Med Biol Eng Comput (2017) 55:1589–1603
DOI 10.1007/s11517-017-1622-1

ORIGINAL ARTICLE

A spatial-frequency-temporal optimized feature sparse representation-based classification method for motor imagery EEG pattern recognition

Minmin Miao[1] · Aimin Wang[1] · Feixiang Liu[1]

Received: 13 June 2016 / Accepted: 25 January 2017 / Published online: 4 February 2017

Abstract Effective feature extraction and classification methods are of great importance for motor imagery (MI)-based brain–computer interface (BCI) systems. The common spatial pattern (CSP) algorithm is a widely used feature extraction method for MI-based BCIs. In this work, we propose a novel spatial-frequency-temporal optimized feature sparse representation-based classification method. Optimal channels are selected based on relative entropy criteria. Significant CSP features on frequency-temporal domains are selected automatically to generate a column vector for sparse representation-based classification (SRC). We analyzed the performance of the new method on two public EEG datasets, namely BCI competition III dataset IVa which has five subjects and BCI competition IV dataset IIb which has nine subjects. Compared to the performance offered by the existing SRC method, the proposed method achieves average classification accuracy improvements of 21.568 and 14.38% for BCI competition III dataset IVa and BCI competition IV dataset IIb, respectively. Furthermore, our approach also shows better classification performance when compared to other competing methods for both datasets.

Keywords Brain–computer interface · Motor imagery · Relative entropy · Sparse regularization · Sparse representation-based classification

✉ Aimin Wang
15006187659@sina.cn; wangam@seu.edu.cn

[1] School of Instrument Science and Engineering, Southeast University, No. 2 Sipailou, Nanjing 210096, China

1 Introduction

As one of the fastest emerging technologies in neurorehabilitation, brain–computer interface (BCI) provides a new non-muscular channel for paralyzed people to communicate with external world [4, 22]. Among various brain measures like magnetoencephalography (MEG), functional magnetic resonance imaging (fMRI), electrooculography (ECoG) and electroencephalography (EEG), noninvasive EEG is preferred to other modalities since it is portable, easy to implement, inexpensive and has a higher temporal resolution. Studies have shown that the rhythmic components of EEG can reflect event-related synchronization and desynchronization (ERS/ERD) over sensorimotor cortex when people perform mental imagination of movements [10, 23, 24, 38]. ERS/ERD can be quantified by band-power changes occurring when subjects do motor imagery (MI) tasks, i.e., imagine their limbs (left hand, right hand and foot).

Due to the feasibility and reliability of BCI systems heavily depend on robust and effective feature extraction of EEG signals and classification for the further translation into device commands, a large number of signal processing and pattern recognition algorithms have been introduced to EEG analysis.

For the former feature extraction issue, common spatial pattern (CSP) is one of the most popular and efficient techniques to extract ERS/ERD-related features and has been widely used for motor imagery BCI systems [1, 6, 8, 12, 14, 16, 25, 31, 34, 37, 39, 40, 42]. CSP algorithm finds spatial filters that simultaneously maximize the variance for one class while minimize the variance for another class, thus providing a natural approach to effectively estimate the discriminative information of MI. In order to accurately capture the band-power changes resulting from ERS/ERD,

most refinements of CSP are focused on the proper selection of optimal frequency bands which are typically subject specific and can hardly be determined manually. So far, three types of approaches have been mainly proposed to fix the problem of filter band selection. The first one is simultaneous optimization of spectral filters within the CSP [8, 12, 16]. The second one is searching the optimal frequency band based on evolutionary algorithm [25, 34, 39]. The third one is selection of significant features from multiple frequency bands. Novi et al. [40] proposed a method called sub-band common spatial pattern (SBCSP), in which the EEG signals are decomposed into sub-bands and CSP features are extracted in each sub-band for classification with score fusion. Subsequently, a more generalized method called filter bank common spatial pattern (FBCSP) was proposed by Ang et al. [1]. To further enhance the performance of FBCSP, a discriminative FBCSP (DFBCSP) [31] method uses Fisher ratio to select subject-specific filter bands. More recently, Zhang et al. [42] proposed a sparse filter band common spatial pattern (SFBCSP) method by exploiting sparse regression for automatic band selection. Apart from the selection of discriminative frequency bands, automated channel selection is also essential to enhance the performance of CSP feature extraction and subsequent classification by removing task-irrelevant and redundant channels. Recently, some efforts have been made to select the optimal EEG channels which are highly subject-dependent [14, 25, 37]. Another issue facing cue-based synchronous BCI [27, 34, 37, 39] is the selection of the time period for EEG classification. Although the time during which the subject begins to perform the task is known, the time when the brain activity is associated with the task is unknown. Some studies have shown that the reaction time defined as the elapsed time between the presentation of a sensory stimulus and the subsequent behavioral response is strongly associated with age [7]. Therefore, the optimal observation period in MI classification depends on subjects and many approaches have been proposed to determine the optimal time window automatically [2, 12, 13]. As described above, the optimal channels, frequency bands and time windows for MI classification are always subject specific. Therefore, an ideal BCI system should consider subject-adapted space–frequency–time patterns for classification.

For the latter classification problem, sparse representation-based classification (SRC) has received much attention recently in the pattern recognition field [9, 20, 36]. The idea of sparse representation was first proposed by Mallat [18], and the sparse representation problem mainly involves finding the most compact representation of a given signal, where the representation is expressed as a linear combination of columns in an overcomplete dictionary matrix. Sparse representation can be used for signal classification when compact description of a test signal is produced using a set of training signals incorporated in the dictionary [27]. This SRC method has been successfully applied to various areas and proved high classification accuracy, such as speech recognition [9], visual tracking [20] and face recognition [36]. Shin et al. [27] applied a SRC method to EEG-based MI BCIs, CSP filtering technique was used to extract band-power features which constructed the columns of the dictionary, and the results indicated that their method was competitive. However, in the preprocessing phase, they used 1–2 s time samples after the cue appeared and fixed the frequency range at 8–15 Hz [27]. This manual selection of time window and frequency band is not optimal for CSP feature extraction. Subsequently, they analyzed noise robustness of the SRC method to evaluate the capability of the SRC for non-stationary EEG signal classification [28]. More recently, they proposed a simple adaptive SRC scheme, in which dictionary update techniques and a dictionary modification method were investigated [29].

Note that the design of a good dictionary matrix is critical in SRC method, the columns of the dictionary matrix are made up of the band powers of the CSP filtered signals in the above SRC methods. So the degree of incoherence (columns from different classes are uncorrelated) of the dictionary heavily depends on the performance of CSP. If a dictionary is incoherent, a test signal from one particular class can be predominantly represented by the columns of the same class. In our proposed spatial-frequency-temporal optimized feature sparse representation-based classification (SFTOFSRC) method, in order to further improve the performance of SRC method for MI EEG classification, we will focus on optimizing CSP features in subject-adapted space–frequency–time patterns and provide a detailed design for a more incoherent dictionary.

2 Methods

2.1 Data description and evaluation scheme

This section provides details on two public datasets and their evaluation schemes to examine the performance of our proposed SFTOFSRC algorithm.

2.1.1 Dataset I: BCI competition III dataset IVa

In this dataset, EEG signals were recorded from five healthy subjects (“aa,” “al,” “av,” “aw,” “ay”) through 118 electrodes [5]. In each trial, a visual cue was shown for 3.5 s, during which three kinds of motor imageries were performed. The motor imageries of right hand and right foot were provided for the competition. The sampling rate was 1000 Hz, and signals were band-pass filtered between 0.05 and 200 Hz. There were 280 trials in total, 140 trials per task for each

subject. The numbers of trials for training and testing across subjects are as follows: subjects (right foot, right hand)—aa(88,80), al(112,112), av(42,42), aw(26,30) and ay(10,18)—were for training and the rest were for testing. For comparison with the results reported in the literature [30], the performance of our proposed method is evaluated by measuring the ratio of trials correctly classified to the total number of test trials for each subject.

2.1.2 Dataset II: BCI competition IV dataset IIb

This dataset was recorded from nine subjects at C3, Cz and C4 electrodes. In each trial, a visual cue was shown for 4.5 s, during which left-hand and right-hand motor imageries were performed. The sampling rate was 250 Hz, and signals were band-pass filtered between 0.5 and 100 Hz. For comparison with the results reported in the literature [42] under the same condition, only the third training sessions of the dataset, i.e., "B0103T," "B0203T,"…, "B0903T" were used. There were 160 trials in total, 80 trials per task for each subject. Unlike the evaluation scheme used for dataset I, the average classification accuracy of 10 × 10-fold cross-validation was used to evaluate the classification performance for this dataset in the literature [42]. Therefore, we also used 10 × 10-fold cross-validation for evaluation. The 10 × 10-fold cross-validation results were obtained by repeating the tenfold cross-validation for ten times. For tenfold cross-validation, in each fold of this procedure, each nine parts (144 trials) were for training and the remaining one part (16 trials) was for testing.

