Spatial-temporal discriminant analysis for ERP-based brain-computer interface

Yu Zhang, Guoxu Zhou, Qibin Zhao, Jing Jin, Xingyu Wang, and Andrzej Cichocki

***Abstract*—Linear discriminant analysis (LDA) has been widely adopted to classify event-related potential (ERP) in brain-computer interface (BCI). Good classification performance of the ERP-based BCI usually requires sufficient data recordings for effective training of the LDA classifier, and hence a long system calibration time which however may depress the system practicability and cause the users resistance to the BCI system. In this study, we introduce a spatial-temporal discriminant analysis (STDA) to ERP classification. As a multiway extension of the LDA, the STDA method tries to maximize the discriminant information between target and non-target classes through finding two projection matrices from spatial and temporal dimensions collaboratively, which reduces effectively the feature dimensionality in the discriminant analysis, and hence decreases significantly the number of required training samples. The proposed STDA method was validated with dataset II of the BCI Competition III and dataset recorded from our own experiments, and compared to the state-of-the-art algorithms for ERP classification. Online experiments were additionally implemented for the validation. The superior classification performance in using few training samples shows that the STDA is effective to reduce the system calibration time and improve the classification accuracy, thereby enhancing the practicability of ERP-based BCI.**

***Index Terms*—Brain-computer interface (BCI), Electroencephalogram (EEG), Event-related potential (ERP), Linear discriminant analysis (LDA), Spatial-temporal discriminant analysis (STDA)**

I. Introduction

A Brain-Computer Interface (BCI) is a system that allows non-muscular connection between a human brain and a computer, thereby providing a new communication channel, particularly for people with severe motor disabilities [1]. BCI can translate the brain signals of subjects to computer commands from various electroencephalogram (EEG) modulations in which sensorimotor rhythm (SMR), steady-state visual evoked potential (SSVEP) and event-related potential (ERP) are mostly used for the BCI development.

Pfurtscheller et al [2] first described the SMR-based BCI which assists subjects to control external devices by utilizing the event-related (de)synchronization (ERD/ERS) of SMRs when the subjects imagine movements of their bodies, typically left and right hands, foot or tongue. Although the SMR-based BCI system has been validated with both healthy and disabled subjects [3], [4], it usually requires relatively long training time for subjects to achieve accurate control over the SMRs, and may even become ineffective for a certain number of subjects due to their poor SMR modulation abilities [5]. The SSVEP-based BCI has been developed with characteristics of simpler calibration procedure and higher information transfer rate (ITR), which depends typically on external visual stimuli in the form of an array of light sources, where each light source flickers with a distinct frequency [6], [7], [8]. Through recognizing frequency components of the SSVEP elicited with the same frequency as the stimulus and usually also higher harmonics, the SSVEP-based BCI can export the attended command with effective communication performance for most subjects, when they focus attentions on a flickering stimulus [9]. However, some subjects still fail to use the SSVEP-based BCI due to their annoyances and fatigues caused by the flickering stimuli [9], [10], and a risk of inducing photoepileptic seizures could be a hidden trouble of such BCI system when using stimulus frequencies in the mid-late beta band (15-25 Hz) [11]. The ERP-based BCI has become relatively more attractive because of robust performance for target detection and also no requirement for subject training [12]. So far, P300 has been most widely employed for the ERP-based BCI development, which is a positive ERP component occurring approximately 300 ms after a rare but task-related stimulus (i.e., 'oddball' paradigm) [13]. In the oddball paradigm with several stimuli, only intensifications of the attended stimuli should elicit the P300s, and thus the desired stimuli (or commands) can be determined through classifying the stimuli that yield the largest P300s [14], [15]. Recently, several novel variants have been proposed to improve the P300-based BCI, which either adopted individually another ERP component (e.g., N200 related to motion-onset) [16] or combined one or more other ERP components (e.g., VPP and N170 related to face perception) with P300 [17], [18], [19].

