Exploiting Sparse Representation in the P300 Speller Paradigm

Hongma Liu
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

Yali Li
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

Shengjin Wang
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

Follow this and additional works at: https://dc.tsinghuajournals.com/tsinghua-science-and-technology

Recommended Citation
Exploiting Sparse Representation in the P300 Speller Paradigm

Hongma Liu*, Yali Li, and Shengjin Wang

Abstract: A Brain-Computer Interface (BCI) aims to produce a new way for people to communicate with computers. Brain signal classification is a challenging issue owing to the high-dimensional data and low Signal-to-Noise Ratio (SNR). In this paper, a novel method is proposed to cope with this problem through sparse representation for the P300 speller paradigm. This work is distinguished using two key contributions. First, we investigate sparse coding and its feasibility for brain signal classification. Training signals are used to learn the dictionaries and test signals are classified according to their sparse representation and reconstruction errors. Second, sample selection and a channel-aware dictionary are proposed to reduce the effect of noise, which can improve performance and enhance the computing efficiency simultaneously. A novel classification method from the sample set perspective is proposed to exploit channel correlations. Specifically, the brain signal of each channel is classified jointly using its spatially neighboring channels and a novel weighted regulation strategy is proposed to overcome outliers in the group. Experimental results have demonstrated that our methods are highly effective. We achieve a state-of-the-art recognition rate of 72.5%, 88.5%, and 98.5% at 5, 10, and 15 epochs, respectively, on BCI Competition III Dataset II.

Key words: sparse representation; sample selection; channel-aware dictionary; P300 speller

1 Introduction

A Brain-Computer Interface (BCI) is a direct interactive pathway between a human brain and an external computer. It allows the user to send commands to a machine without body movement, which is specifically useful for disabled people[1]. In a BCI system, the brain signals of the user are recorded with electroencephalography (EEG) recorder and then magnified and preprocessed to reduce the noise effect. Next, the signals are converted to commands via classification, and finally, the commands are sent to a device for operation[2]. Research on BCI is a fast-growing field since it is a revolutionary breakthrough in human-computer interfacing and can improve our understanding of the human brain. A portable and cheap EEG recorder, such as Emotive, is suitable for daily research. Figure 1 illustrates the typical electrode layout of a 64-channel EEG recorder.

Two types of technologies are used for monitoring brain activity, invasive and noninvasive methods. Our research mainly focuses on noninvasive methods because they allow brain signals to be recorded more easily and do not harm the subjects. EEG is the most widely used noninvasive method owing to its convenience for

*Hongma Liu, Yali Li, and Shengjin Wang are with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China. E-mail: liuhongma.thu@hotmail.com; liyali13@tsinghua.edu.cn; wgsgj@tsinghua.edu.cn.

*To whom correspondence should be addressed.

Manuscript received: 2019-12-01; accepted: 2019-12-28
conducting experiments and lower cost.

Many Event-Related Potentials (ERPs) have been discovered and studied. Of all the ERPs, we are interested in P300 because it is thought to reflect neuro-electric activity related to cognitive processes, such as attention allocation and activation of immediate memory[3]; moreover, P300 has been proven subject-independent and intense across various people. Furthermore, P300 is practically applicable, such that it can help a subject to spell a word using brain signals, which is exciting for future use and can greatly change the way people interact with computers. It represents the natural response of humans to some specific external stimuli. The stimulus should follow the oddball paradigm where expected stimulus is a rare event among all stimuli[4]. In that case, the expected stimulus produces a positive deflection in EEG after approximately 300 ms owing to the onset as illustrated in Fig. 2.

The P300 speller that follows the oddball paradigm has been proposed by Blackertz et al[5]. They used a 6×6 character matrix as shown in Fig. 3. We will discuss this matrix in detail in Section 3.1.

The current studies on the analysis of P300 can be divided into two categories, statistical learning and deep learning. Both the categories are data-driven and largely disregard the inner structure of a brain signal. In our work, we propose a new classification method based on the sparse representation of the original signals and their reconstruction error. First, sample selection is adopted to construct a dictionary using discriminative samples, which is quite different from traditional sparse coding methods and particularly effective for noisy brain signals. Next, we propose the Channel-Aware Dictionary (CAD) method that splits a dictionary based on channels and determines the most discriminative channels for our task. Then, we refine the channel selection method by considering channel correlations and classifying the brain signal of one channel jointly with its spatially neighboring channels. Experimental results on existing datasets show that we improve the performance by 25% and 17.5% for 5 and 15 epochs, respectively, compared with the baseline method.

The remainder of this paper is organized as follows. After a brief review of the related work in Section 2, we introduce the proposed sparse representation and channel selection method in Section 3. In Section 4, we demonstrate the experimental results of our method. Finally, the results are analyzed and discussed in Section 5.

2 Related Work

2.1 Brain signal processing and classification

For its prospective practical applications, much research has been devoted to the potential of P300. In general, the process of the P300 signal can be separated into data preprocessing, feature extraction, and classification.