2.2 The proposed approach

The proposed EEG signal analysis system for MI classification is illustrated in Fig. 1. The flowchart describes procedures for both training and testing. Here we define the entire training data as Ω and testing data as ψ. Data preprocessing, spatial optimization, significant features extraction on frequency-temporal domains are implemented on Ω to obtain the necessary information for effective classification, such as optimal channels and the indexes of significant features on frequency-temporal domains. Finally, dictionary optimization is implemented to further enhance the performance of SRC.

2.2.1 Preprocessing

The continuous experimental data are first intercepted into single-trial data. Then, common average reference (CAR) [19] is adopted to re-reference them. Henceforth, a fifth-order Butterworth band-pass filter is used to remove noises over 40 Hz and slow baseline signal under 6 Hz. Finally, to reduce the computational cost of subsequent processing without omitting important time-series information, BCI competition III dataset IVa is down-sampled to 250 Hz.

2.2.2 Spatial optimization

It is generally known that the ERD/ERS phenomenon appears in a specific brain region so that only a few electrodes are activated during MI [17]. Besides, the irrelevant and redundant channels may increase the computation burden and reduce the classification accuracy. We propose a new relative entropy-based dissimilarity metric for selecting channels with great contributions to the binary class classification. Since the ERD/ERS phenomenon is time localized, temporal segmentation is done at first, length of each time segment is equivalent to 1 s, and they are 50% overlapping. Then, for each channel, log-power $p_{\mathrm{ch},t}$ of each segment is calculated using the equation below.

$$p_{\mathrm{ch},t} = \log\left(\mathrm{var}\left(x_{\mathrm{ch},t}\right)\right) \quad \text{对数功率} \tag{1}$$

where $\mathbf{x}_{\mathrm{ch},t}$ is signal segment of time segment t of channel ch, var( ) denotes variance of the variable. Afterward, $p_{\mathrm{ch},t}$ of all trials in Ω are grouped according to MI classes. To select the most discriminative channels, the dissimilarity between two classes needs to be measured. We choose a dissimilarity metric that compares the probability distribution functions of $p_{\mathrm{ch},t}$, namely relative entropy [33]. Given two probability distributions, $\kappa^1(j)$ and $\kappa^2(j)$, the relative entropy is defined as

$$D\left(\kappa^1(j)|\kappa^2(j)\right) = \sum_j \kappa^1(j)\ln\left(\frac{\kappa^1(j)}{\kappa^2(j)}\right) \tag{2}$$

The relative entropy provides a nonnegative value which evaluates the dissimilarity between two distributions; a high relative entropy value means large dissimilarity. For the probability distribution of EEG trials, Gaussian model is usually adopted [3]. In this study, the Gaussian probability distributions of two classes are defined as

$$\kappa^1_{\mathrm{ch},t}(j) = \exp\left[\frac{-\left(p_{\mathrm{ch},t}(j) - \mu^1_{\mathrm{ch},t}\right)^2}{2\mathrm{var}^1_{\mathrm{ch},t}}\right] \tag{3}$$

两类高斯概率分布

$$\kappa^2_{\mathrm{ch},t}(j) = \exp\left[\frac{-\left(p_{\mathrm{ch},t}(j) - \mu^2_{\mathrm{ch},t}\right)^2}{2\mathrm{var}^2_{\mathrm{ch},t}}\right] \tag{4}$$

where $\mu^1_{\mathrm{ch},t}$ and $\mu^2_{\mathrm{ch},t}$ are means of $p_{\mathrm{ch},t}$ in two classes, $\mathrm{var}^1_{\mathrm{ch},t}$ and $\mathrm{var}^2_{\mathrm{ch},t}$ are corresponding variances, j denotes the index of a trial in Ω. Finally, the relative entropy criteria is defined as

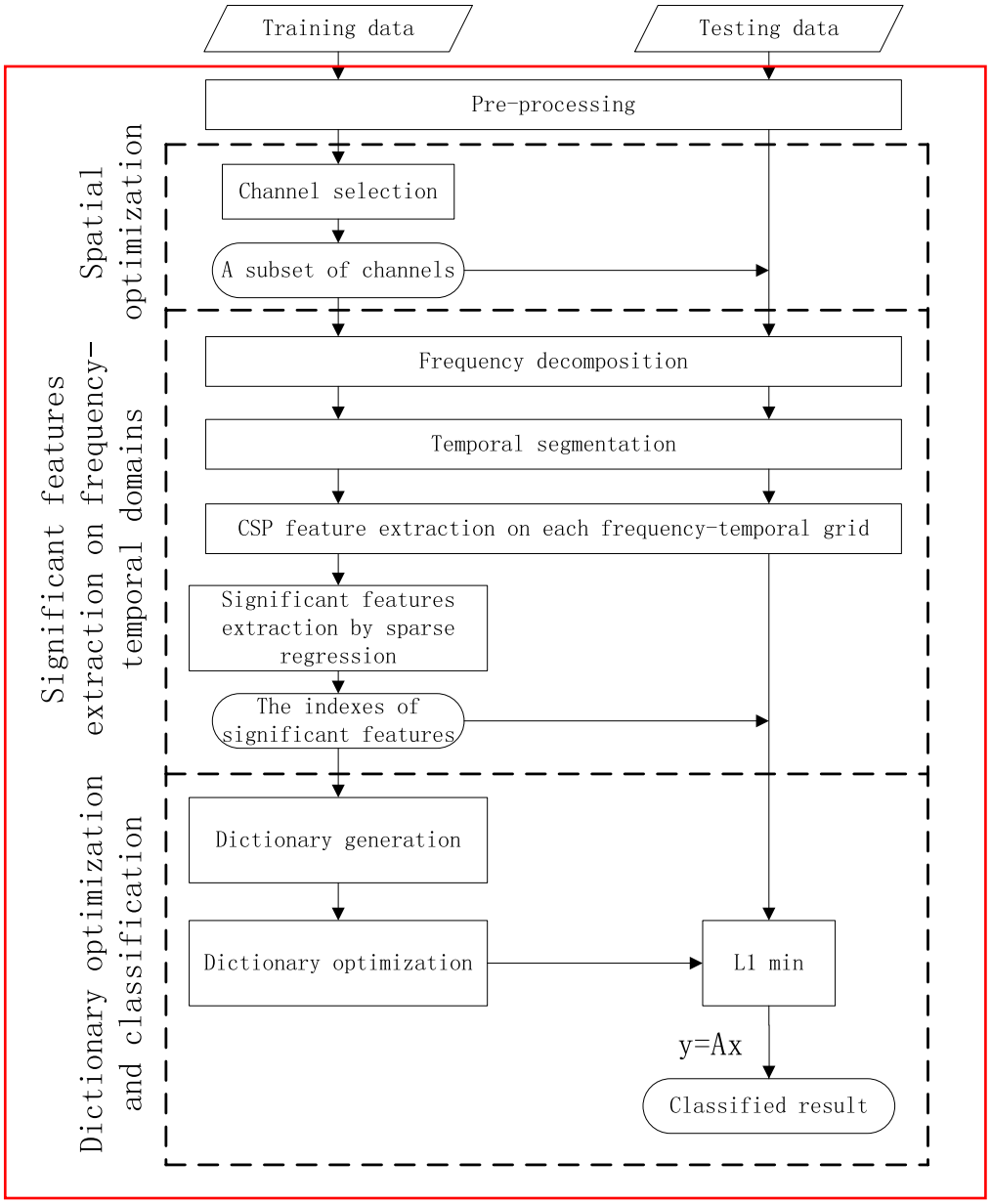

Fig. 1 Flowchart of proposed BCI system. A, x and y denotes the dictionary of SRC, coefficient vector and feature vector of a testing signal, respectively

通道ch的时间段t的相对熵值

$$\mathrm{RE}_{\mathrm{ch},t} = \sum_{j=1}^{S} \left[\kappa^1_{\mathrm{ch},t}(j) \ln \left(\frac{\kappa^1_{\mathrm{ch},t}(j)}{\kappa^2_{\mathrm{ch},t}(j)} \right) + \kappa^2_{\mathrm{ch},t}(j) \ln \left(\frac{\kappa^2_{\mathrm{ch},t}(j)}{\kappa^1_{\mathrm{ch},t}(j)} \right) \right] \tag{5}$$

where S is the total number of trials in Ω and $\mathrm{RE}_{\mathrm{ch},t}$ is the relative entropy value of time segment t of channel ch. Then, for each channel ch, the maximum relative entropy value of all time segments is reserved. The channels with high relative entropy values mean that they are class-discriminative, possibly improving classification performance. Therefore, it is natural to select channels with high relative entropy values. In this study, we sort the relative entropy values in descending order and the first M channels are selected.

