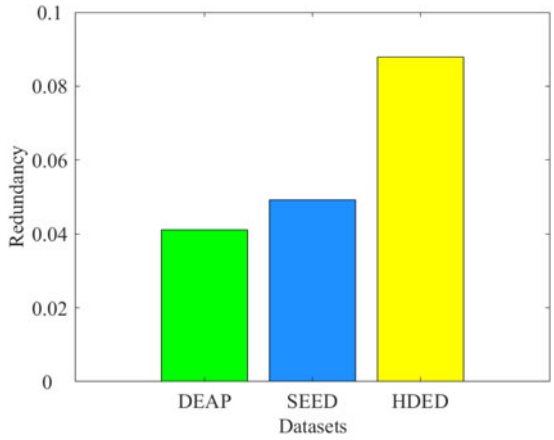


Fig. 2. The redundancy comparison between the three EEG data sets.

6.1 Dataset Redundancy Comparison

To demonstrate that the redundancy issue in the EEG features becomes more serious as the number of EEG channels increases, we first compared the redundancies in the feature data matrices extracted from the three data sets. The average redundancy is shown in Fig. 2. As observed from Fig. 2, as the number of EEG channels increases, the amount of redundant information continues to increase. Moreover, the gap in redundancy between the results on the HDED and SEED data sets is much greater than that between the results on the DEAP and SEED data sets, which indicates that redundant information removal is more important for a high-density EEG than it is for a low-density EEG.

This may be due to the volume conduction of the human head. The human head can be treated as a volume conductor, and thus, the recorded signal at each electrode is the result of several underlying sources, and this introduces interchannel dependence [39]. Compared with the low-density EEG data sets (DEAP and SEED), the electrodes of the high-density EEG data set (HDED) are more densely distributed on the scalp. The spatial density of the electrodes results in more interchannel dependence, which leads to highly redundant information in the feature data matrix.

6.2 Performance Comparison

Next, we analyze the performance of the ten supervised feature selection methods on the three EEG data sets. The redundancies and average classification accuracies of the EEG feature subsets obtained by the ten methods are shown in Figs. 3 and 4, respectively. It should be noted that the maximum number of selected EEG features on the x axis is determined by the highest classification accuracy. For example, the highest classification accuracy of GRMOR on the DEAP data set can be achieved when the number of selected EEG features is 30. Even if the number of selected EEG features is more than 30, the recognition accuracies of the EEG feature subsets obtained by all the methods do not exceed the highest accuracy.

We can see that the EEG feature subset of the GRMOR method consistently yielded the best performance in terms of classification accuracy and redundancy. Specifically, we made the following observations.

First, the traditional mRMR and CMIM methods did not function well. As shown in Fig. 3, the two methods can

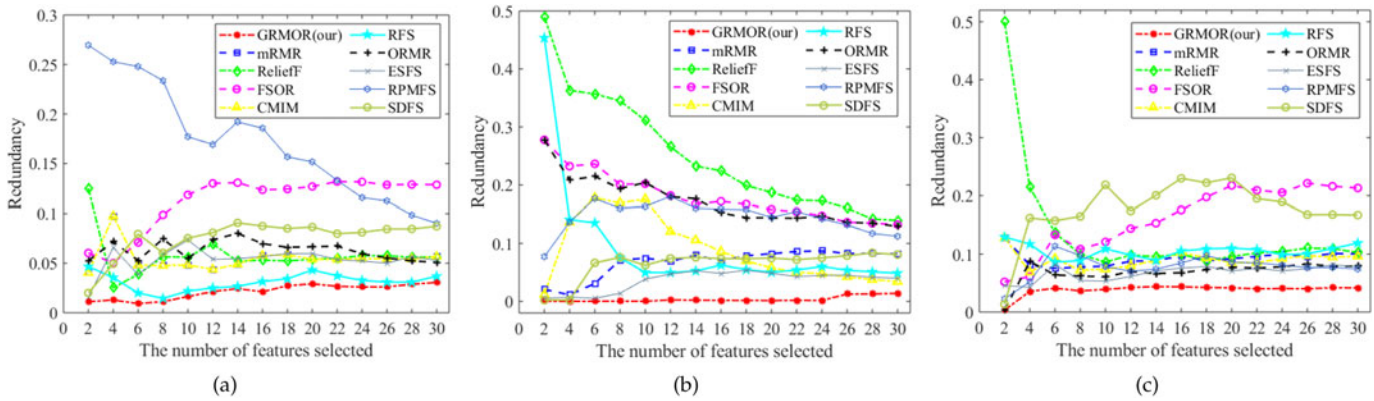


Fig. 3. Redundancies with different numbers of selected EEG features on the three EEG emotional data sets: (a) DEAP; (b) SEED; (c) HDED.