In addition to stimulus paradigm modification, classification algorithm optimization is also one of the key points to develop an improved ERP-based BCI [20]. Linear discriminant analysis (LDA) is probably most used and has demonstrated its strength for ERP classification [21], [22]. The LDA usually

This study was supported in part by the Nation Nature Science Foundation of China 61074113, 61203127, 61103122, 61202155, Shanghai Leading Academic Discipline Project B504, and Fundamental Research Funds for the Central Universities WH1114038.

Y. Zhang is with the Key Laboratory for Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China, and also the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan (e-mail: zhangyu0112@gmail.com).

G. Zhou, Q. Zhao are with the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan (e-mail: zhouguoxu@brain.riken.jp, qbzhao@brain.riken.jp).

J. Jin and X. Wang are with the Key Laboratory for Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mail: jinjingat@gmail.com, xywang@ecust.edu.cn).

A. Cichocki is with the Laboratory for Advanced Brain Signal Processing, RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan, and also the System Research Institute, Polish Academy of Sciences, Warsaw 00-901, Poland (e-mail: a.cichocki@riken.jp).

works well for ERP classification due to the very similar covariance matrices of Gaussian distributions corresponding to the ERP and non-ERP (i.e., target and non-target) features [22]. However, an effective training for the LDA classifier usually requires five to ten times as many training samples per class as the feature dimensionality [20], [23], and hence a long system calibration time which may depress the system practicability and cause the subjects resistance to the BCI system. To enhance practicability of the ERP-based BCI, the calibration stage should be kept as short as possible, thus only few training samples could be recorded. How to design an effective classification algorithm to classify ERP accurately using limited calibration time is a significant issue for improving the ERP-based BCI. Recently, semi-supervised schemes have been proposed to reduce the calibration time for classification in BCIs and achieved good classification performance, especially with limited training samples [24], [25]. On the other hand, regularization methods [22], [26] have also been exploited to improve the generalization capacity of classifier for BCIs in small sample size scenarios, which this study will focus on. As a regularized version of the LDA, stepwise LDA (SWLDA) was originally introduced to ERP classification by Farwell and Donchin [12]. The SWLDA has been recently referred as the state-of-the-art P300 classification algorithm and shown its superiority for the P300 classification over various classification algorithms, such as Pearson's correlation method (PCM), linear support vector machine (LSVM) and Gaussian support vector machine (GSVM) [21]. The SWLDA is commonly employed to alleviate effects of small sample size on LDA transforms, and is less likely to corrupt the classification accuracy using fewer training samples since those insignificant features are removed from the discriminant model by forward and backward stepwise analysis with statistical tests [21], [26]. Recently, another regularized LDA with shrinkage technique, called shrinkage LDA (SKLDA), was proposed to mitigate effects resulting from the high-dimensionality of features compared to the number of training samples on sensorimotor rhythms classification [27]. Blankertz et al [22] subsequently introduced the SKLDA to ERP classification with a detailed interpretation for solving the problem of small sample size, and showed its superior classification performance over the traditional LDA with only few training samples. The SKLDA remedies the ill-conditioned covariance matrix with an appropriately selected shrinkage parameter, and effectively enhances generalization capability of classifier, thereby giving good ERP classification performance even when using insufficient training samples [22].

It is worth noting that the traditional LDA and the two aforementioned regularized versions adopted the vectorized features (i.e., one-way samples) for ERP classification, where each feature vector was the concatenation of temporal points from spatial channels. Such feature vector has typically high dimensionality. Consequently, the number of training samples recorded within limited calibration time is significantly insufficient for covariance matrix estimation (see Fig. 1), since the number of unknown parameters that have to be estimated in covariance matrix is quadratic in the dimensionality [22].