2.2.3 *Significant features extraction on frequency-temporal domains*

After channel selection, the EEG is decomposed to a series of frequency-temporal components which cover different frequency and time ranges at local scale. The decomposition is achieved by first filtering the EEG to basic frequency bands and then segregating signal of each frequency band into basic time segments. In this study, totally 16 band-pass filters of uniform bandwidth 4 Hz are used to cover the frequency components from 6 to 40 Hz, and the overlapping between each other is 2 Hz. Temporal segmentation uses rectangular time windows, length of each time segment is 1 s, and the overlapping between each other is 0.5 s. In each frequency–time grid (time segment T of frequency band f),

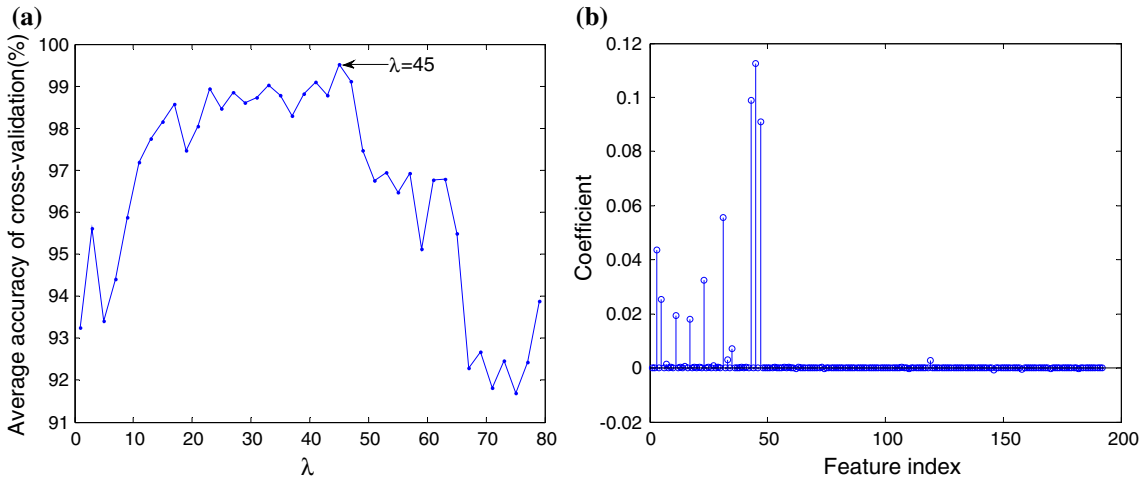

Fig. 2 Effect of varying λ on the average accuracy of cross-validation for subject "aa" (**a**) and the obtained sparse vector **v** corresponding to the optimal λ (**b**)

CSP is applied to Ω. We then obtain a spatial filter $\mathbf{W}^{f,T}$, where $\mathbf{W}^{f,T} \in R^{M\times M}$ denotes the filter of time segment T of frequency band f. Afterward, eigenvectors corresponding to the largest and smallest eigenvalues are selected to form an optimal spatial filter $\mathbf{F}^{f,T}$. A given EEG signal sample $\mathbf{X}^{f,T}$ is then transformed using $\mathbf{F}^{f,T}$ as follows

$$\mathbf{Z}^{f,T} = \mathbf{F}^{f,T} \times \mathbf{X}^{f,T} \tag{6}$$

Log-powers of $\mathbf{Z}^{f,T}$ are extracted as features. As described above, the EEG signal is decomposed into multiple frequency-temporal components. Two dimensional features of all components are then concatenated to form a feature vector **Q**. It should be noted that ERD/ERS phenomenon is typically frequency and time localized. In addition, ERD/ERS always occurs at different frequencies and during different time intervals for different subjects [11]. Therefore, it is natural to assume that the significances of features located in different components are different and the most discriminative features also vary among different subjects. In order to obtain the optimal feature subset for classification automatically as well as reduce the feature dimension, we propose to use sparse regularization which is a supervised method. Firstly, we construct a feature matrix $\mathbf{K} = \left[\mathbf{Q}'_1; \mathbf{Q}'_2; \ldots; \mathbf{Q}'_E\right]$, where E is the number of trials in Ω. The widely used Lasso estimate [32] is then adopted for significant features selection as follows

$$\mathbf{v} = \arg\min_{\mathbf{v}} \left(\frac{1}{2}\|\mathbf{K}\mathbf{v} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{v}\|_1\right) \tag{7}$$

where $\|\cdot\|_1$ denotes the l_1 norm, $\mathbf{y} \in R^E$ is a vector containing class labels {1, 2}, **v** is a sparse vector to be learned, λ is a positive regularization parameter for controlling the sparsity of **v**. The truncated Newton interior-point method [15] is adopted to solve the optimization problem in (Eq. 7). Note that λ plays an important role in feature selection and subsequent classification. The elements of vector **Q** corresponding to those zero values in **v** are excluded and the remainder is selected. A too large λ may remove useful features while a too small one cannot eliminate redundant information effectively. In this study, the optimal subject-specific λ is determined by 5 × 2-fold cross-validation [42] on Ω. Linear discriminant analysis (LDA) method is adopted for cross-validation due to its simplicity. Figure 2a presents an example to see the effect of varying λ on average accuracy of the 5 × 2-fold cross-validation for subject "aa." The optimal average accuracy is achieved at $\lambda = 45$ that is chosen for the subsequent testing procedure. Figure 2b shows the sparse vector **v** obtained when $\lambda = 45$.

2.2.4 *Dictionary optimization and classification*

Based on Ω, we define a component dictionary matrix $\mathbf{A}_i = [\mathbf{a}_{i,1}, \ldots, \mathbf{a}_{i,N_i}]$ for each class $i (i = 1, 2)$, where each column vector contains selected features and N_i is the total number of trials in class i. By concatenating the two matrices, the complete dictionary $\mathbf{A} = [\mathbf{A}1, \mathbf{A}2]$ is generated as shown in Fig. 3. Note that the EEG signals are noisy and non-stationary, there may be some abnormal samples in each class, i.e., the feature patterns of them are extremely different from the majority of samples in the same class. These abnormal samples may cause some incorrect representations and deteriorate the performance of SRC method. In this study, a probability-based method is used to eliminate these abnormal samples. For each component dictionary matrix $\mathbf{A}_i$, the mean of all column vectors is computed as $\mathbf{C}_i$, and then the Euclidean distance between $\mathbf{C}_i$ and $\mathbf{a}_{i,j}, j = 1, \ldots, N_i$ is calculated as follows

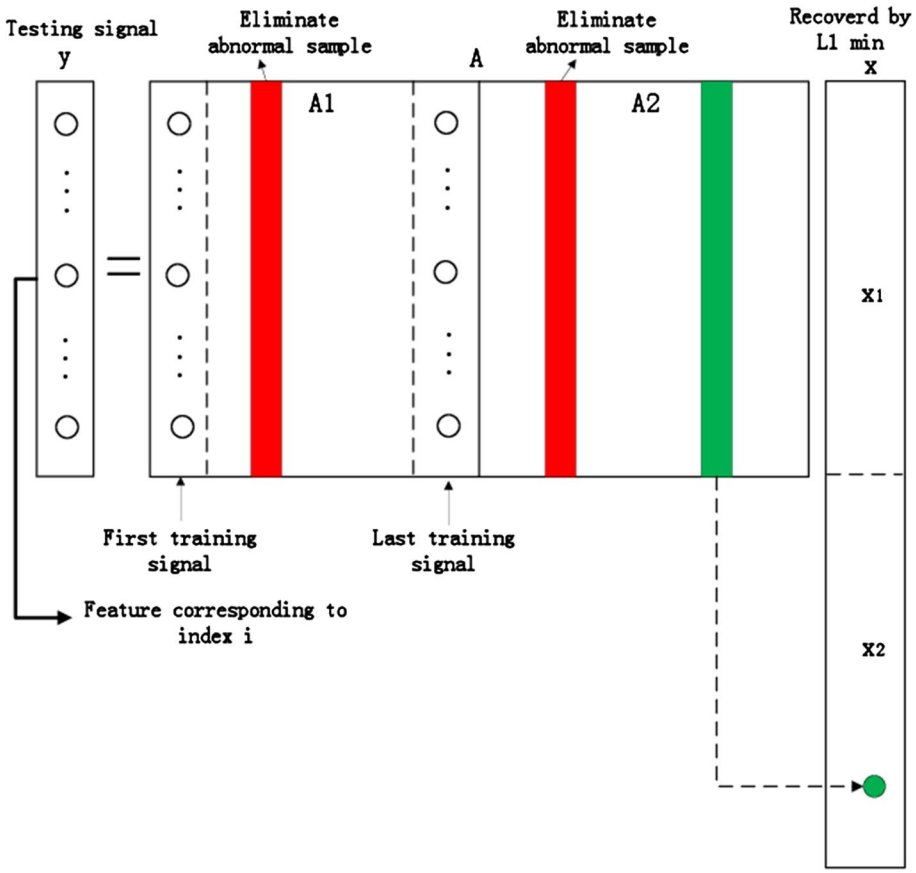

Fig. 3 Dictionary design and linear sparse representation model for SRC

$$D_i(j) = \left\|\mathbf{a}_{i,j} - \mathbf{C}_i\right\|_2, \quad j = 1, \ldots, N_i \tag{8}$$

The probability of $\mathbf{a}_{i,j}$ can be obtained by (Eq. 9).

$$P(\mathbf{a}_{i,j}) = \frac{\exp\left[\frac{-(D_i(j)-\mu_i)^2}{2\mathrm{var}_i}\right]}{\sqrt{2\pi\,\mathrm{var}_i}}, \quad j = 1, \ldots, N_i \tag{9}$$

Here, μ_i and var_i are the mean and variance of D_i. A small probability value means that the column vector is different from others in the same class. In this study, the column vectors whose probabilities are smaller than 5% are eliminated empirically. After dictionary optimization, a test signal $\mathbf{y}$ can be sparsely represented as a linear combination of the columns in $\mathbf{A}$ as follows

$$\mathbf{y} = \sum_{i=1,2} x_{i,1}\mathbf{a}_{i,1} + x_{i,2}\mathbf{a}_{i,2} + \cdots + x_{i,G_i}\mathbf{a}_{i,G_i} \tag{10}$$

where $x_{i,j}, j = 1, \ldots, G_i$ are scalar coefficients and G_i is the number of column vectors after dictionary optimization. The problem in (Eq. 10) is solved using L1 norm minimization [27] as:

$$\min\|\mathbf{x}\|_1 \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x} \tag{11}$$

Figure 4 presents an example to see how a testing signal is represented by column vectors in the dictionary for subject "al." We can see that a testing signal belonging to class 1 can be mainly represented by the column vectors in $\mathbf{A}1$ and a testing signal belonging to class 2 can be mainly represented by the column vectors in $\mathbf{A}2$. Finally, a residual-based rule [36] is used for classification in this study.