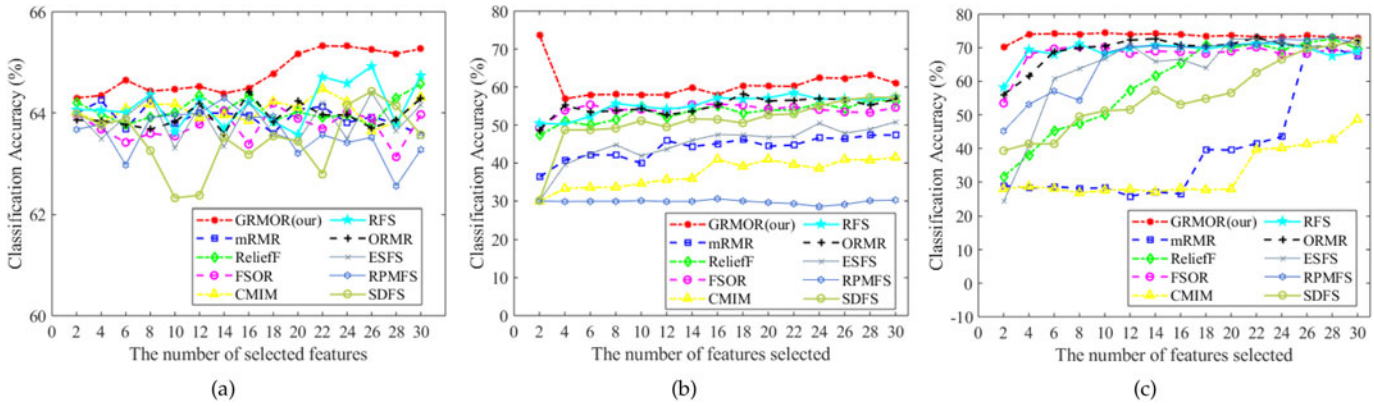


Fig. 4. Classification accuracies with different numbers of selected EEG features on the three EEG emotional data sets: (a) DEAP; (b) SEED; (c) HDED.

minimize the redundancy in the selected EEG feature subset to a certain extent. However, they utilize a greedy strategy to evaluate feature dependence; thus, global redundancy minimization cannot be achieved, and the selected feature subset is not an optimal output. The issue was more serious with the high-density EEG data set due to the high redundancy. As shown in Fig. 4c, the mRMR and CMIM methods performed worse than other methods on the high-density EEG data set.

Second, as shown in Figs. 3b, 3c, 4b, and 4c, ReliefF had poor emotion recognition performance when we selected few EEG features. As the number of selected EEG features increased, the performance of ReliefF improved significantly. However, ReliefF still performed worse than GRMOR even if a sufficient number of EEG features was selected as the EEG feature subset.

Third, we can observe from Fig. 3 that the redundancy of the EEG feature subsets obtained by the FSOR method increased rapidly as the number of selected EEG features increased. This can be attributed to the fact that the FSOR method ignores the dependence between EEG features during the construction of the learning model, and so much redundant information is retained in the selected EEG feature subset. From Fig. 3, it can be seen that the redundancy of the ORMR method is roughly lower than that of the FSOR method. However, the classification accuracies of the ORMR method are not higher than those of the FSOR method. Since the ORMR method is a two-step strategy based on the mRMR and FSOR methods, it also has the same disadvantages as the mRMR and FSOR methods.

Last but not least, as shown in Figs. 3 and 4, the proposed GRMOR method took the global redundancy information into account and achieved a greater improvement in emotion recognition performance than all other embedded-type methods (ESFS, RPMFS, RFS, SDFS, ORMR, and FSOR) did. Compared with other embedded-type methods, the number of EEG features required by the GRMOR method to obtain the maximum classification accuracy is the least, which shows that the GRMOR method can rapidly select the optimal EEG feature subset with a minimal number of EEG features and minimal redundancy.

6.3 Feature Ranking

To investigate the importance of different EEG feature types in emotion recognition, we further analyze the feature scores ranked by the GRMOR method and create a histogram of the average feature scores in Fig. 5. Each average feature score is computed by averaging all the scores of an EEG feature type. For example, on the DEAP data set, the dimension of DE is 160 (32 electrodes*5 frequency bands). Hence, we calculate the mean of the 160 feature rankings as the average feature score of DE.

Based on the assumption that the scores of valuable EEG features are higher than those of useless EEG features in terms of successful emotion recognition, we compare the importance of the different EEG features for emotion recognition. As shown in Fig. 5, the experimental results suggest that there is no one optimal EEG feature type or a most