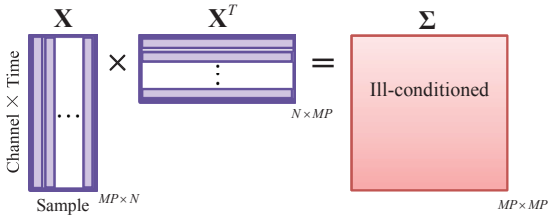


Fig. 1. Illustration for the problem of ill-conditioned covariance matrix in the LDA. Assume the training data $\mathbf{X}$ consists of N samples where each sample is recorded from M channels with P temporal points in each channel. For ERP classification, the required number of parameters $(MP \times (MP-1)/2)$ for the covariance matrix estimation is typically much larger than the number N of training samples recorded from limited calibration time.

A poor estimation for the covariance matrix occurs spontaneously and results in low generalization performance of the trained classifier. As an alternative to the regularized versions of LDA mentioned above, this study introduces a spatial-temporal discriminant analysis (STDA) algorithm to ERP classification, which is basically a multiway extension of the LDA [28]. The STDA learns two projection matrices collaboratively from the spatial and temporal feature subspaces by adopting matrix features (i.e., spatial-temporal two-way samples) instead of vectorized features, and thus effectively reduces feature dimensionality in the discriminant analysis. With the learned projection matrices, each spatial-temporal sample is transformed to a new one-way sample with much lower dimensionality which improves significantly the estimation of covariance matrix in the subsequent ERP classification. Both dataset II of the BCI Competition III and dataset recorded from our own experiments are used to validate the proposed STDA algorithm for ERP classification, especially using few training samples, in contrast to the traditional LDA, SWLDA and SKLDA, etc. Online classification performance of ERP is also evaluated and compared for the aforementioned methods. Superior classification accuracy with insufficient training samples indicates that the STDA is effective to reduce the calibration time of the ERP-based BCI with minimal accuracy degradation, and hence improve the system practicability.

II. Materials and Methods

A. EEG acquisition

1) Dataset-1: The dataset-1 was from dataset II of the BCI competition III (http://www.bbci.de/competition/iii/) and provided by Wadsworth Center, Albany, NY. EEG signals were recorded at 240 Hz sampling rate from 64-scalp positions with high-pass and low-pass filters 0.1 Hz and 60 Hz. As the P300 speller proposed by Farwell and Donchin [12], a 6×6 matrix consisting of characters was presented to the subject as the stimulus paradigm on a computer screen (see Fig. 2 (a)). The row/column of the matrix was intensified for 100 ms with an inter-stimulus interval (ISI) of 75 ms, and the subject was asked to focus attention on the cued character. Row and column intensifications were block randomized in blocks of 12. Only intensifications of the row and column containing the cued character should elicit P300 potentials. The desired character can be determined through detecting

the row and column with largest P300 responses. The sets of intensifications were repeated 15 times and thus a total of 180 intensifications were presented to the subject for each character spelling. The stimulus matrix was blank during the interval of 2.5 s between each two character spellings. Two subjects (subject A and B) EEG data are available, and each subject has a train dataset containing 85 characters and a test dataset containing 100 characters. See [29] for more detailed description of this dataset.

Since this study focuses mainly on the problem of small sample size and labels of the test datasets are not provided in the dataset II, we only use the train datasets of the two subjects for our analysis. The 16 channels F3, Fz, F4, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, PO7, PO8, Oz were used for the subsequent feature extraction and classification. A 700 ms data segment was extracted from the beginning of each intensification and band-pass filtered from 0.1 Hz to 30 Hz by a sixth order forward-backward Butterworth bandpass filter. Each extracted segment was downsampled to 40 Hz by selecting each 6th point from the filtered data, thus the signal of each channel consisted of 28 points. A total of 15300 (180×85) such segments (2550 targets and 12750 non-targets) were extracted for each subject, and each 180 segments (30 targets and 150 non-targets) correspond to one character spelling. The EEG segments recorded from spellings of the first 40 characters were used as training data, while those from spellings of the remaining 45 characters as test data.