3 Results

In this section, we present our experimental results of the proposed method. All the experiments were implemented on a PC with Intel(R) Xeon(R) 2.40 GHz CPU and 8.0 GB RAM using MATLAB 2015a software.

3.1 Determining the optimal number of channels

We calculated the discriminative power of each channel defined by relative entropy for all subjects in dataset I, and the topographical distributions are displayed in Fig. 5. We observe that the channels with high discriminative powers are always located in neighboring areas of C3 and C4 electrodes. However, the spatial patterns of different subjects still vary a lot. Note that there is an unfixed parameter, namely the channel

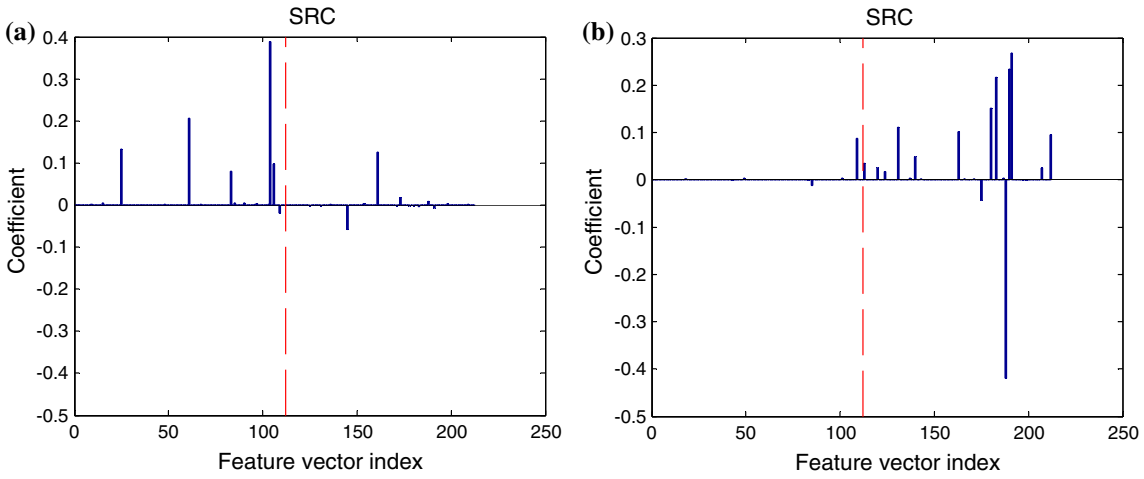

Fig. 4 Results of SRC for subject "al." Coefficient vector **x** of a testing signal belonging to class 1 (**a**). Coefficient vector **x** of a testing signal belonging to class 2 (**b**)

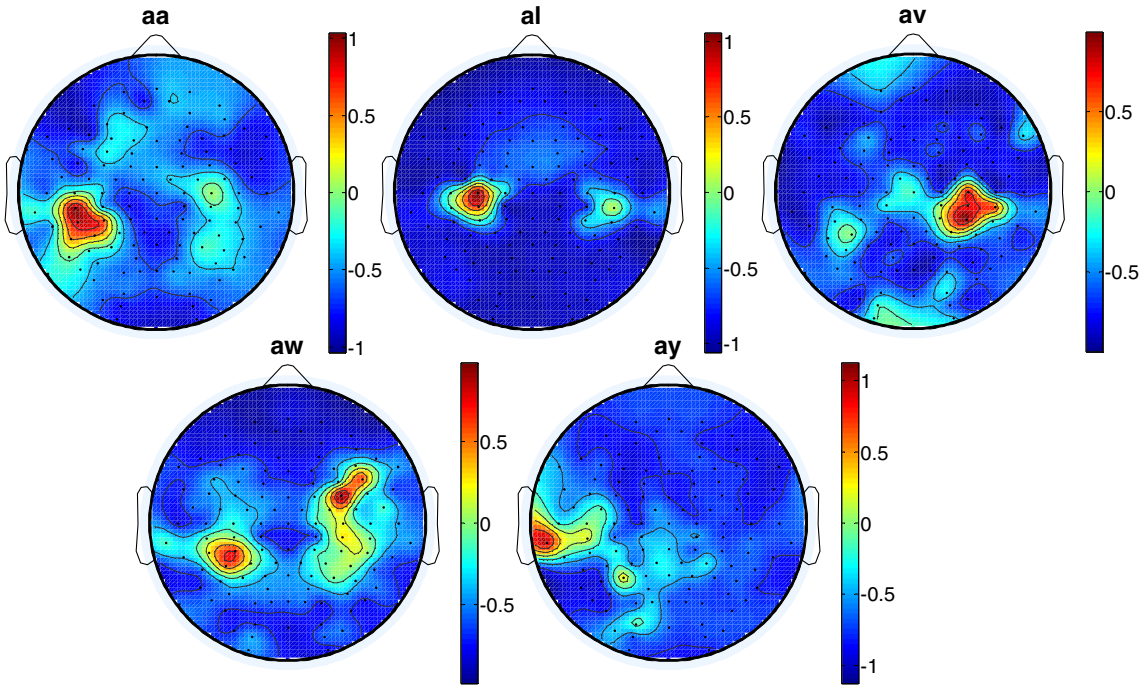

Fig. 5 Discriminant topographical maps of five subjects in dataset I. The *color bar* values are scaled from −1 to 1 (color figure online)

selection number M. To determine the optimal number of channels for subsequent processing, we also calculated the average classification accuracy of 5 × 2-fold cross-validation with M varying from 6 to 118 on Ω. The classification performance was evaluated by conventional CSP approach and LDA but with a varying number of common spatial filter vectors, ranging from 1 to 3. Figure 6 shows the effects of varying M on the classification performance for all subjects in dataset I. As can be seen in this figure, there is significant decrease or no significant increase for each subject when more than 20 channels are selected. Thus, we used the EEG data from these selected 20 channels for further feature extraction and classification. Note that dataset II was recorded at only three electrodes, so channel selection was not implemented.

3.2 Significant features on frequency-temporal domains

As described in Sect. 2.2.3, CSP method is applied on every frequency-temporal component and significant