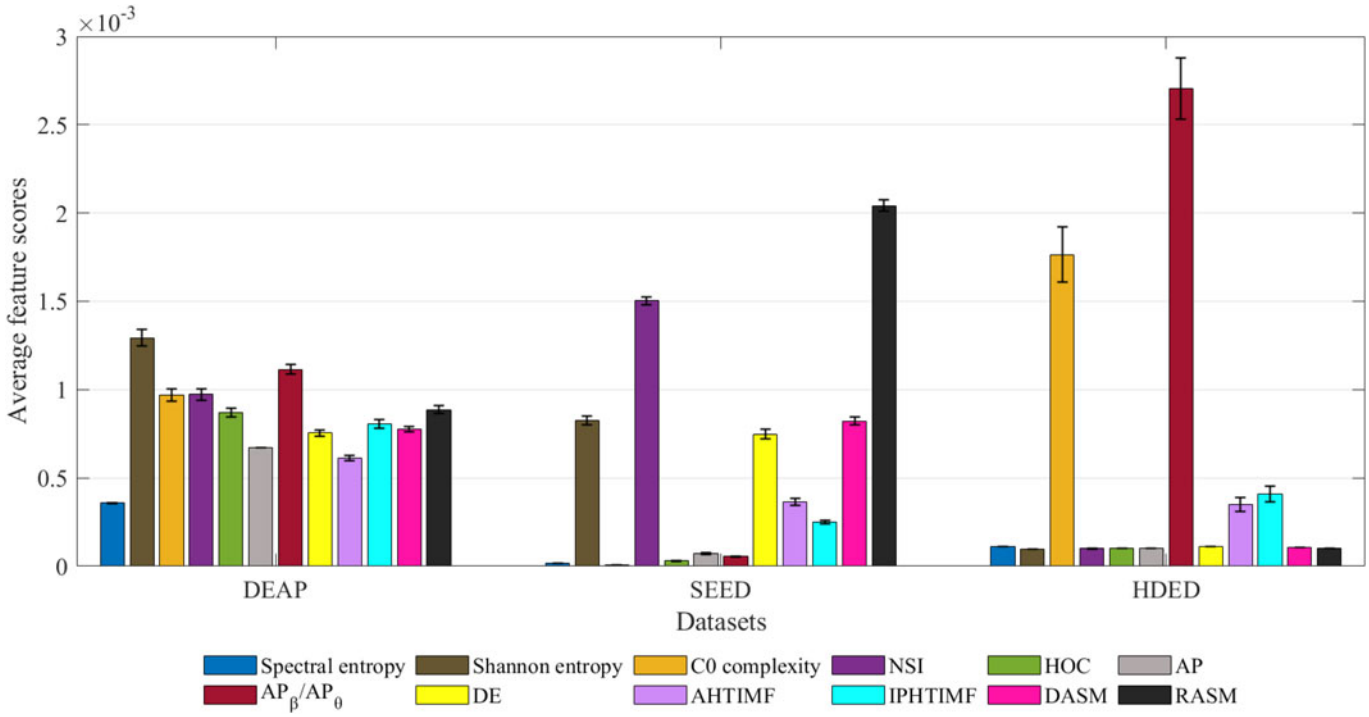


Fig. 5. The importance of different EEG feature types on the emotion recognition task is evaluated based on the feature scores ranked by the proposed GRMOR method.

valuable subgroup of multiple EEG feature types that is applicable to all the data sets. Hence, for an EEG-based emotion recognition task, we recommend using the GRMOR method to adaptively select an optimal subgroup from the group of multiple types of EEG features, rather than using a specific EEG feature type.

To further discuss the characteristics of selected features of GRMOR, we analyze the electrode locations of selected features from the perspective of different brain regions (frontal, left temporal, right temporal, occipital, and central). Refer to the study [87], all the electrodes were divided into five brain regions. The distribution of electrodes corresponding to the top ten features of the GRMOR method is shown in Fig. 6. In contrast to SEED and HDED, no feature in DEAP occurs in the occipital area (visual function area), and most features in DEAP exist in the left temporal and right temporal areas (audio function areas), which can be attributed to the different stimulus materials. DEAP, SEED, and HDED adopted music, film, and film as stimulus materials, respectively.

To all the data sets, similar electrode positions can be observed in Fig. 6, such as the CP5 electrode in DEAP, the CP5 electrode in SEED, and the 50th electrode in HDED. From the topographic map distribution of EEG electrodes, these three electrodes are located on the surface of the scalp corresponding to left inferior temporal gyrus area, and hence part of the source signals of the EEG recordings collected by these three electrodes may come from the left inferior temporal gyrus area. The observation is consistent with previous researches [69], [88]. In [88], the functional magnetic resonance imaging data of 26 male healthy subjects were collected to investigate the brain response differential between positive and negative experiences. The results in [88] showed that stronger activations occurred in the inferior temporal gyrus area during the happiness experience. In [69], the selected 12 EEG channels containing the CP5 electrode have the best performance on the SEED dataset than the performance of other EEG electrode selection settings (4 channels, 6 channels, 9 channels, and full 62 channels). The other electrode selection settings (4 channels, 6 channels, and 9 channels) did not include the CP5 electrode. In conclusion, the results in Fig. 6 and other emotion-related researches indicate that the CP5/50th electrode is a critical channel, and the features derived from the CP5/50th electrode are discriminative for affective computing.

6.4 Selected Feature Subset Comparison

To discuss the reasons for the differences in emotion recognition performance between the various methods, we analyze the feature subsets selected by the six feature selection methods. Taking the HDED data set as an example, the feature subsets with their corresponding channels are listed in Table 5. We can observe from Fig. 4c that the best classification performance can be obtained by only 4 features selected by GRMOR. Hence, we only list four selected features in Table 5. Table 5 shows that the feature types in the feature subsets of ReliefF, mRMR, and CMIM are too singular. The features selected by mRMR and CMIM are all DASM features, and the features chosen by ReliefF are all time-frequency features. Combining the results in Figs. 3c and 4c, we can find that the features selected by ReliefF, mRMR, CMIM, ESFS, RPMFS, and SDFS are ineffective and redundant for emotion recognition.

In addition, we find that the RFS, FSOR, ORMR, and GRMOR methods all selected the C0 complexity feature in the 59th channel, which shows that this feature is an informative feature for the HDED data set. Compared with the