As brain signals are often noisy, data preprocessing is extremely important for noise reduction. The noise is mainly composed of two parts: voltage drift, which has a frequency that is close to zero; and power line interference at the frequency of 50 Hz. The signals that carry useful P300 are often restricted in certain bands. To reduce the noise effect, a band-pass filter is often designed to eliminate the interference; moreover, independent component analysis is widely used to remove artifacts caused by eye or muscle movements[6]. This is a special case of blind source separation that can separate components from different sources.

The brain signals are continuously recorded in multiple channels. As the sampling rate is usually high,
the original signals consist of high-dimensional data that contain much useless and redundant information. Many methods have been presented for feature extraction to retain the most useful information, the simplest and most effective of which is down-sampling in the time domain\cite{7}; moreover, the Common Spatial Pattern (CSP) algorithm\cite{8,9} is one of the most widely used methods for feature extraction, especially for motion imagery tasks, because it uses the channel correlations. It separates a multi-channel signal into additive subcomponents that have maximum differences in variance between two windows\cite{10}. Principal component analysis is used to extract the principal component of brain signals in Ref.\cite{11}.

The extracted features are then fed into classifiers. Various methods have been used for brain signal classification. Among them, a Support Vector Machine (SVM) is most frequently used due to its generalization capability and effectiveness in handling high-dimensional data\cite{12}. An SVM model finds the best hyperplane for separating samples from different classes to a maximum margin. Ensemble SVM methods\cite{7} achieved the best performance in the BCI III competition. They ensemble multiple linear SVM classifiers trained on a small part of the training data; moreover, Linear Discriminant Analysis (LDA) classifiers are used in Refs.\cite{11,13}. LDA finds a linear combination of features that separates different classes of signals. In addition, Cecotti and Graser\cite{14} applied a Convolutional Neural Network (CNN) in the P300 speller paradigm. This network is composed of space convolution, time convolution, and fully connected layers. It is the first successful example of a CNN being used in a P300 classification task. Liu et al.\cite{15} applied batch normalization to address the over-fitting problem for P300 signal detection. Reference\cite{16} exploits auto-encoders to force features that are stable across subjects and trials. The sparse coding method is applied in some motor imagery tasks. References\cite{9,17} develop a sparse representation of EEG signals to achieve better accuracy.

In general, a variety of methods for brain signal processing have been proposed; however, the current method used in the P300 speller paradigm is entirely based on a discriminative model and largely disregards the specific characteristics of brain signals. The existing research proves that an ensemble of SVM classifiers achieves the highest accuracy for P300 classification\cite{7}. An SVM is a discriminative model based on statistical machine learning, which does not consider the inner structure of the brain signals. A small disturbance of one sample may affect the performance of the classifier significantly, because it does not have an incremental learning mechanism for overcoming the change or the increase of the samples. The classifiers need to be retrained once there are changes in subjects or samples, which is time consuming and inconvenient for practical use. In our work, we apply sparse coding in the P300 speller system and vastly improve the performance after considering the characteristics of brain signals. Experimental results demonstrate the effectiveness of our method.

2.2 Sparse representation-based classification

Recently, sparse representation-based classification and dictionary learning methods have been widely studied and have shown impressive performances in many different tasks. Li et al.\cite{18} proposed a Discriminative Fisher Embedding Dictionary Learning (DFEDL) algorithm that establishes Fisher embedding models on atoms and coefficients to promote the discriminative capabilities of the dictionary and coding coefficients. Atoms of the same class are forced to reconstruct the corresponding training samples, while the profiles (row vectors of the coding coefficient matrix) and coding coefficients are also imposed by the Fisher criterion. Zhang et al.\cite{19} proposed an Analysis Discriminative Dictionary Learning (ADDL) framework that integrates the analysis dictionary learning, analysis representation, and analysis classifier into a unified model to promote the discriminative capabilities and improve computational efficiency. References\cite{19,20} inspired by projective Dictionary Pair Learning (DPL)\cite{21} integrates the dictionary learning, block-diagonal representation, and classification into a unified model to save training and testing time, and it is referred to as Locality-Constrained Projective Dictionary Learning (LC-PDL). Reference\cite{22} notes that fine-tuning a pre-trained network that is widely used in deep learning may fail when the distribution of the test dataset is quite different from that of the training. It proposed a Joint Projection and Low-Rank Dictionary Learning (JP-LRDL) method that exploits a class-specific dictionary and imposes a graph constraint on the coding coefficients to improve discrimination; however, these methods largely disregard the noise of the samples and do not perform the brain signal classification task satisfactorily.
3 Method

In this paper, we propose a novel sparse coding method based on signal representation and reconstruction for brain signal classification that achieves better accuracy, scalability, and flexibility. It is also quite different from other sparse coding methods\(^9, 17\) because sample selection and channel selection are carefully designed to effectively reduce the effect of noise.

3.1 Dataset description

We use the BCI Competition III Dataset II\(^5\), which is the dataset most commonly used by researchers. It records the brain signals of two subjects. For each character to be spelled, six rows and six columns are randomly intensified. Each row or column is flashed for 100 ms, and then the matrix is blank for 75 ms. The subjects are asked to focus on the desired character position; thus, only two of the 12 stimuli are different from the other stimuli and can cause a P300 response. The sets of 12 intensifications for one character are repeated for 15 epochs; therefore, there are a total of 180 intensifications for one character, among which 30 are positive samples and can drive a P300 response, whereas the other 150 are negative samples. Each subject is recorded for 85 characters for training and 100 characters for testing.