2) Dataset-2: The dataset-2 was recorded from our own experiments. Seven healthy right-handed volunteers (S1-S7, aged from 24 to 49, all males) participated in the experiment. The subjects were seated in a comfortable chair 60 cm from a 17 inch LCD monitor (60 Hz refresh rate and 1280×1024 screen resolution) in a shield room. EEG signals were recorded at 256 Hz sampling rate using the g.USBamp amplifier with high-pass and low-pass filters of 0.1 Hz and 30 Hz. The following 16 electrodes (according to the 10-20 international system) were used for signal recording and analysis: F3, Fz, F4, T7, C3, Cz, C4, T8, P7, P3, Pz, P4, P8, PO7, PO8, Oz, referenced to the average of the two mastoids and grounded to the electrode Fpz. Fig. 2 (b) shows the experimental layout which consists of eight arrow commands to simulate a wheelchair control. Each subject completed two experimental sessions, where each session consisted of eight runs. In each run, a randomly cued target arrow was first presented for 1 s in the middle of the screen followed by a 1 s black screen period, and a block-randomized sequence of stimulus intensifications was subsequently started. The intensification block was repeated five times, in each of which each arrow was intensified once with a duration of 100 ms and ISI of 80 ms. During the experiment, subjects were asked to focus attention on the target arrows and silently count the number of times they were intensified.

A 700 ms data segment after baseline corrected by 100 ms pre-stimulus interval was extracted from each stimulus intensification. A total of 640 such segments consisting of 80 targets and 560 non-targets were derived for each subject from the two experimental sessions, and each 40 segments (5 targets and 35 non-targets) corresponded to one command selection (16 command selections in total). Each segment was then downsampled to 21 Hz after 12-point moving average, thus the signal of each channel consisted of 15 points. For each subject, 8 command selections were randomly chosen from the 16 command selections for classifier training, while the remaining 8 command selections were used for test to evaluate the ERP classification performance. Such program was repeated 100 times and the average classification accuracy was then calculated.

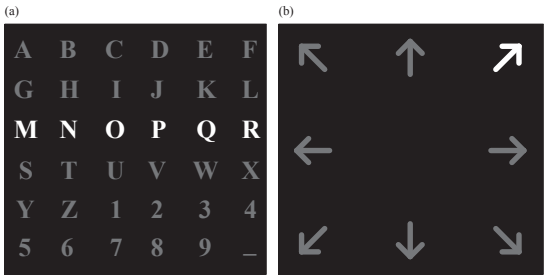


Fig. 2. P300 paradigms for recording the dataset-1 (a) and the dataset-2 (b).

B. Linear discriminant analysis (LDA)

LDA is probably the most popular algorithm for ERP classification in the BCI application, since it has relatively low computational requirement and usually provides good classification results [21]. Assume we are given a set of samples (i.e., feature vectors) $\mathbf{x}_i \in \mathbb{R}^D (D = MP)$ $(i = 1, 2, \ldots, N)$ where each sample is the concatenation of P temporal points from each of M channels, and the corresponding class labels $l_i \in \{1, 2\}$. The means and empirical covariance matrices of the two classes are computed from:

$$\begin{aligned} \boldsymbol{\mu}_c &= \frac{1}{N_c} \sum_{i \in \mathcal{I}_c} \mathbf{x}_i, \\ \boldsymbol{\Sigma}_c &= \frac{1}{N_c - 1} \sum_{i \in \mathcal{I}_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^T, c = 1, 2, \end{aligned} \quad (1)$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ denote respectively the target and non-target classes, N_c represents the number of samples in class $\mathcal{I}_c$. Estimate the common covariance matrix by $\boldsymbol{\Sigma} = \frac{N_1}{N} \boldsymbol{\Sigma}_1 + \frac{N_2}{N} \boldsymbol{\Sigma}_2$, the projection vector of the LDA is defined as:

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1). \quad (2)$$

Note that the LDA is equivalent to Fisher's discriminant analysis (FDA) and least squares regression [22], [30]. A potential problem of the LDA is that the covariance matrices $\boldsymbol{\Sigma}_c$ become ill-conditioned in small training sample size. This happens particularly for the classification of high-dimensional ERP features and results in poor accuracy.