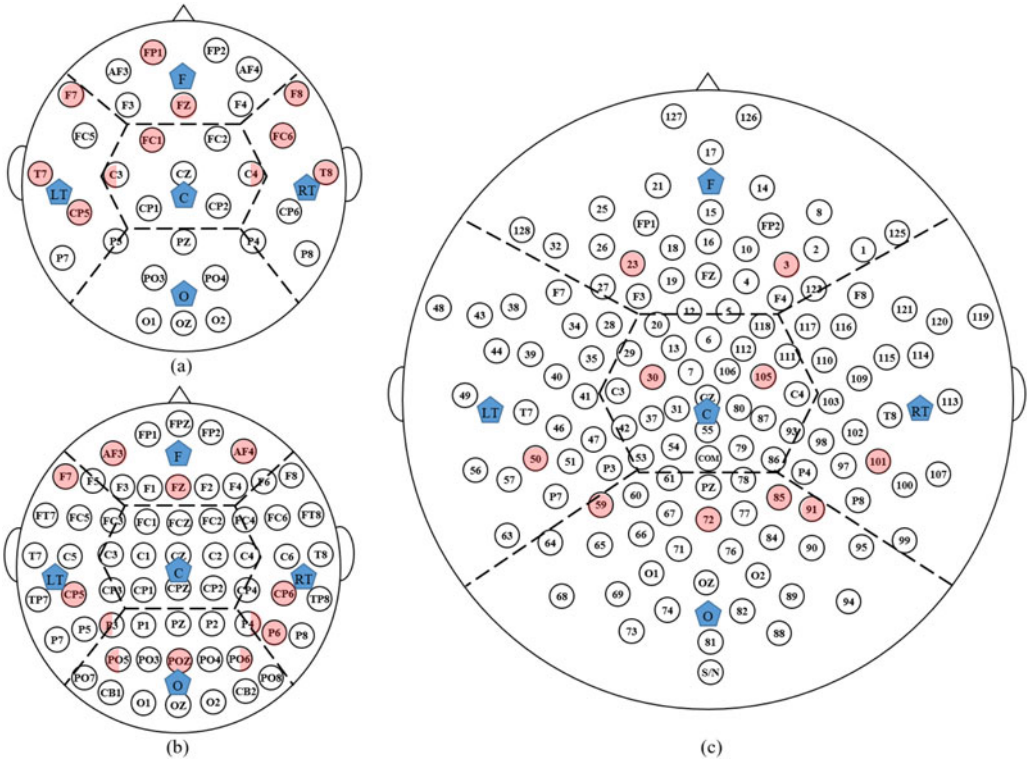


Fig. 6. The distribution of electrodes (corresponding to the top ten features of the GRMOR method) on the topographical map: (a) DEAP; (b)SEED; (c) HDED. Half painted circles represent the features between symmetric electrodes. Pentagons represent different brain regions.

TABLE 5
The Four Selected EEG Features With the Highest Rankings on the HDED Data Set

Method / Order	1	2	3	4
ReliefF	AHTIMF (E49)	ITHTIMF (E49)	AHTIMF (E27)	ITHTIMF (E27)
mRMR	DASM (E60-E85)	DASM (E35-E110)	DASM (E31-E80)	DASM (E34-E116)
CMIM	DASM (E60-E85)	DASM (E35-E110)	DASM (E50-E101)	DASM (E24-E124)
RFS	C0 complexity (E59)	IPHTIMF (E101)	AHTIMF (E101)	AP_β/AP_θ (E112)
ESFS	DE (γ band) (E18)	NSI (E13)	DE (α band) (E9)	AHTIMF (E96)
RPMFS	Shannon entropy (E83)	RASM (E71-E76)	C0 complexity (E12)	IPHTIMF (E78)
SDFS	DASM (E29-E111)	HOC (E80)	DE (θ band) (E116)	HOC (E19)
FSOR	C0 complexity (E59)	HOC (E101)	AP_β/AP_θ (E105)	HOC (E60)
ORMR	C0 complexity (E59)	ITHTIMF (E89)	HOC (E10)	HOC (E12)
GRMOR	C0 complexity (E59)	AP_β/AP_θ (E105)	ITHTIMF (E101)	AP_β/AP_θ (E85)

GRMOR method, the feature subsets of the FSOR and ORMR methods both contain two HOC features, and the feature subset of the RFS method contains two time-frequency features in the 101th channel. However, as shown in Figs. 3c and 4c, the feature subsets containing HOC performed worse than the feature subset obtained by the GRMOR method, which indicates that the feature subsets containing HOC may not be the optimal feature selection results. The feature subset of the RFS method contains the redundant information (IPHTIMF and AHTIMF) in the 101th channel.

We can also observe from Fig. 6c that the first three EEG features of the GRMOR method were extracted from different domains and brain regions. Different types of informative features provide useful and nonredundant discriminative information for emotion recognition. C0 complexity, ITHTIMF, and AP_β/AP_θ can be used to represent characteristics of time series, dynamical changes[12], [13], and reduced frontal cortical control over subcortical affective approach drive [89], respectively.

6.5 Parameter Sensitivity

Moreover, we investigate the unique parameter λ in the GRMOR method. λ is applied to control the magnitude of the difference between the orthogonal regression term and global redundancy minimization term. Varying the values of the parameter λ from 10^{-3} to 10^3, the average classification accuracies on three data sets are shown in Fig. 7. It can be observed from Fig. 7 that the GRMOR method obtains competitive performance on the SEED and HDED data sets when setting $\lambda = 0.001$. Hence, the final emotion recognition performance of the proposed GRMOR method is sensitive to the value of λ.

6.6 Convergence Demonstration

Here, we experimentally study the convergence speed of the optimization algorithm. The behaviors of the objective

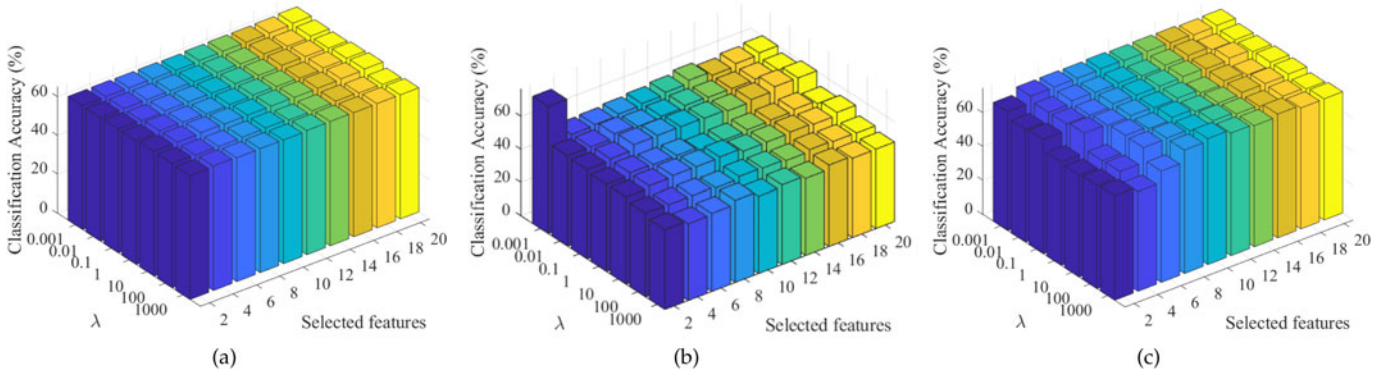


Fig. 7. Classification accuracies of the GRMOR algorithm with respect to different numbers of selected EEG features under the varying parameter λ alternatively on the three EEG data sets: (a) DEAP; (b) SEED; (c) HDED.