The brain signals are high-dimensional data due to high sampling rates and multiple channel configurations. The signal is digitized at a sample rate of 240 Hz and has 64 channels. If we choose a 667-ms window after the onset, the dimension of a sample is \(10 \times 240\) (240×0.667×64), which is too high for classification. We use a down-sampling method for feature extraction. Because 10 Hz is chosen as the high cut-off frequency, according to the Nyquist-Shannon sampling theorem, a sample rate higher than 20 Hz is sufficient for determining the entire signal; thus, the original signals are down-sampled at 20 Hz. In this way, we only obtain a 14-dimension signal for each channel and an 896-dimension signal for all 64 channels for a 667-ms window, which is much smaller than 10 240 and will be feasible for the following classification.

3.2 Proposed sparse representation-based classification method

The framework of sparse representation is shown in Fig. 4. We first exploit sample selection and choose discriminative training samples to build the dictionary. Test samples are then sparse represented and classified by assigning the class with the lowest reconstruction error.

We denote \(N_{\text{elec}}\) as the number of channels and \(d\) as the dimensionality of each channel. The dimensionality of the entire signal is \(D_s\), and we have

\[
D_s = N_{\text{elec}} \times d \quad (1)
\]

In the dataset as we introduced in Section 3.1, \(N_{\text{elec}} = 64, d = 14,\) and \(D_s = 896.\) Suppose \(y \in \mathbb{R}^{D_s}\) is a sample to be tested, \(D = [D_1, D_2] \in \mathbb{R}^{D_s \times 2N}\) is the dictionary used in sparse coding, where \(D_i \in \mathbb{R}^{D_s \times N}\) \((i = 1, 2)\) is the training samples selected from class \(i,\) \(x \in \mathbb{R}^{2N}\) is the sparse code of the test signal, and \(N\) is the number of positive/negative training signals. Given \(D\) and \(y,\) we solve the following convex optimization problem to obtain the corresponding sparse representation \(x:\)

\[
x^* = \arg\min_{x} \|D_x - y\|^2_F + \lambda \|x\|_0 \quad (2)
\]

The loss function is the sum of the reconstruction error and the sparsity of the representation, where \(\lambda\) is a regularization parameter for balancing the different losses. The posed problem is indeed NP-hard and can only be solved using approximation algorithms. One

<table>
<thead>
<tr>
<th>Table 1 BCI Competition III Dataset II size.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training dataset size</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Subject A</td>
</tr>
<tr>
<td>P300</td>
</tr>
<tr>
<td>No P300</td>
</tr>
<tr>
<td>Subject B</td>
</tr>
<tr>
<td>P300</td>
</tr>
<tr>
<td>No P300</td>
</tr>
</tbody>
</table>

Fig. 4 Framework of sparse representation used in the P300 speller paradigm. Training samples are used to build the dictionary after sample selection. Test samples are then sparse represented and classified according to the sparse code and reconstruction error.
solution is a convex relaxation of the problem, obtained using the \( \ell_1 \)-norm instead of \( \ell_0 \).

\[
x^* = \arg \min_x \|Dx - y\|_2^2 + \lambda \|x\|_1
\]  

This equation is a standard \( \ell_1 \)-norm minimization problem, which can be solved using the matching pursuit algorithm.[23]

The dictionary \( D \) has a size of \( 2N \) columns. The \( N \) left columns correspond to \( N \) positive training samples (P300), whereas the \( N \) right columns are \( N \) negative samples (no P300). It is easy to see that a positive test signal (P300) can be better reconstructed by positive training signals; thus, the sparse code has more non-zero items in the first half, whereas the sparse code for a negative test sample (no P300) has more non-zero items in the second half. Hence, \( y \) can be classified according to the class-wise reconstruction residual,

\[
\text{Class}(y) = \arg \min_i \|D_i x_i - y\|_F^2
\]

where \( x_i \in \mathbb{R}^{N \times 1} \) \( (i = 1, 2) \) is the coefficient vector corresponding to sub-dictionary \( D_i \). The intersection of the row and column is the character spelled by the subjects.

3.3 Data cleaning and discriminative training samples selection

Because brain signals are quite noisy, using all training signals as the bases to build the dictionary will lead to a poor classification performance. Actually, sparse coding requires the two subspaces constructed by positive and negative bases to be mutually orthogonal. In the face recognition application, images with high resolution and quality are manually selected to construct the dictionary. Experimentation has shown that the signals used in the dictionary should have a high Signal-to-Noise Ratio (SNR) and be sufficiently discriminative.[24]

Hence, we propose the Sample Selection (SS) method to reduce the noise effect by selecting discriminative training signals. It is obvious that we cannot select them manually as in face recognition because it is much more difficult to judge whether the brain signal is sufficiently qualified. We design a simple linear SVM classifier instead to provide a score for each signal. For each training sample \( x_i \), which means the \( j \)-th epoch of the \( i \)-th sample, the score \( s_i \) is given by the following equation:

\[
s_i = \frac{1}{J} \sum_{j=1}^{J} x_{ij}^T \cdot w + b
\]

where \( J \) is the number of epochs, \( w \) is the support vector, and \( b \) is the bias term. The scores after the ranking are plotted in Fig. 5 as a function of the sample number. The first 2350 scores in red are positive samples (P300), whereas the scores in blue are negative samples (no P300), and they are sorted separately. The score of the P300 sample is always positive, and therefore represents the confidence value for correctly classifying this sample. If the score of the P300 sample is negative, it is more likely to be less qualified and easily misclassified; thus, we choose the \( N \) samples with the highest scores as the positive dictionary bases and the \( N \) samples with lowest scores as the negative dictionary bases. The dictionary size \( N \) is determined by the experiments in Section 4.1.