C. Regularized versions of LDA

As two regularized versions of the LDA, the SWLDA and SKLDA try to mitigate effects of small training sample size on classification through reducing the feature dimensionality and shrinking the covariance matrix, respectively. Both of them have been increasingly applied to BCI [26], [31], [32], [33], and shown their superior powers over the traditional LDA for ERP classification [22].

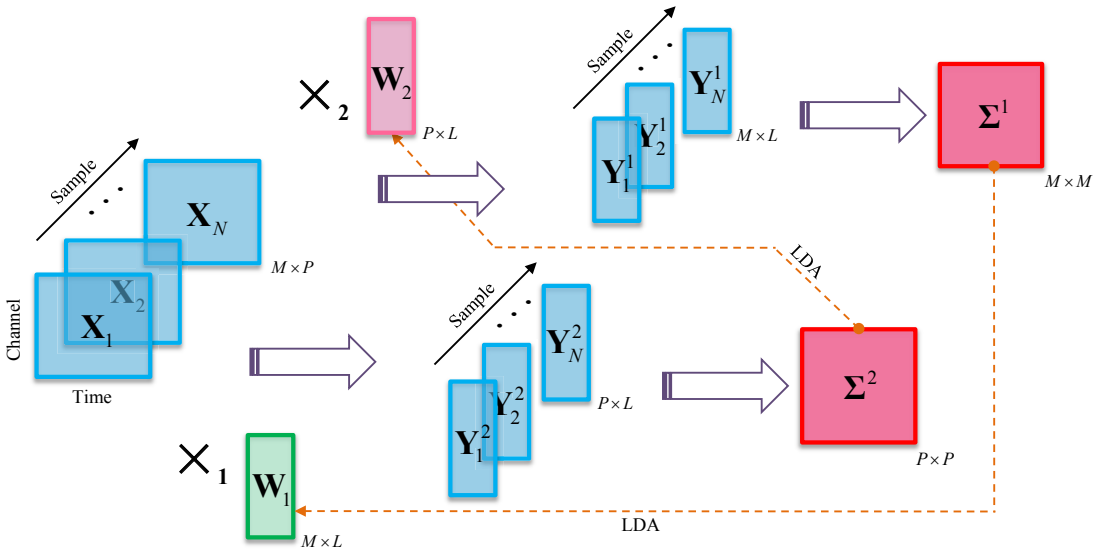


Fig. 3. Illustration of collaborative optimization procedure in the spatial-temporal discriminant analysis (STDA). With the constructed spatial-temporal two-way samples $\mathbf{X}_i \in \mathbb{R}^{M\times P}$ $(i = 1, 2, \ldots, N)$, the STDA method implements alternatingly discriminant analysis to learn projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ using a predefined low dimensionality L (typically, $L^2 \ll MP$) in the spatial and temporal dimensions. Due to the spatial-temporal construction, the dimensionalities of the covariance matrices $\mathbf{\Sigma}^1$ and $\mathbf{\Sigma}^2$ are $M \times M$ and $P \times P$ in the discriminant analysis, respectively, and both the required numbers $M \times (M-1)/2$ and $P \times (P-1)/2$ of parameters for estimation of the two covariance matrices are decreased compared to $MP \times (MP-1)/2$ in using the traditional LDA (cf. Fig. 1). Consequently, the new one-way samples transformed by Eq. (11) with the optimized $\mathbf{W}_1$ and $\mathbf{W}_2$ will have much lower dimensionality L^2 than that of the directly vectorized samples, which improves further the covariance matrix estimation in the subsequent ERP classification, since the required number of parameters is only $L \times (L-1)/2$.