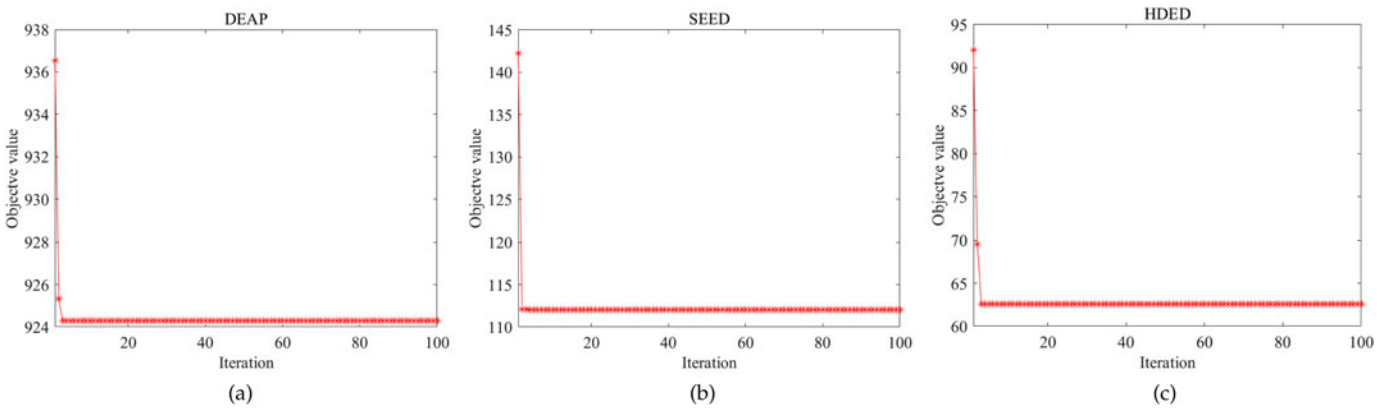


Fig. 8. Convergence of the GRMOR algorithm on the three data sets: (a) DEAP; (b) SEED; (c) HDED.

values of the GRMOR algorithm ($\lambda = 1$) at each iteration are shown in Fig. 8. The number of iterations was set to 100. As observed from Fig. 8, the objective values on the three EEG data sets decrease monotonically and always converge within a few iterations (i.e., less than 5), which indicates that GRMOR converges sufficiently.

6.7 Computational Complexity Comparison

The computational complexity of GRMOR can be mainly attributed to the calculation of W. To reduce the computational cost of GRMOR, direct calculation of JW in the third step of Algorithm 1 is used instead of calculating J and then multiplying by W. Based on the GPI method, the GRMOR algorithm takes $\mathcal{O}(dkn)$ to compute the JW. Additionally, the computational complexity of ALM is also $\mathcal{O}(dkn)$. Hence, the total computational complexity of GRMOR is $\mathcal{O}(dkn)$. It should be noted that d, n, k, and m ($1 \leq m < d$) are the number of total EEG features, EEG samples, emotional categories, and selected EEG features, respectively.

The computational costs of the ten feature selection methods in our study are shown in Table 6. m ($1 \leq m < d$) is the number of selected EEG features. As shown in Table 6, when the value of k is small, the GRMOR method and the traditional filter-type approaches (ReliefF, mRMR, and CMIM) have similar computational complexities.

TABLE 6
Comparisons of Computational Complexity

Method	Computational Complexity
ReliefF	$\mathcal{O}(dn)$
mRMR	$\mathcal{O}(dmn)$
CMIM	$\mathcal{O}(dmn)$
RFS	$\mathcal{O}(d^3 + dn^2 + d^2n + n^3 + dkn)$
ESFS	$\mathcal{O}(dkn + k^2n)$
RPMFS	$\mathcal{O}(d^2)$
SDFS	$\mathcal{O}(d^2m)$
FSOR	$\mathcal{O}(dkn)$
ORMR	$\mathcal{O}(dmn + vkn)$
GRMOR	$\mathcal{O}(dkn)$

7 Conclusion

To remove redundant EEG features and select discriminative features for emotion recognition, we propose a novel EEG feature selection method for emotion recognition. Compared with popular feature selection methods, for highly correlated EEG features, GRMOR can effectively evaluate the dependence among all EEG features from a global view and then select a discriminative and nonredundant EEG feature subset for emotion recognition. However, the solution of the objective function of the GRMOR method is an unbalanced orthogonal Procrustes problem on the Stiefel manifold, which makes it difficult to obtain an optimal

solution. To solve this problem, an alternative algorithm based on GPI and a general ALM is employed to optimize the objective function.

Moreover, to evaluate the effectiveness of GRMOR, we utilize three EEG emotional data sets. The experimental results demonstrate that the proposed method can effectively discard the redundant EEG features from the original set of EEG features and obtain a discriminative EEG feature subset for emotion recognition. Compared with nine commonly used feature selection methods, the effectiveness and superiority of the GRMOR method on the low-density EEG and high-density EEG are verified both theoretically and experimentally.

Although GRMOR has been proven to be a promising tool for redundant feature removal and informative feature selection from highly correlated EEG features under supervised situations, whether or not it is applicable for semi-supervised or unsupervised situations still needs to be investigated. In further research, we will focus on the extension of GRMOR to situations with semi-supervised or unsupervised EEG feature selection.

ACKNOWLEDGMENTS

This work was supported by the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China (Grant No. SKLMCC2020KF001), and the General Program of Beijing Municipal Natural Science Foundation (Grant No. 4212037).