Note that sample selection only needs to be done once when training. It helps to select discriminative and typical samples to build the dictionary. During the test phase, we only need to optimize according to Eq. (3) based on the pre-built dictionary.

3.4 CAD for discriminative channel selection

In Section 3.2, the brain signals of each channel are filtered and down-sampled in that order, and then the features of all channels are concatenated to a long vector; however, not all channels are beneficial and some channels are quite noisy. The process of sparse coding is similar to fitting. If the noisy channels constitute a large percentage of the total number of channels, noise fitting will have a considerable effect on the sparse code, which will be less credible when used for classification.

Thus, we propose a new method called the CAD to select discriminative channels. For preprocessing, the
brain signals of each channel are filtered and then downsampled but not concatenated. Then, the dictionary is divided into $N_{\text{elec}}$ sub-dictionaries. Each sub-dictionary, $C_i \in \mathbb{R}^{d \times 2N}$ ($i = 1, 2, \ldots, N_{\text{elec}}$), corresponds to one channel. Optimization of the following equation is repeated $N_{\text{elec}}$ times for one test signal using $N_{\text{elec}}$ different dictionaries:

$$x_i^* = \arg \min_{x_i} \|C_i x_i - y_i\|^2_F + \lambda_i \|x_i\|$$  \hspace{1cm} (6)

An intuitive idea is that we can evaluate the performance of the channels separately and select the top $K$ discriminative channels, by which noisy channels will be excluded from P300 classification.

However, processing the brain signal of different channels separately may cause an unstable sparse representation and lead to a sub-optimal solution. Moreover, the adoption of multiple channels with higher classification rates might not increase the overall classification rate due to a high correlation between channels. Hence, channel correlation should be considered for brain signal analysis, and we propose Neighboring channels Group Sparse Coding (NGSC) from the sample set perspective to achieve a better performance.

### 3.4.1 NGSC from the sample set perspective

Inspired by Ref. [25], we can improve the classification accuracy if we can classify each channel jointly with its spatially neighboring channels. These channels form a set of samples from the same class. We believe we are the first to exploit sample set sparse representation in the P300 classification task. The sample set can virtually synthesize new samples, which is helpful for exploiting the channel correlation and complementary information among the spatially neighboring channels to overcome the instability of sparse representations; Suppose $Y = [y_0, y_1, \ldots, y_{n-1}] \in \mathbb{R}^{d \times n}$ is the sample matrix, where $y_0 \in \mathbb{R}^{d \times 1}$ corresponds to the channel to be classified and $y_i \in \mathbb{R}^{d \times 1}$ ($i = 1, 2, \ldots, n - 1$) are the spatially neighboring channels for forming a sample set. Because our goal is to determine the class $y_0$ based on the sample set $Y$, the sample set sparse representation can be rewritten as follows:

$$x^*, w^* = \arg \min_{x,w} \|C x - Y w\|^2_F + \lambda \|x\|_1$$  \hspace{1cm} (7)

where $C$ is a CAD corresponding to $y_0$, $\lambda$ is a scalar constant for balancing the reconstruction residual and representation sparsity, and $w = [w_0, w_1, \ldots, w_{n-1}]^T \in \mathbb{R}^{n \times 1}$ is the weight vector of the sample set with the constraint $\|w\|_1 = 1$ for avoiding the trivial solution $w = 0$ and $x = 0$, and making the computation stable.

Suppose $\hat{y} \in \mathbb{R}^{d \times 1}$ is an arbitrary reference point residing in the space spanned by sample set $Y$, that is

$$\exists b \in \mathbb{R}^{n \times 1}, \|b\| = 1, \hat{y} = \sum_{i=0}^{n-1} b_i y_i$$  \hspace{1cm} (8)

We introduce the following lemma to transform the original problem Eq. (7).

**Lemma 1**

$$\exists e \in \mathbb{R}^{n \times 1}, \hat{y} + T c = Y w, \text{s.t.} \|w\| = 1$$  \hspace{1cm} (9)

where $T = [y_0 - \hat{y}, y_1 - \hat{y}, \ldots, y_{n-1} - \hat{y}] \in \mathbb{R}^{d \times n}$ is the difference matrix of the sample set $Y$ with respect to the reference point $\hat{y}$, and $c = [c_0, c_1, \ldots, c_{n-1}]^T \in \mathbb{R}^{n \times 1}$ is the vector of the coefficient.