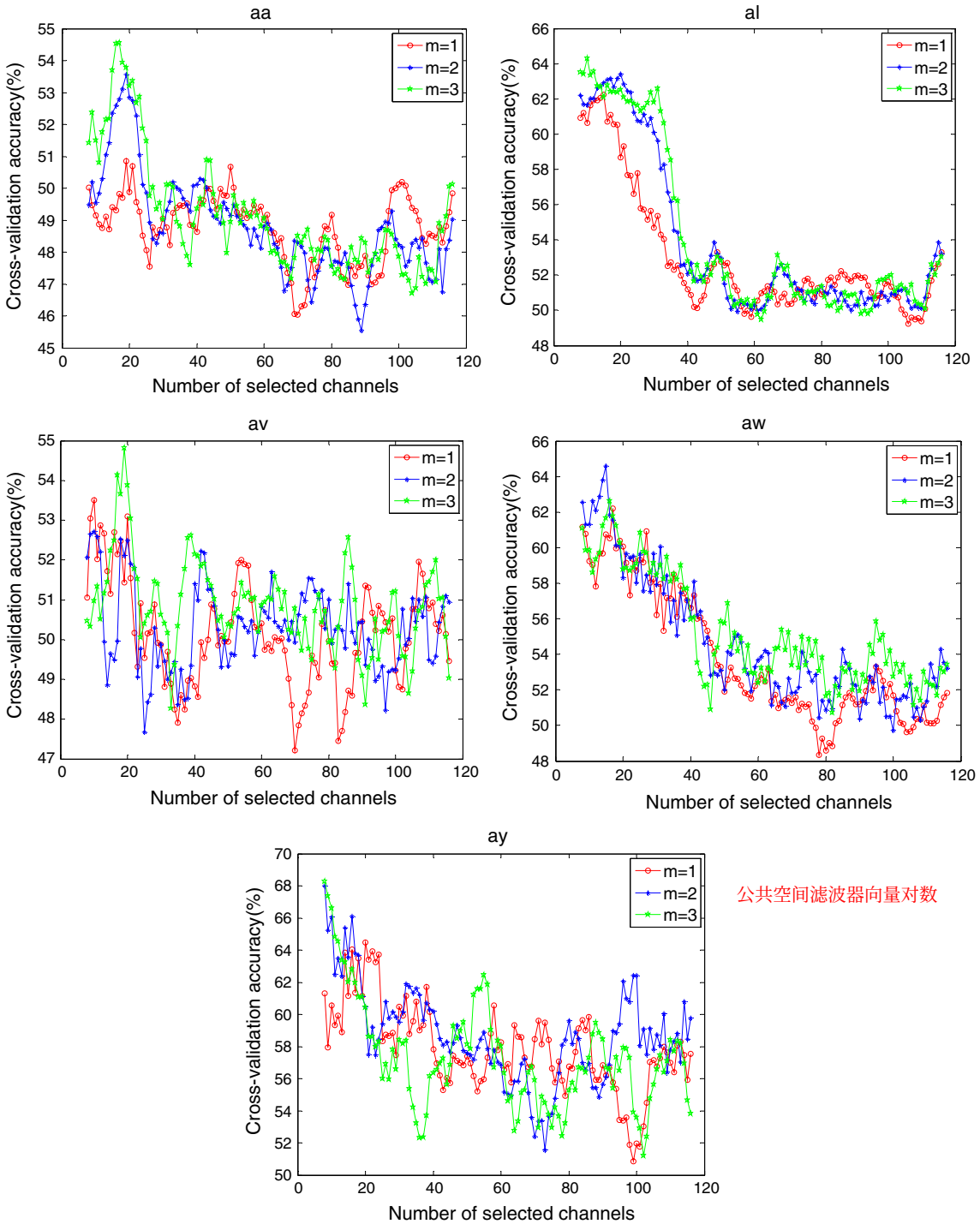

Fig. 6 Effects of varying channel selection number M on the average classification accuracies of 5 × 2-fold cross-validation for subjects in dataset I. m is the number of pairs of common spatial filter vectors. Moving window smoothing is applied to the results for each subject

features are selected by a supervised sparse regularization method. To have an intuitive observation of the distribution of significant frequency-temporal components, the sum of the absolute values of two sparse coefficients belonging to each frequency-temporal component was calculated for significance quantification. Figure 7 presents the distribution of significant frequency-temporal components for each subject in dataset I. Figure 8 shows the distribution of significant frequency-temporal components for each subject in dataset II. Note that 10 × 10-fold cross-validation was

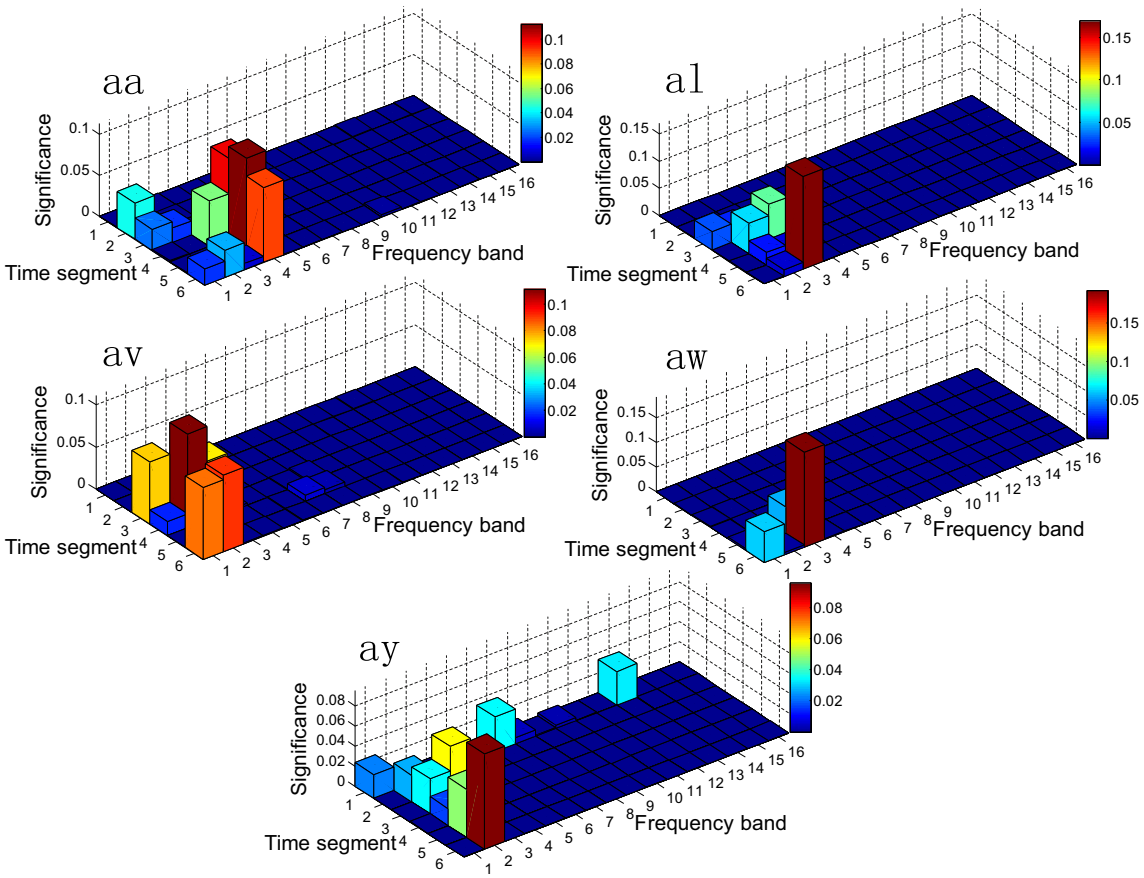

Fig. 7 Distribution of significant frequency-temporal components for each of the five subjects in dataset I. Frequency bands indicated as 1, 2,…, 16 correspond to filter sub-bands 6–10, 8–12,…, 36–40 Hz. Time segments indicated as 1, 2,…, 6 correspond to the temporal windows 0–1, 0.5–1.5,…, 2.5–3.5 s. Different *colors* correspond to different values of significance (color figure online)

implemented for dataset II, the sparse vector corresponding to the best classification accuracy was used in Fig. 8. It can be seen from Fig. 7 that, for all the subjects in dataset I, most significant features are located in 6–16 Hz frequency band which covers mu rhythm. It has been observed from Fig. 8 that, for all the subjects in dataset II, most significant features are located in 8–14 and 16–30 Hz which cover both mu and beta rhythms. It is also noticeable from both figures that the distribution of significant frequency-temporal components is sparse and subject-specific.

3.3 Classification results of dataset I

Table 1 gives the classification accuracies derived by CSP [6], CSSP [16], FBCSP [1], DFBCSP [31], OSSFN [41], BSSFO [30], SRC [27] and our proposed SFTOFSRC, respectively, for the five subjects in dataset I. Apart from SRC and SFTOFSRC, the results of other competing methods have already been reported in the literature [30]. The proposed SFTOFSRC method outperformed SRC significantly over all subjects, and the average classification accuracy improvement was 21.568%. In addition, SFTOFSRC method further yielded higher average accuracy than those of other competing methods. The average classification accuracy improvements achieved by SFTOFSRC were 14.328, 2.518, 15.768, 15.088, 9.138, 9.658 and 0.658% in comparison with the CSP, CSSP, FBCSP, DFBCSP, OSSFN + FBCSP, OSSFN + DFBCSP and BSSFO methods. Table 2 presents the classification accuracies obtained by SFTOFSRC with and without dictionary optimization for dataset I. As observed from this table, the accuracies of all subjects after employing dictionary optimization have significantly improved and the average improvement reached 3.22%.

3.4 Classification results of dataset II

Table 3 summarizes classification accuracies for the nine subjects in dataset II. Apart from SRC and SFTOFSRC, the results of other competing methods have already been reported in the literature [42]. The classification performance of SFTOFSRC was significantly better than that