REFERENCES

[1] P. Zhong, D. Wang, and C. Miao, "EEG-based emotion recognition using regularized graph neural networks," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.2994159.
[2] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," *IEEE Trans. Cognitive Develop. Syst.*, vol. 9, no. 3, pp. 281–290, Sep. 2017.
[3] S. M. Alarcão and M. J. Fonseca, "Emotions recognition using EEG signals: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 374–393, Jul.–Sep. 2017.
[4] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul.–Sep. 2014.
[5] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2266–2274, Mar. 2019.
[6] C. A. Frantzidis, C. Bratsas, C. L. Papadelis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "Toward emotion aware computing: An integrated approach using multichannel neurophysiological recordings and affective visual stimuli," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 3, pp. 589–597, May 2010.
[7] B. Hjorth, "EEG analysis based on time domain properties," *Electroencephalogr. Clin. Neurophysiol.*, vol. 29, no. 3, pp. 306–310, 1970.
[8] E. Kroupi, A. Yazdani, and T. Ebrahimi, "EEG correlates of different emotional states elicited during watching music videos," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2011, pp. 457–466.
[9] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 186–197, Mar. 2010.
[10] R. Duan, J. Zhu, and B. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. 6th Int. IEEE/EMBS Conf. Neural Eng.*, 2013, pp. 81–84.
[11] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Investigating the role of glottal features in classifying clinical depression," in *Proc. 25th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2003, pp. 2849–2852.
[12] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3498–3510, Dec. 2012.
[13] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, I. Zunaidi, and D. Hazry, "EEG feature extraction for classifying emotions using FCM and FKM," *Int. J. Comput. Commun.*, vol. 1, no. 2, pp. 21–25, 2007.
[14] H. Becker, J. Fleureau, P. Guillotel, F. Wendling, I. Merlet, and L. Albera, "Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 244–257, Apr.–Jun. 2020.
[15] H. Wang, X. Wu, and L. Yao, "Identifying cortical brain directed connectivity networks from high-density EEG for emotion recognition," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.3006847.
[16] F. Wang, S. Wu, W. Zhang, Z. Xu, Y. Zhang, C. Wu, and S. Coleman, "Emotion recognition with convolutional neural network and EEG-based EFDMS," *Neuropsychologia*, vol. 146, 2020, Art. no. 107506.
[17] C. Zhan, D. She, S. Zhao, M.-M. Cheng, and J. Yang, "Zero-shot emotion recognition via affective structural embedding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1151–1160.
[18] J. Zhang, M. Chen, S. Zhao, S. Hu, Z. Shi, and Y. Cao, "ReliefF-based EEG sensor selection methods for emotion recognition," *Sensors*, vol. 16, no. 10, p. 1558, 2016.
[19] J. Chen, B. Hu, P. Moore, X. Zhang, and X. Ma, "Electroencephalogram-based emotion assessment system using ontology and data mining techniques," *Appl. Soft Comput.*, vol. 30, pp. 663–674, 2015.
[20] X. Wang, D. Nie, and B. Lu, 'EEG-based emotion recognition using frequency domain features and support vector machines," in *Int. Conf. Neural Inf. Process.*, 2011, pp. 734–743.
[21] M. Zabihi, S. Kiranyaz, T. Ince, and M. Gabbouj, "Patient-specific epileptic seizure detection in long-term EEG recording in paediatric patients with intractable seizures," in *Proc. IET Intell. Sig. Process. Conf.*, 2013, pp. 1–7.
[22] X. Xu, F. Wei, Z. Zhu, J. Liu, and X. Wu, 'EEG feature selection using orthogonal regression: Application to emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Sig. Process.*, 2020, pp. 1239–1243.
[23] C. Hou, F. Nie, D. Yi, and Y. Wu, "Feature selection via joint embedding learning and sparse regression," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1324–1329.
[24] X. Wu, X. Xu, J. Liu, H. Wang, B. Hu, and F. Nie, "Supervised feature selection with orthogonal regression and feature weighting," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2020.2991336.
[25] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Inf. Fusion*, vol. 50, pp. 158–167, 2019.
[26] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
[27] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to instability problems with sequential wrapper-based approaches to feature selection," Dept. Comput. Sci. Courses, Trinity College, Dublin, Tech. Rep. 2002-28, 2002.
[28] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.
[29] D. Cai, X. He, J. Han, and H.-J. Zhang, "Orthogonal Laplacianfaces for face recognition," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3608–3614, Nov. 2006.
[30] R. Zhang, F. Nie, and X. Li, "Feature selection under regularized orthogonal least square regression with optimal scaling," *Neurocomputing*, vol. 273, pp. 547–553, 2018.
[31] S. P. van den Broek, F. Reinders, M. Donderwinkel, and M. Peters, "Volume conduction effects in EEG and MEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 106, no. 6, pp. 522–534, 1998.
[32] P. L. Nunez *et al.*, 'EEG coherency: I: Statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 5, pp. 499–515, 1997.
[33] R. Vigário, J. Sarela, V. Jousmiki, M. Hamalainen, and E. Oja, "Independent component approach to the analysis of EEG and MEG recordings," *IEEE Trans. Biomed. Eng.*, vol. 47, no. 5, pp. 589–593, May 2000.