**Proof**

$$\hat{y} + T c = \hat{y} + (y_0 - \hat{y}) c_0 + \ldots + (y_{n-1} - \hat{y}) c_{n-1} = (1 - \sum_{i=0}^{n-1} c_i) \hat{y} + \sum_{i=0}^{n-1} c_i y_i$$  \hspace{1cm} (10)

Substituting Eq. (8) into Eq. (10), we obtain

$$\hat{y} + T c = (1 - \sum_{i=0}^{n-1} c_i) \hat{y} + \sum_{i=0}^{n-1} c_i y_i = \sum_{i=0}^{n-1} \left( b_i + c_i \right) y_i$$  \hspace{1cm} (11)

Let $u_i = b_i + c_i - b_i \sum_{j=0}^{n-1} c_j$, then

$$\|w\| = \sum_{i=0}^{n-1} u_i = \sum_{i=0}^{n-1} \left( b_i + c_i - b_i \sum_{j=0}^{n-1} c_j \right).$$

As $\|b\| = 1$, we then obtain

$$\|w\| = 1 + \sum_{i=0}^{n-1} c_i - \sum_{j=0}^{n-1} c_j = 1.$$

Therefore, $\hat{y} + T c = Y w$, s.t. $\|w\| = 1$.  \hspace{1cm}  \(\blacksquare\)

**Lemma 1** is quite important because the coefficient vector $c$ has no constraints, in contrast to the original $w$ with constraint $\|w\| = 1$.

Substituting Eq. (9) into Eq. (7), we obtain

$$x^*, c^* = \arg \min_{x,c} \|C x - \hat{y} - T c\|^2_F + \lambda \|x\|_1$$  \hspace{1cm} (12)

Recall that we want to determine the class of $y_0$ by exploiting the sample set $Y$, hence we should pay more attention to $y_0$ to prevent bias to other samples. In this paper, we emphasize the role of $y_0$ using a Gaussian function with $y_0$ as the center,

$$\hat{y} = \frac{1}{Z} \sum_{i=0}^{n-1} \exp \left( -\frac{\|y_i - y_0\|^2}{2\sigma^2} \right) \cdot y_i$$  \hspace{1cm} (13)
where $\sigma$ is the standard deviation of the Gaussian function and $Z$ is a constant for normalization. Because the center of the Gaussian function is set as $y_0$, the reference point $\hat{y}$ will consider all channels in the sample set while emphasizing $y_0$.

### 3.4.2 Weighted regulation strategy

Note that the method described in Section 3.4.1 is quite sensitive to the noisy samples in the set; moreover, the channels in the sample set are not equally contributed and we should reduce the effect of channels that are not similar to the reference signal. Because the combination coefficient $c$ measures the contribution of each sample, we can regularize $c$ according to the distance between $y_i$ and the reference point $\hat{y}$ as

$$x^*, c^* = \arg \min_{x, c} \|Cx - y - Tc\|_F^2 + \lambda \|x\|_1 + \alpha \|\Gamma c\|_2^2$$

(14)

where $\lambda$ and $\alpha$ are constant scalars for balancing the reconstruction residual, representation sparsity and the noisy sample penalty, respectively, $\Gamma$ is the biasing Tikhonov matrix to the reference point $\hat{y}$, that is

$$\Gamma = \begin{bmatrix} \|y_0 - \hat{y}\|_2 & 0 \\ 0 & \ddots & 0 \\ 0 & \|y_{2N-1} - \hat{y}\|_2 \end{bmatrix}$$

(15)

The weighted regulation strategy can help to constrain the noisy channel to make a smaller contribution to the final model, which consequently makes the optimization process more stable.

### 3.4.3 Optimization

Equation (14) can be solved by the alternating iterative method. Determining $x$ while keeping $c$ fixed is the original sparse representation problem. Determining $c$ while keeping $x$ fixed is a weighted least-squares problem that can be solved as follows:

$$c = (T^T T + \alpha \Gamma^T \Gamma)^{-1} T^T (Cx - \hat{y})$$

(16)

The two steps are iterated alternatively until the solution converges. After obtaining the optimal solution $x^*$ and $c^*$, we can classify $y_0$ according to a class-wise reconstruction residual similar to Eq. (4).

Then, we can sort the channels by the reconstruction residual and select the top $K$ discriminative channels, by which noisy channels will be excluded from P300 classification. Here, $K$ is a tradeoff between the recognition rate and computational cost. For practical use, we can determine the number of channels according to the requirements of accuracy and the limitation of complexity. NGSC, from the sample set perspective, can better evaluate the discriminative capabilities of each channel. We will show the $K$ most discriminative channels in Section 4.2.

Moreover, the computational complexity for Eq. (3) is $O(2N \times N_{elec}^2 \times d^2)$. After channel selection, only discriminative channels will be used during the test phase, and the computational cost will be significantly reduced to $O(2N \times K^2 \times d^2)$, where $K$ is the number of channels we select.

### 4 Experiment

In this section, we present the results using a sparse representation-based method. At first, signals are preprocessed using a 8-order band-pass Chebyshev Type I filter with cut-off frequencies between 0.1 Hz and 10 Hz. Then, the filtered signals are down-sampled at 20 Hz. After that, the signals in the training dataset are used to build a dictionary. Samples in the test dataset are then sparse represented and classified according to the sparse code and reconstruction error. That is our baseline experiment. We then compare the accuracy of using SS and NGSC with the accuracy of the baseline and state-of-the-art methods.