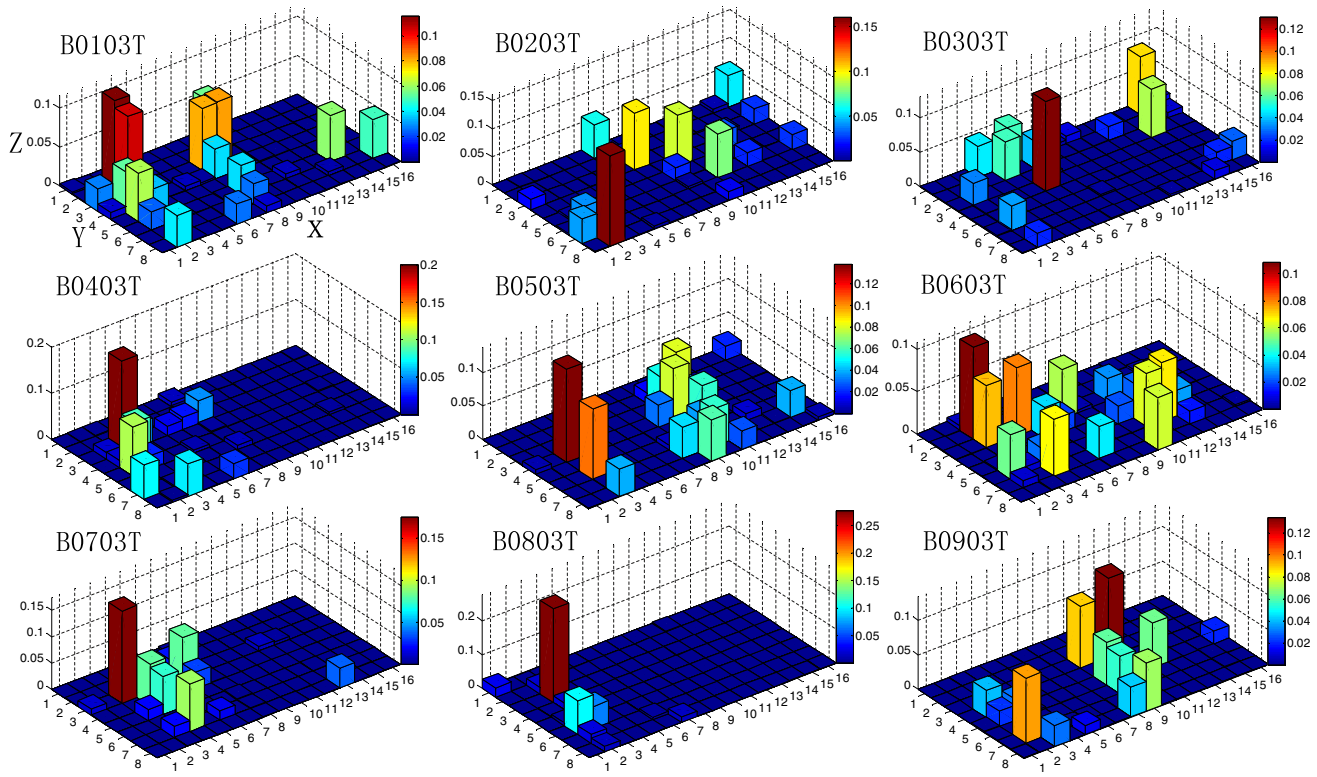

Fig. 8 Distribution of significant frequency-temporal components for each of the nine subjects in dataset II. *X* labels indicated as 1, 2,…, 16 correspond to filter sub-bands 6–10, 8–12,…, 36–40 Hz. *Y* labels indicated as 1, 2,…, 8 correspond to the temporal windows 0–1, 0.5–1.5,…, 3.5–4.5. *Z* labels denote significances. Different *colors* correspond to different values of significance (color figure online)

Table 1 Classification accuracies (%) obtained by CSP, CSSP, FBCSP, DFBCSP, OSSFN, BSSFO, SRC and our proposed SFTOFSRC method for dataset I

Method	aa	al	av	aw	ay	Mean	SD
CSP	66.96	89.29	52.55	47.77	52.38	61.79	16.98
CSSP	79.46	92.86	52.55	91.52	51.59	73.60	20.33
FBCSP	69.64	80.36	47.96	55.36	48.41	60.35	14.21
DFBCSP	69.64	82.14	54.08	50.89	48.41	61.03	14.41
OSSFN + FBCSP	75.00	83.93	53.06	74.11	48.81	66.98	15.22
OSSFN + DFBCSP	75.00	83.93	52.05	74.11	47.22	66.46	15.93
BSSFO	79.46	94.64	57.65	**91.96**	**53.57**	75.46	19.06
SRC	46.43	84.29	45.41	48.21	48.41	54.55	16.67
SFTOFSRC	**81.25**	**100**	**65.31**	81.25	52.78	**76.118**	17.92

All accuracies are obtained by calculating the ratio of trials correctly classified to the total number of trials in testing data ψ. The highest accuracy is marked in boldface; SD means standard deviation

Table 2 Classification accuracies (%) obtained by SFTOFSRC with dictionary optimization and SFTOFSRC without dictionary optimization for dataset I

Method	aa	al	av	aw	ay	Mean	SD
SFTOFSRC with DO	**81.25**	**100**	**65.31**	**81.25**	**52.78**	**76.118**	17.92
SFTOFSRC without DO	80.36	94.64	61.22	79.46	48.81	72.898	17.94

All accuracies are obtained by calculating the ratio of trials correctly classified to the total number of trials in testing data ψ. The higher accuracy is marked in boldface; DO means dictionary optimization and SD means standard deviation

Table 3 Classification accuracies (%) obtained by CSP, FBCSP, DFBCSP, SFBCSP, SRC and our proposed SFTOFSRC method for dataset II

Subject	CSP	FBCSP	DFBCSP	SFBCSP	SRC	SFTOFSRC
B0103T	76.56 ± 5.88	77.50 ± 4.61	77.94 ± 4.48	78.15 ± 4.79	69.43 ± 5.28	**83.83** ± 5.29
B0203T	55.56 ± 6.97	55.94 ± 6.05	57.25 ± 5.89	58.75 ± 5.84	53 ± 6.47	**62.61** ± 6.99
B0303T	52.62 ± 8.10	53.75 ± 6.95	55.19 ± 6.50	55.81 ± 6.09	51.24 ± 5.33	**61.32** ± 6.09
B0403T	98.06 ± 1.50	**98.87** ± 1.13	98.81 ± 1.43	98.85 ± 1.23	93.15 ± 1.83	98.11 ± 1.23
B0503T	88.19 ± 4.77	90.44 ± 2.42	**92.44** ± 2.14	92.06 ± 2.35	81.23 ± 5.47	91.91 ± 4.24
B0603T	69.37 ± 5.60	78.06 ± 3.44	81.69 ± 3.17	82.32 ± 3.62	64.15 ± 3.84	**85.78** ± 3.72
B0703T	83.44 ± 4.85	86.50 ± 2.90	89.50 ± 1.83	90.25 ± 1.25	76.96 ± 2.25	**94.39** ± 3.79
B0803T	86.56 ± 3.02	88.75 ± 2.78	88.62 ± 2.93	88.87 ± 3.21	80.03 ± 2.93	**93.26** ± 4.37
B0903T	81.75 ± 4.63	83.88 ± 3.39	84.75 ± 3.33	85.50 ± 3.61	79.05 ± 2.24	**91.14** ± 2.14
Average	76.90 ± 5.04	79.30 ± 3.74	80.69 ± 3.52	81.17 ± 3.55	72.03 ± 3.96	**86.41** ± 4.21

All accuracies are achieved with 10 × 10-fold cross-validation of the 160 samples of each subject, so results are presented in "Mean + SD" form. The highest mean accuracy is marked in boldface

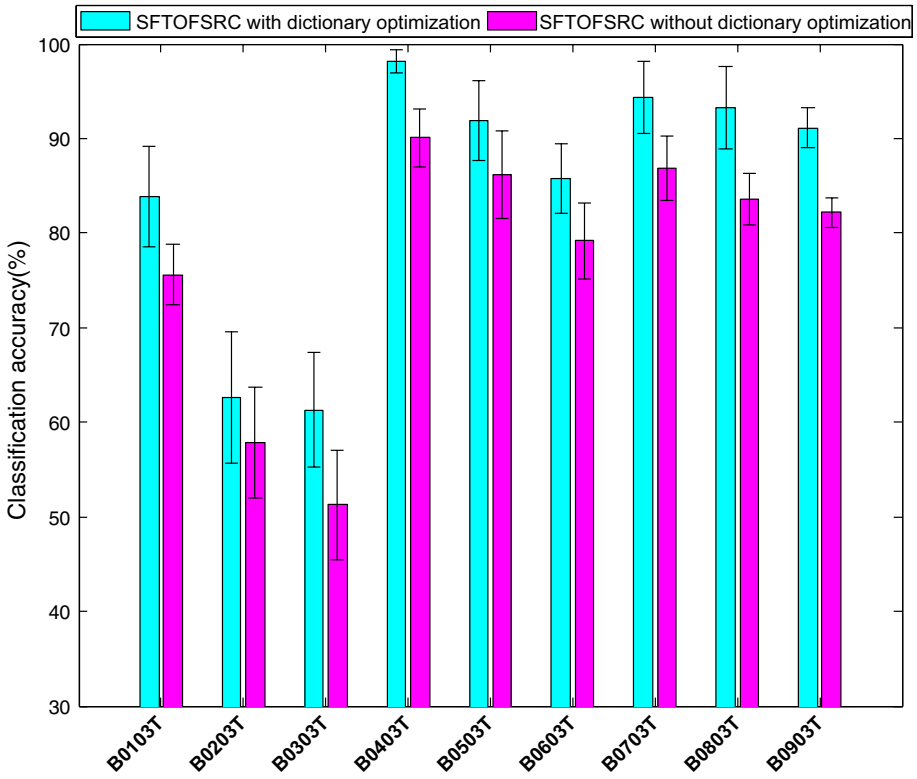

Fig. 9 Comparison of classification accuracies (%) obtained by SFTOFSRC with dictionary optimization and SFTOFSRC without dictionary optimization for dataset II

of SRC, and the average classification accuracy improvement was 14.38%. Furthermore, the average classification accuracy improvements achieved by SFTOFSRC were 9.51, 7.11, 5.72 and 5.24% in comparison with the CSP, FBCSP, DFBCSP and SFBCSP methods. Figure 9 shows classification accuracies obtained by SFTOFSRC with and without dictionary optimization for all the subjects in dataset II. As can be seen, the accuracies of all subjects after employing dictionary optimization have significantly improved and the average improvement reached 9.4389%.