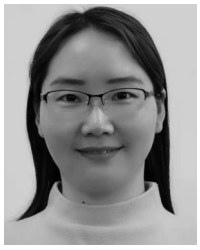
[34] C. W. Hesse and C. J. James, "On semi-blind source separation using spatial constraints with applications in EEG analysis," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2525–2534, Dec. 2006.
[35] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller, "A fully automated correction method of EOG artifacts in EEG recordings," *Clin. Neurophysiol.*, vol. 118, no. 1, pp. 98–104, 2007.
[36] Y. Wongsawat, S. Oraintara, T. Tanaka, and K. R. Rao, "Lossless multi-channel EEG compression," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2006, pp. 1611–1614.
[37] B. Hejrati, A. Fathi, and F. Abdali-Mohammadi, "Efficient lossless multi-channel EEG compression based on channel clustering," *Biomed. Sig. Process. Control*, vol. 31, pp. 295–300, 2017.
[38] X. Xu, X. Chen, and Y. Zhang, "Removal of muscle artefacts from few-channel EEG recordings based on multivariate empirical mode decomposition and independent vector analysis," *Electron. Lett.*, vol. 54, no. 14, pp. 866–868, 2018.
[39] X. Chen, X. Xu, A. Liu, M. J. McKeown, and Z. J. Wang, "The use of multivariate EMD and CCA for denoising muscle artifacts from few-channel EEG recordings," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 359–370, Feb. 2018.
[40] R. Hussein, A. Mohamed, and M. Alghoniemy, "Scalable real-time energy-efficient eeg compression scheme for wireless body area sensor network," *Biomed. Sig. Process. Control*, vol. 19, pp. 122–129, 2015.
[41] H. U. Amin, A. S. Malik, N. Kamel, and M. Hussain, "A novel approach based on data redundancy for feature extraction of EEG signals," *Brain Topogr.*, vol. 29, no. 2, pp. 207–217, 2016.
[42] B. Nakisa, M. N. Rastgoo, D. Tjondronegoro, and V. Chandran, "Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors," *Expert Syst. App.*, vol. 93, pp. 143–155, 2018.
[43] M. Odabaee *et al.*, "Neonatal EEG at scalp is focal and implies high skull conductivity in realistic neonatal head models," *Neuroimage*, vol. 96, pp. 73–80, 2014.
[44] A. M. Narayanan and A. Bertrand, "Analysis of miniaturization effects and channel selection strategies for EEG sensor networks with application to auditory attention detection," ., vol. 67, no. 1, pp. 234–244, Apr. 2019.
[45] A. Soler, P. A. Muñoz-Gutiérrez, M. Bueno-López, E. Giraldo, and M. Molinas, "Low-density EEG for neural activity reconstruction using multivariate empirical mode decomposition," *Front. Neurosci.*, vol. 14, p. 175, 2020.
[46] N. Mammone *et al.*, "Estimating the asymmetry of brain network organization in stroke patients from high-density EEG signals," in *Proc. Neural Approaches Dyn. Sig. Exchanges*, 2020, pp. 475–483.
[47] F. Nie, S. Yang, R. Zhang, and X. Li, "A general framework for auto-weighted feature selection via global redundancy minimization," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2428–2438, Dec. 2018.
[48] D. Wang, F. Nie, and H. Huang, "Feature selection via global redundancy minimization," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 10, pp. 2743–2755, Oct. 2015.
[49] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, no. 11, pp. 1531–1555, 2004.
[50] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classification, Algorithms Appl.*, p. 37, 2014.
[51] R. Jenke, A. Peer, and M. Buss, "Effect-size-based electrode and feature selection for emotion recognition from EEG," in *Proc. IEEE Int. Conf. Acoust., Speech Sig. Process.*, 2013, pp. 1217–1221.
[52] R. M. Mehmood, R. Du, and H. J. Lee, "Optimal feature selection and deep learning ensembles method for emotion recognition from human brain EEG sensors," *IEEE Access*, vol. 5, pp. 14 797–14 806, Jul. 2017.
[53] Z. Lan, O. Sourina, L. Wang, Y. Liu, R. Scherer, and G. R. Müller-Putz, "Stable feature selection for EEG-based emotion recognition," in *Proc. Int. Conf. Cyberworlds*, 2018, pp. 176–183.
[54] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, 2016.
[55] Y. Kong, J. Yan, and H. Xu, "Rega based feature selection emotion recognition using EEG signals," in *Proc. Chin. Automat. Congr.*, 2017, pp. 6588–6593.
[56] X. Li, H. Zhang, R. Zhang, and F. Nie, "Discriminative and uncorrelated feature selection with constrained spectral analysis in unsupervised learning," *IEEE Trans. Image Process.*, vol. 29, no. 1, pp. 2139–2149, Oct. 2019.
[57] F. Nie, H. Huang, X. Cai, and C. H. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.
[58] H. Tao, C. Hou, F. Nie, Y. Jiao, and D. Yi, "Effective discriminative feature selection with nontrivial solution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 796–808, Apr. 2016.
[59] X. Cai, F. Nie, and H. Huang, "Exact top-k feature selection via ℓ 2, 0-norm constraint," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1240–1246.
[60] H. Zhao, Z. Wang, and F. Nie, "Orthogonal least squares regression for feature extraction," *Neurocomputing*, vol. 216, pp. 200–207, 2016.
[61] H. Zhao, S. Wang, and Z. Wang, "Multiclass classification and feature selection based on least squares regression with large margin," *Neural Comput.*, vol. 30, no. 10, pp. 2781–2804, 2018.
[62] P. Glaister, "The use of orthogonal distances in generating the total least squares estimate," *Math. Comput. Educ.*, vol. 39, no. 1, pp. 21, 2005.
[63] H. Wason, R. Das, and M. Sharma, "Magnitude conversion problem using general orthogonal regression," *Geophys. J. Int.*, vol. 190, no. 2, pp. 1091–1096, 2012.
[64] L. Leng, T. Zhang, L. Kleinman, and W. Zhu, "Ordinary least square regression, orthogonal regression, geometric mean regression and their applications in aerosol science," *J. Phys.: Conf. Ser.*, vol. 78, no. 1,2007, Art. no. 012084.
[65] X. Wang, Y. Liu, F. Nie, and H. Huang, "Discriminative unsupervised dimensionality reduction," in *24th Int. Joint Conf. Artif. Intell.*, 2015, pp. 3925–3931.
[66] F. Nie, S. Yang, R. Zhang, and X. Li, "A general framework for auto-weighted feature selection via global redundancy minimization," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2428–2438, May 2019.
[67] L. Eldén and H. Park, "A procrustes problem on the Stiefel manifold," *Numerische Mathematik*, vol. 82, no. 4, pp. 599–619, 1999.
[68] F. Nie, R. Zhang, and X. Li, "A generalized power iteration method for solving quadratic problem on the Stiefel manifold," *Sci. China Inf. Sci.*, vol. 60, no. 11, 2017, Art. no. 112101.
[69] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Ment. Develop.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.
[70] S. Koelstra *et al.*, "Deap: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.–Mar. 2012.
[71] T. Song, W. Zheng, P. Song, and Z. Cui, "EEG emotion recognition using dynamical graph convolutional neural networks," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 532–541, Jul.–Sep. 2020.
[72] A. Zhang, B. Yang, and L. Huang, "Feature extraction of EEG signals using power spectral entropy," in *Proc. Int. Conf. BioMed. Eng. Inform.*, 2008, pp. 435–439.
[73] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, Jan. 1991.
[74] Y. Zhou, L. Xie, G. Yu, F. Liu, Y. Zhao, and Y. Huang, "The study of c0 complexity on epileptic absence seizure," in *Proc. 7th Asian-Pacific Conf. Med. Biolo. Eng.*, 2008, pp. 420–425.
[75] X. Chen, *et al.*, "Removal of muscle artifacts from the EEG: A review and recommendations," *IEEE Sensors J.*, vol. 19, no. 14, pp. 5353–5368, Jul. 2019.
[76] S. Koehler *et al.*, "Increased EEG power density in alpha and theta bands in adult ADHD patients," *J. Neural Transmiss.*, vol. 116, no. 1, pp. 97–104, 2009.
[77] S. Zhang and H. Dai, "Application of spectral estimation method on validation of simulation model of an aircraft," *Comput. Simul.*, vol. 9, pp. 48–49, 2003.
[78] Z. Li, X. Wu, X. Xu, H. Wang, Z. Guo, Z. Zhan, and L. Yao, "The recognition of multiple anxiety levels based on electroencephalograph," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2019.2936198.
[79] P. Flandrin, G. Rilling, and P. Goncalves, "Empirical mode decomposition as a filter bank," *IEEE Sig. Process. Lett.*, vol. 11, no. 2, pp. 112–114, Feb. 2004.