#### 4.1 Parameter analysis

The sample selection method was introduced in Section 3.3. Here, we determine the dictionary size according to the experimental results. We depict the mean accuracy of two subjects using different dictionary sizes for sample selection in Table 2.

Among all the parameters used, the best performance is obtained when 2000 samples are used to construct the dictionary—1000 positive samples (P300) and 1000 negative samples (no P300)—which achieves an accuracy of 50.5% for five epochs and 88.5% for 15 epochs. The sample sizes for different classes are equal, so we do not need to normalize this feature when comparing reconstruction residuals. The proposed method is better than the baseline method implemented without sample selection, where the accuracy for 5 and 15 epochs is 47.5% and 81%, respectively.

Meanwhile, low accuracy is obtained if the number of bases used is too small (corresponding to a dictionary

#### Table 2 Mean accuracy of two subjects using different dictionary sizes for sample selection. (%)

<table>
<thead>
<tr>
<th>Epoch Without sample selection</th>
<th>Dictionary size</th>
<th>Rand</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>47.5</td>
<td>45.0</td>
</tr>
<tr>
<td>15</td>
<td>81.0</td>
<td>73.0</td>
</tr>
</tbody>
</table>
size equal to 1000). In this case, bases are not sufficiently descriptive and some patterns may not be represented correctly. The last column in the table presents the performance when 2000 random samples are used as the bases, which is worse than the baseline and proves the effectiveness of our method for choosing samples.

Note that a dictionary size of 2000 is quite small compared with the original 15 300 samples, making our optimization progress much faster.

4.2 Channel selection and electrode ranking results

Figures 6a and 6c show the recognition rate plotted against the channel number to demonstrate the accuracy of using 64 channels independently for Subjects A and B, respectively. The recognition rate of the best and worst channel for Subject A is listed in Table 3. Channel 11 (Cz) is most discriminative, as it obtains a 45% recognition rate after 15 epochs without using the information of other channels, whereas Channel 45 (Tp7) is so noisy that the recognition rate is similar to that of random guessing. This result proves that not all channels will benefit the P300 classification task and channel selection is quite necessary.

Figures 6b and 6d present the corresponding accuracy topographies. We find that channels in the parietal lobe are more discriminative for our P300 classification task.

We select the top eight channels during the test phase. Here, eight is a tradeoff between the recognition rate and the computational cost. For practical use, we can determine the number of channels according to the requirements of accuracy and the limitation of computational complexity. The top eight channels for Subjects A and B are listed in Table 4. Both subjects share Cz, Fc4, C1, and Fc2 as the top-ranked channel.

### Table 3 Recognition rate of Channels Cz and Tp7. (%)

<table>
<thead>
<tr>
<th>Channel</th>
<th>Epoch 1</th>
<th>Epoch 5</th>
<th>Epoch 10</th>
<th>Epoch 13</th>
<th>Epoch 14</th>
<th>Epoch 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cz</td>
<td>3.3</td>
<td>20.0</td>
<td>28.3</td>
<td>45.0</td>
<td>48.3</td>
<td>45.0</td>
</tr>
<tr>
<td>Tp7</td>
<td>5.0</td>
<td>1.7</td>
<td>1.7</td>
<td>1.7</td>
<td>3.3</td>
<td>8.3</td>
</tr>
</tbody>
</table>

Fig. 6 (a) Performance for using each channel independently for Subject A. (b) Corresponding accuracy topography for Subject A. (c) Performance for using each channel independently for Subject B. (d) Corresponding accuracy topography for Subject B.
The electrode ranking is not quite the same for Subjects A and B due to individual differences. Note that channel selection is very important for a practical BCI system, where we want to use as few channels as possible for a better user experience.

### 4.3 Recognition rates on the test dataset

For the dataset mentioned in Section 3.1, the performance is evaluated based on the recognition rate of the predicted characters in the test sets. Hence, it is a 36-class classification problem. We demonstrate the character recognition rates of the proposed method on the test dataset in Table 5.

It is intuitive that the accuracy increases with the number of epochs used, as shown in Table 5. Additionally, we find that the accuracy of Subject B is better than that of A, which indicates a difference between individuals in the P300 classification task.

Sparse coding representation based on a reconstruction strategy is quite suitable for brain signal analysis and BCI applications. In fact, the bases used in the dictionary can be viewed as different patterns of P300 (P3a, P3b, etc.). The non-zero items in the sparse code reflect the internal structure of the signal.

SS introduced in Section 3.3 helps to select discriminative and typical samples for building the dictionary, which greatly improve the quality of the training samples and reduce the effect of noise. We see that the mean recognition rates are increased by 3% and 7.5% at 5 and 15 epochs, respectively, after sample selection.

NGSC from the sample set perspective, introduced in Section 3.4.1, also improves the performance significantly by 25% and 17.5% at 5 and 15 epochs, respectively, compared with the baseline. The CAD helps to increase the SNR by selecting the top performance channels, and NGSC uses the correlation of spatially neighboring channels to make the sparse representation more stable. The result shows that sample selection and channel selection are quite effective for brain signal processing.