4 Discussion

In the existing SRC method [27], CSP algorithm is used to construct the dictionary and a good dictionary is critical for effective classification. CSP has proven to be very useful in extracting ERS/ERD-related features. However, ERD/ERS phenomenon which appears in a specific brain region is typically frequency and time localized. Hence some steps like channel selection, band-pass filtering and time interval selection are necessary and have great effect on final CSP features extraction. Furthermore, due to the individual

distinctions in physiology, anatomy and brain state, the discriminative frequency bands, temporal segments of maximum separability and channels with high discriminative powers are subject specific and can hardly be determined manually. A simple manual setting in work [27] is not optimal for CSP features extraction. To overcome the drawback of SRC and design a more incoherent dictionary, we propose to optimize CSP features in subject-adapted space–frequency–time patterns by two data-driven approaches.

In spatial optimization, subject-specific optimal channels are selected based on our proposed relative entropy criteria. As shown in Fig. 5, the active channels are always located over primary motor cortex which has been long considered as the best position for recording prominent change of mu or beta rhythm [35]. However, the topography of subject "ay" seems to be abnormal compared to the other four subjects. The reasons may be twofold as follows: (1) The ERD/ERS rhythm activity is contaminated by strong artifact which leads to focus around electrode "TTP9h." (2) Subject "ay" has an extremely small training dataset which contains only 10 right-foot samples and 18 right-hand samples, Gaussian model can hardly be estimated well with such few training samples, thus the channels selected by relative entropy criteria do not show enough neurophysiological plausibility. We notice that some predefined channel configurations have been used for sensorimotor rhythms (SMR)-related features extraction in previous works [21, 26]. In the literature [21], optimal average performance of BCI competition III dataset IVa was obtained when channel configuration 7 (see Fig. 10) was adopted. We also used this fixed channel configuration for subject "ay" and the same classification accuracy 52.78% was obtained. From our point of view, most discriminative channels such as "C5," "C3," "CP5," "CP3" and "P3" are still selected automatically by our method for subject "ay," though the pattern is a bit abnormal. Figure 6 illustrates the importance of not using all the channels for CSP feature extraction and classification, as the non-event-related channels add ambiguous information to the data, thus resulting in separability attenuation of different classes. Therefore, selecting channels with high discriminative powers can help to improve classification performance and reduce computation burden.

In work [1, 40, 42], significant features are extracted from multiple frequency bands. However, the frequency band and time interval have the interacted effect on each other and the mere optimization of frequency band without the simultaneous consideration of time interval effect may fail to find the optimal features for classification. Inspired by their works, we further extend the idea to frequency–time domains. From Figs. 7 and 8, we can see that the automatically selected features for classifying limb MIs are mainly located in mu and beta rhythms. In addition, due to the non-stationary characteristic of EEG signal, the most useful features are not located in an entire frequency band but some local frequency-temporal components. We believe that decomposing a multi-channel EEG into multiple frequency-temporal components for more precise analysis helps to improve the classification performance.

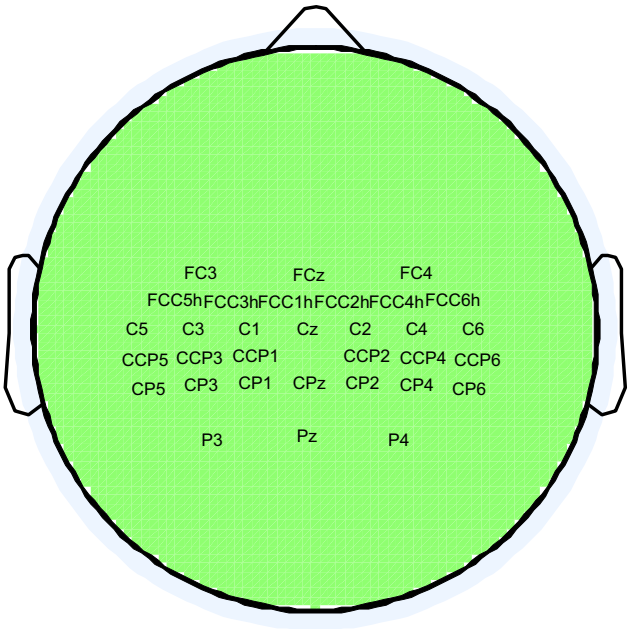

Fig. 10 Predefined 32 channel configuration

As shown in Tables 1 and 3, our proposed method outperforms the existing SRC significantly. The main reason is that selecting significant channels and features from multiple frequency-temporal components in two data-driven manners can improve the effect of CSP filtering. As we know, the power of the CSP filtered signal is used to construct the dictionary, so optimizing features in subject-adapted space–frequency–time patterns helps to design a more incoherent dictionary for classification. Table 2 and Fig. 9 illustrate that dictionary optimization can further improve the classification accuracy when abnormal samples are involved in the component dictionary. For dataset I, the classification accuracies presented in [27] are much better than our method except subject "al." Note that leave-one-out (LOO) cross-validation method was used in their work to tackle the problem of small training data size. So the dictionary in their work included more features of training samples and was more balanced for two classes. Such dictionary design is more suitable for SRC. However, in order to test our algorithm's ability to cope with a small training data size and compare the results against other studies, the ratio of trials correctly classified to the total number of testing trials for each subject is used for evaluation. The reason why subject "al" which has 112 right-hand trials and 112 right-foot trials in the training

data still achieves high accuracy is that the dictionary also contains sufficient samples and is balanced. According to Tables 1 and 3, SFTOFSRC shows better performance when compared to other methods which are also devoted to finding ERD/ERS-related channels, frequency bands or time intervals for MI classification. These results indicate that our proposed method is competitive among similar methods. However, it should be noted that the classification results of dataset I shown in Table 1 are still significantly lower than the winner of the BCI competition and some other researches like [21]. The reasons can be concluded as follows: (1) We do not take any adaptation or extension of training data by classified testing samples which is adopted by the BCI competition winners to tackle the problem of the small training data size. (2) As described above, a balanced dictionary with sufficient training samples is more suitable for SRC, while in this study few training samples with unbalanced sample sizes are used for dictionary generation, this is also the main reason why the classification results of subject "ay" are very low. Hence, optimizing the dictionary of SRC by extending small training dataset would be our forthcoming research issue to further improve the classification performance.

Besides classification accuracy, another important issue is speed of the algorithm. In our proposed method, discriminative channels, significant features on frequency-temporal domains and SRC classifier are obtained from training data. After training, these information can be directly used to the testing data. For a testing sample, features needs to be extracted on every frequency-temporal component. We compared the average execution times of one testing trial using SRC and SFTOFSRC. For subject "aa," SRC took 29.3 ms and SFTOFSRC took 47.3 ms. Since CSP algorithm needs to be implemented on each frequency–time component, the computational burden of SFTOFSRC is obviously heavier than SRC which can be treated as a trade-off between accuracy and computation efficiency. However, CSP algorithm is known to be quite computationally efficient [37]. We believe that such computational load would not affect the online performance of BCI study as the computing efficiency of the computer is getting better.

5 Conclusion

In this study, we proposed a new SFTOFSRC method for MI EEG pattern recognition. Firstly, optimal channels were selected based on our proposed relative entropy criteria. Then, significant features on frequency-temporal domains were selected automatically by sparse regularization. Afterward, selected features from multiple frequency–time components were concatenated to form a column vector for SRC. Finally, a probability-based method was used to optimize the dictionary. Its accuracy improvement over SRC, comprehensive consideration on spatial-frequency-temporal domains and capability to adapt to different subjects make the proposed approach a competitive candidate for future BCI system.

Acknowledgements Special thanks to the editors and anonymous reviewers for their positive and constructive comments and suggestions on our manuscript. The study was supported by the Jiangsu Province Science and Technology Support Program of China (No. BE2012740).