[80] L. Schmidt and L. Trainor, "Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions," *Cogn. Emot.*, vol. 15, no. 4, pp. 487–500, 2001.
[81] Y. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, May 2010.
[82] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *Proc. Eur. Conf. Mach. Learn.*, 1994, pp. 171–182.
[83] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.
[84] X. Xu and X. Wu, "Feature selection under orthogonal regression with redundancy minimizing," in *Proc. IEEE Int. Conf. Acoust., Speech Sig. Process.*, 2020, pp. 3457–3461.
[85] L. Chen, J. Tang, and B. Li, "Embedded supervised feature selection for multi-class data," in *Proc. SIAM Int. Conf. Data Mining* 2017, pp. 516–524.
[86] Z. Wang, F. Nie, L. Tian, R. Wang, and X. Li, "Discriminative feature selection via a structured sparse subspace learning module," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, 2020, pp. 3009–3015.
[87] Z. Bian, Q. Li, L. Wang, C. Lu, S. Yin, and X. Li, "Relative power and coherence of EEG series are related to amnestic mild cognitive impairment in diabetes," *Front. Aging Neurosci.*, vol. 6, p. 11, 2014.
[88] U. Habel, M. Klein, T. Kellermann, N. J. Shah, and F. Schneider, "Same or different? Neural correlates of happy and sad mood in healthy males," *Neuroimage*, vol. 26, no. 1, pp. 206–214, 2005.
[89] P. Putman, J. van Peer, I. Maimari, and S. van der Werff, "EEG theta/beta ratio in relation to fear-modulated response-inhibition, attentional control, and affective traits," *Biol. Psychol.*, vol. 83, no. 2, pp. 73–78, 2010.

Xueyuan Xu (Student Member, IEEE) is currently working toward the PhD degree in computer application technology from the School of Artificial Intelligence, Beijing Normal University, China. He has authored or coauthored more than ten papers in the journals and conferences, such as the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Affective Computing*, *IEEE Transactions on Instrumentation and Measurement*, *IEEE Sensors Journal*, and *IEEE International Conference on Acoustics, Speech, and Signal Processing*. His research interests include affective computing, feature selection, and blind source separation.

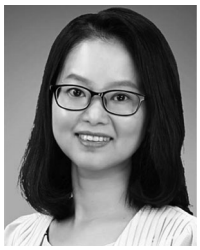
Tianyuan Jia is currently working toward the master's degree in information and signal processing from the School of Artificial Intelligence, Beijing Normal University, China. Her main research interests include affective computing, biological signal processing, and feature selection.

Qing Li is currently working toward the PhD degree in computer application technology from the School of Artificial Intelligence, Beijing Normal University, China. Her research interests include brain medical imaging analysis, deep learning, and multimodal fusion.



Fulin Wei is currently working toward the PhD degree in computer application technology from the School of Artificial Intelligence, Beijing Normal University, China. Her main research interests include affective computing, feature selection, and biological signal processing.

Long Ye received the PhD degree in communication and information system from the Communication University of China, Beijing, China, in 2012. He is currently a professor with the State Key Laboratory of Media Convergence and Communication, Communication University of China. His research interests include video sentiment analysis and computing, intelligent media technology, and virtual reality technology.



Xia Wu (Member, IEEE) received the PhD degree in basic psychology in 2007 from the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, where he is currently a professor with the School of Artificial Intelligence. Her main research interests include affective computing, feature extraction, and feature selection.