Table 6 compares the results of the presented method with other methods in the literature.

---

**Table 4** 8 top-ranked channels for Subjects A and B.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Electrode ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Cz Fc4 Fcz Pz C1 Fc2 Fc1 Cpz</td>
</tr>
<tr>
<td>B</td>
<td>Fc2 C2 Fc4 Cz Cp2 F1 C1 Cp1</td>
</tr>
</tbody>
</table>

---

**Table 5** Recognition rate on the test dataset for the proposed methods. (%)

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject</th>
<th>Epoch</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>A</td>
<td></td>
<td>18</td>
<td>23</td>
<td>28</td>
<td>31</td>
<td>43</td>
<td>45</td>
<td>46</td>
<td>48</td>
<td>58</td>
<td>65</td>
<td>66</td>
<td>68</td>
<td>72</td>
<td>78</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td></td>
<td>28</td>
<td>31</td>
<td>39</td>
<td>44</td>
<td>52</td>
<td>52</td>
<td>57</td>
<td>59</td>
<td>64</td>
<td>71</td>
<td>76</td>
<td>79</td>
<td>79</td>
<td>85</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>23.0</td>
<td>27.0</td>
<td>33.5</td>
<td>37.5</td>
<td>47.5</td>
<td>48.5</td>
<td>51.5</td>
<td>53.5</td>
<td>61.0</td>
<td>68.0</td>
<td>71.0</td>
<td>73.5</td>
<td>75.5</td>
<td>81.5</td>
<td>81.0</td>
</tr>
<tr>
<td>BL+SS</td>
<td>A</td>
<td></td>
<td>21</td>
<td>25</td>
<td>36</td>
<td>40</td>
<td>43</td>
<td>50</td>
<td>55</td>
<td>63</td>
<td>68</td>
<td>70</td>
<td>71</td>
<td>73</td>
<td>75</td>
<td>87</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td></td>
<td>36</td>
<td>41</td>
<td>49</td>
<td>54</td>
<td>58</td>
<td>61</td>
<td>62</td>
<td>66</td>
<td>70</td>
<td>74</td>
<td>75</td>
<td>76</td>
<td>84</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>28.5</td>
<td>33.0</td>
<td>42.5</td>
<td>47.0</td>
<td>50.5</td>
<td>55.5</td>
<td>58.5</td>
<td>64.5</td>
<td>69.0</td>
<td>72.0</td>
<td>73.0</td>
<td>74.5</td>
<td>79.5</td>
<td>89.0</td>
<td>88.5</td>
</tr>
<tr>
<td>BL+SS+NGSC</td>
<td>A</td>
<td></td>
<td>23</td>
<td>26</td>
<td>44</td>
<td>57</td>
<td>66</td>
<td>69</td>
<td>72</td>
<td>73</td>
<td>78</td>
<td>85</td>
<td>89</td>
<td>88</td>
<td>90</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td></td>
<td>42</td>
<td>46</td>
<td>66</td>
<td>70</td>
<td>79</td>
<td>82</td>
<td>82</td>
<td>85</td>
<td>88</td>
<td>92</td>
<td>92</td>
<td>93</td>
<td>95</td>
<td>97</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td></td>
<td>32.5</td>
<td>36.0</td>
<td>55.0</td>
<td>63.5</td>
<td>72.5</td>
<td>75.5</td>
<td>77.0</td>
<td>79.0</td>
<td>83.0</td>
<td>88.5</td>
<td>90.5</td>
<td>90.5</td>
<td>92.5</td>
<td>97.5</td>
<td>98.5</td>
</tr>
</tbody>
</table>

---

**Table 6** Recognition rates comparison with other methods in the literature. (%)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>45</td>
<td>72</td>
<td>61</td>
<td>55</td>
<td>57</td>
<td>62</td>
<td>49</td>
<td>66</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>78</td>
<td>83</td>
<td>82</td>
<td>79</td>
<td>81</td>
<td>81</td>
<td>76</td>
<td>91</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>88</td>
<td>97</td>
<td>97</td>
<td>94</td>
<td>93</td>
<td>96</td>
<td>91</td>
<td>98</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>76</td>
<td>75</td>
<td>77</td>
<td>72</td>
<td>70</td>
<td>77</td>
<td>69</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>92</td>
<td>91</td>
<td>92</td>
<td>84</td>
<td>86</td>
<td>89</td>
<td>81</td>
<td>92</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>96</td>
<td>96</td>
<td>94</td>
<td>89</td>
<td>90</td>
<td>92</td>
<td>90</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>5</td>
<td>60.5</td>
<td>73.5</td>
<td>69.0</td>
<td>63.5</td>
<td>63.5</td>
<td>69.5</td>
<td>59.0</td>
<td>72.5</td>
<td>88.5</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>85.0</td>
<td>87.0</td>
<td>87.0</td>
<td>81.5</td>
<td>83.5</td>
<td>85.0</td>
<td>78.5</td>
<td>88.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>92.0</td>
<td>96.5</td>
<td>95.5</td>
<td>91.5</td>
<td>91.5</td>
<td>94.0</td>
<td>90.5</td>
<td>98.5</td>
<td></td>
</tr>
</tbody>
</table>
method with those of other algorithms in the literatures for frequently used epochs. We achieve the mean recognition rate of 72.5%, 88.5%, and 98.5% at 5, 10, and 15 epochs, respectively, which outperforms most existing algorithms including the winner of the BCI competition[7] and the neural network-based method[14]. Moreover, the presented method has some advantages for practical use. The existing methods, such as SVM or the neural network-based method, do not have an incremental learning mechanism to fit the change of samples and have to train different classifiers for different subjects; however, what we proposed is a subject-independent method in which a new sample can be directly added to the dictionary as long as it passes the sample selection test. Compared with other dictionary learning-based methods, our method shows better performance because we pay more attention to the noise of the brain signal and select the most discriminative samples and channels for training.