References

1. Ang KK, Chin ZY, Wang CC, Guan CT, Zhang HH (2012) Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. Front Neurosci. doi:10.3389/fnins.2012.00039
2. Ang KK, Chin ZY, Zhang HH, Guan CT (2012) Mutual information-based selection of optimal spatial-temporal patterns for single-trial EEG-based BCIs. Pattern Recogn 45:2137–2144. doi:10.1016/j.patcog.2011.04.018
3. Arvaneh M, Guan CT, Ang KK, Quek C (2013) Optimizing spatial filters by minimizing within-class dissimilarities in electroencephalogram-based brain–computer interface. IEEE Trans Neural Netw Learn Syst 24:610–619. doi:10.1109/tnnls.2013.2239310
4. Bhattacharyya S, Konar A, Tibarewala DN (2014) Motor imagery, P300 and error-related EEG-based robot arm movement control for rehabilitation purpose. Med Biol Eng Compu 52:1007–1017. doi:10.1007/s11517-014-1204-4
5. Blankertz B, Muller KR, Krusienski DJ, Schalk G, Wolpaw JR, Schlogl A, Pfurtscheller G, Millan JDR, Schroder M, Birbaumer N (2006) The BCI competition III: validating alternative approaches to actual BCI problems. IEEE Trans Neural Syst Rehabil Eng 14:153–159. doi:10.1109/tnsre.2006.875642
6. Blankertz B, Tomioka R, Lemm S, Kawanabe M, Muller KR (2008) Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Process Mag 25:41–56. doi:10.1109/msp.200790.900,9
7. Der G, Deary IJ (2006) Age and sex differences in reaction time in adulthood: results from the United Kingdom Health and Lifestyle Survey. Psychol Aging 21:62–73. doi:10.1037/0882-7974.21.1.62
8. Dornhege G, Blankertz B, Krauledat M, Losch F, Curio G, Muller KR (2006) Combined optimization of spatial and temporal filters for improving brain–computer interfacing. IEEE Trans Biomed Eng 53:2274–2281. doi:10.1109/tbme.2006.883649
9. Gemmeke JF, Virtanen T, Hurmalainen A (2011) Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans Audio Speech Lang Process 19:2067–2080. doi:10.1109/tasl.2011.2112350
10. Hadjidimitriou S, Zacharakis A, Doulgeris P, Panoulas K, Hadjileontiadis L, Panas S (2010) Sensorimotor cortical response during motion reflecting audiovisual stimulation: evidence from fractal EEG analysis. Med Biol Eng Compu 48:561–572. doi:10.1007/s11517-010-0606-1
11. Hamedi M, Salleh SH, Noor AM (2016) Electroencephalographic motor imagery brain connectivity analysis for BCI: a review. Neural Comput 28:999–1041. doi:10.1162/NECO_a_00838
12. Higashi H, Tanaka T (2013) Common spatio–time–frequency patterns for motor imagery-based brain machine interfaces. Comput Intell Neurosci. doi:10.1155/2013/537218

13. Ince NF, Goksu F, Tewfik AH, Arica S (2009) Adapting subject specific motor imagery EEG patterns in space–time–frequency for a brain computer interface. Biomed Signal Process Control 4:236–246. doi:10.1016/j.bspc.2009.03.005
14. Kee CY, Ponnambalam SC, Loo CK (2015) Multi-objective genetic algorithm as channel selection method for P300 and motor imagery data set. Neurocomputing 161:120–131. doi:10.1016/j.neucom.2015.02.057
15. Kim SJ, Koh K, Lustig M, Boyd S, Gorinevsky D (2007) An interior-point method for large-scale l(1)-regularized least squares. IEEE J Sel Top Signal Process 1:606–617. doi:10.1109/jstsp.2007.910971
16. Lemm S, Blankertz B, Curio G, Muller KR (2005) Spatio-spectral filters for improving the classification of single trial EEG. IEEE Trans Biomed Eng 52:1541–1548. doi:10.1109/tbme.2005.851521
17. Li MA, Guo SD, Yang JF, Sun YJ (2016) A novel EEG feature extraction method based on OEMD and CSP algorithm. J Intell Fuzzy Syst 30:2971–2983. doi:10.3233/ifs-151896
18. Mallat SG, Zhifeng Z (1993) Matching pursuits with time–frequency dictionaries. IEEE Trans Signal Process 41:3397–3415. doi:10.1109/78.258082
19. McFarland DJ, McCane LM, David SV, Wolpaw JR (1997) Spatial filter selection for EEG-based communication. Electroencephalogr Clin Neurophysiol 103(3):386–394. doi:10.1016/s0013-4694(97)00022-2
20. Mei X, Ling HB (2011) Robust visual tracking and vehicle classification via sparse representation. IEEE Trans Pattern Anal Mach Intell 33:2259–2272. doi:10.1109/tpami.2011.66
21. Meng JJ, Huang G, Zhang DG, Zhu XY (2013) Optimizing spatial spectral patterns jointly with channel configuration for brain–computer interface. Neurocomputing 104:115–126. doi:10.1016/j.neucom.2012.11.004
22. Muller-Putz GR, Scherer R, Pfurtscheller G, Rupp R (2005) EEG-based neuroprosthesis control: a step towards clinical practice. Neurosci Lett 382:169–174. doi:10.1016/j.neulet.2005.03.021
23. Pfurtscheller G, Neuper C (1997) Motor imagery activates primary sensorimotor area in humans. Neurosci Lett 239:65–68. doi:10.1016/s0304-3940(97)00889-6
24. Pfurtscheller G, Neuper C, Ramoser H, Muller-Gerking J (1999) Visually guided motor imagery activates sensorimotor areas in humans. Neurosci Lett 269:153–156. doi:10.1016/s0304-3940(99)00452-8
25. Satti AR, Coyle D, Prasad G, IEEE (2009) Spatio-spectral and temporal parameter searching using class correlation analysis and particle swarm optimization for a brain computer interface. In: 2009 IEEE international conference on systems, man and cybernetics. IEEE international conference on systems man and cybernetics conference proceedings. pp 1731–1735. doi:10.1109/icsmc.2009.5346679
26. Sannelli C, Dickhaus T, Halder S, Hammer E-M, Mueller K-R, Blankertz B (2010) On optimal channel configurations for SMR-based brain–computer interfaces. Brain Topogr 23(2):186–193. doi:10.1007/s10548-010-0135-0
27. Shin Y, Lee S, Lee J, Lee HN (2012) Sparse representation-based classification scheme for motor imagery-based brain–computer interface systems. J Neural Eng. doi:10.1088/1741-2560/9/5/056002
28. Shin Y, Lee S, Ahn M, Cho H, Jun SC, Lee HN (2015) Noise robustness analysis of sparse representation based classification method for non-stationary EEG signal classification. Biomed Signal Process Control 21:8–18. doi:10.1016/j.bspc.2015.05.007
29. Shin Y, Lee S, Ahn M, Cho H, Jun SC, Lee HN (2015) Simple adaptive sparse representation based classification schemes for EEG based brain–computer interface applications. Comput Biol Med 66:29–38. doi:10.1016/j.compbiomed.2015.08.017
30. Suk HI, Lee SW (2013) A novel Bayesian framework for discriminative feature extraction in brain–computer interfaces. IEEE Trans Pattern Anal Mach Intell 35:286–299. doi:10.1109/tpami.2012.69
31. Thomas KP, Guan CT, Lau CT, Vinod AP, Ang KK (2009) A new discriminative common spatial pattern method for motor imagery brain–computer interfaces. IEEE Trans Biomed Eng 56:2730–2733. doi:10.1109/tbme.2009.2026181
32. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol 58:267–288
33. Vu VQ, Yu B, Kass RE (2009) Information in the nonstationary case. Neural Comput 21:688–703. doi:10.1162/neco.2008.01-08-700
34. Wei QG, Wei ZH (2015) Binary particle swarm optimization for frequency band selection in motor imagery based brain–computer interfaces. Bio-Med Mater Eng 26:S1523–S1532. doi:10.3233/bme-151451
35. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM (2002) Brain–computer interfaces for communication and control. Clin Neurophysiol 113:767–791. doi:10.1016/s1388-2457(02)00057-3
36. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y (2009) Robust face recognition via sparse representation. IEEE Trans Pattern Anal Mach Intell 31:210–227. doi:10.1109/tpami.2008.79
37. Wu B, Yang F, Zhang JC, Wang YW, Zheng XX, Chen WD (2012) A frequency-temporal-spatial method for motor-related electroencephalography pattern recognition by comprehensive feature optimization. Comput Biol Med 42:353–363. doi:10.1016/j.compbiomed.2011.11.014
38. Xu Q, Zhou H, Wang YJ, Huang J (2009) Fuzzy support vector machine for classification of EEG signals using wavelet-based features. Med Eng Phys 31:858–865. doi:10.1016/j.medengphy.2009.04.005
39. Xu P, Liu TJ, Zhang R, Zhang YS, Yao DZ (2014) Using particle swarm to select frequency band and time interval for feature extraction of EEG based BCI. Biomed Signal Process Control 10:289–295. doi:10.1016/j.bspc.2013.08.012
40. Novi Q, Guan C, Dat TH, Xue P, IEEE (2007) Sub-band common spatial pattern (SBCSP) for brain–computer interface. In: 2007 3rd International IEEE/EMBS conference on neural engineering, vols 1 and 2. doi:10.1109/cne.2007.369647
41. Zhang HH, Chin ZY, Ang KK, Guan CT, Wang CC (2011) Optimum spatio-spectral filtering network for brain–computer interface. IEEE Trans Neural Networks 22:52–63. doi:10.1109/tnn.2010.2084099
42. Zhang Y, Zhou GX, Jin J, Wang XY, Cichocki A (2015) Optimizing spatial patterns with sparse filter bands for motor-imagery based brain–computer interface. J Neurosci Methods 255:85–91. doi:10.1016/j.jneumeth.2015.08.004

Minmin Miao is currently pursuing his Ph.D. degree in School of Instrument Science and Engineering, Southeast University. His research interests include brain–computer interface, signal processing and pattern recognition.

Aimin Wang is now working as a professor in School of Instrument Science and Engineering, Southeast University. His research interests include precision instrument and machinery, and intelligent control.

Feixiang Liu is now doing M.S. research in School of Instrument Science and Engineering, Southeast University. His research interests include brain–computer interface and pattern recognition.