In our baseline experiment, the recognition rate is not satisfactory. We improve the performance significantly using sample selection and a CAD, which are both designed to reduce the noise effect and increase the SNR.

4.4 Efficiency comparison

We run our experiments using Matlab 2017a on the 2.40 GHz CPU of a 16-core Intel Xeon server with 32 GB memory. Table 7 compares the average training and testing times for different methods. The results show that our proposed methods require much less time for training compared with neural network-based methods and other dictionary learning-based methods. The test time is greatly reduced thanks to sample selection and channel selection.

In conclusion, the presented method can better meet the user’s expectation of a fast response time while enjoying high accuracy. Hence, it is more suitable for real-world applications.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training time (s)</th>
<th>Test time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESVM[7]</td>
<td>20.48</td>
<td>2.75</td>
</tr>
<tr>
<td>CNN[14]</td>
<td>1168</td>
<td>3.49</td>
</tr>
<tr>
<td>LDA[26]</td>
<td>14.27</td>
<td>1.38</td>
</tr>
<tr>
<td>DFEDL[18]</td>
<td>45.50</td>
<td>1.17</td>
</tr>
<tr>
<td>ADDL[19]</td>
<td>52.31</td>
<td>0.94</td>
</tr>
<tr>
<td>LC-PDL[20]</td>
<td>20.24</td>
<td>0.66</td>
</tr>
<tr>
<td>JP-LRDL[22]</td>
<td>189.00</td>
<td>2.45</td>
</tr>
<tr>
<td>Ours</td>
<td>18.57</td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>

5 Conclusion

Classification is an important step for a BCI system. Current research mainly focuses on discriminative methods, such as SVM, LDA, or using a data-driven neural network, because they are easy to implement and always obtain fine results. However, a brain signal has its own characteristics and its inner structure should be valued. It should not be processed as common signals such as an image or voice.

We believe that few efforts have been devoted to sparse coding-based methods in P300 classification. In this paper, we investigate its feasibility for brain signal classification and demonstrate its great effectiveness. Training signals are used to build the dictionaries and test signals are classified according to their sparse representation and reconstruction errors; moreover, sample selection and a CAD are proposed to reduce the noise effect and improve the SNR. Training samples are evaluated and the discriminative ones are chosen to achieve high-quality dictionaries. In addition, channel correlations are also exploited by jointly classifying the brain signals with its spatially neighboring channels and a novel weighted regulation strategy is proposed to overcome outliers in the group. Channel selection and NGSC improve the recognition rate and enhance the computing efficiency simultaneously.

Our method exploits sparse representation of the original brain signal, which is quite flexible for classification. In the future, we will introduce discriminative restrictions into the sparse representation framework. Sparse code is a good representation of the original signal and can be combined with more classification methods; moreover, channel selection can be improved by weighting all channels to make full use of all information. The correlation of channels can also be further exploited by considering symmetry and different grouping methods.

Acknowledgment

This work was supported by the National High Technology Research and Development (863) Program of China (No. 2012AA011004) and the National Science and Technology Support Program (No. 2013BAK02B04).

References


Yali Li received the BEng degree with Excellent Graduates Award from Nanjing University, China in 2007 and the PhD degree from Tsinghua University, Beijing, China in 2013. Currently she is a research assistant in the Department of Electronic Engineering, Tsinghua University. Her research interests include image processing, pattern recognition, computer vision, and video analysis.

Hongma Liu received the BEng degree from Southeast University, Nanjing, China in 2012. He is currently pursuing the PhD degree at the Department of Electronic Engineering, Tsinghua University, Beijing, China. His current research interests include pattern recognition, computer vision, BCIs, and object recognition.

Tsinghua Science and Technology, August 2021, 26(4): 440–451
Shengjin Wang received the BEng degree from Tsinghua University, China in 1985 and the PhD degree from the Tokyo Institute of Technology, Tokyo, Japan in 1997. From May 1997 to August 2003, he was a member of the research staff in the Internet System Research Laboratories, NEC Corporation, Japan. Since September 2003, he has been a professor with the Department of Electronic Engineering, Tsinghua University, where he is currently also a director of the Research Institute of Image and Graphics. He has published more than 50 papers on image processing. He is the holder of ten patents. His current research interests include image processing, computer vision, video surveillance, and pattern recognition.