1) Stepwise LDA (SWLDA): The SWLDA performs feature dimensionality reduction by selecting effective features to be included in the discriminant model. As three free parameters, two p-values controlling adding and removing of features, and the maximal number of features have to be predefined for the SWLDA. For comparison, the same settings as in [21] (add feature: p-value < 0.1; remove feature: p-value > 0.15; maximal number of features: 60) were used in this study.

2) Shrinkage LDA (SKLDA): The SKLDA improves the traditional LDA through adjusting the extreme eigenvalues of the covariance matrix towards the average eigenvalue [34], [35]. Given a shrinkage parameter $\lambda_c \in [0, 1]$, the poorly estimated covariance matrix $\mathbf{\Sigma}_c$ can be remedied as:

$$\tilde{\mathbf{\Sigma}}_c = (1-\lambda_c)\mathbf{\Sigma}_c + \lambda_c v_c \mathbf{I}, \tag{3}$$

where $v_c = \frac{tr(\mathbf{\Sigma}_c)}{D}$ denotes the average eigenvalue of $\mathbf{\Sigma}_c$ with D being the dimensionality of the feature space, and $\mathbf{I}$ is an identity matrix. Let $(\mathbf{x}_n)_i$ and $(\boldsymbol{\mu}_c)_i$ be the ith element of the feature vector $\mathbf{x}_n$ and mean vector $\boldsymbol{\mu}_c$, respectively, and denote by s_{ij} the element in the ith row and jth column of the covariance matrix $\mathbf{\Sigma}_c$, and define $z_{ij}(n) = ((\mathbf{x}_n)_i - (\boldsymbol{\mu}_c)_i)((\mathbf{x}_n)_j - (\boldsymbol{\mu}_c)_j)$, then the optimal shrinkage parameter can be analytically solved as [34]:

$$\lambda_c^* = \frac{N_c}{(N_c-1)^2} \frac{\sum_{i,j=1}^{D} \mathrm{var}_n(z_{ij}(n))}{\sum_{i\neq j} s_{ij}^2 + \sum_i (s_{ii} - v_c)^2}, \tag{4}$$

D. Spatial-temporal discriminant analysis (STDA)

Although the SWLDA and SKLDA outperform the traditional LDA for ERP classification with insufficient training samples, both of the two methods train the classifiers by still adopting vectorized (one-way) samples with high dimensionality (like the traditional LDA), which is actually the key factor resulting in the problem of small sample size within limited calibration time. Instead of using vectorized samples by concatenation of temporal points from multiple channels for ERP classification, we can construct spatial-temporal two-way samples as feature matrices $\mathbf{X}_i \in \mathbb{R}^{M\times P}$ $(i = 1, 2, \ldots, N)$ by keeping the spatial features in the first way (rows) while the temporal features in the second way (columns). Based on the spatial-temporal samples, we introduce a spatial-temporal discriminant analysis (STDA) to collaboratively learn two projection matrices from the spatial and temporal ERP feature subspaces. The learned projection matrices are then used to transform the original spatial-temporal samples to new one-way samples with much lower dimensionality than that of the directly vectorized samples used in the aforementioned LDA and regularized LDA algorithms. Estimation of the covariance matrices for ERP classification is significantly improved due to the collaboratively spatial-temporal optimization in the STDA method.

To describe the STDA method conveniently, we first define the following two operators:

$$\begin{aligned} \mathbf{X} \times_1 \mathbf{W}_1^T &= \mathbf{W}_1^T \mathbf{X}, \\ \mathbf{X} \times_2 \mathbf{W}_2^T &= \mathbf{X}\mathbf{W}_2, \end{aligned} \tag{5}$$

where $\mathbf{X} \in \mathbb{R}^{D_1\times D_2}$ is a spatial-temporal sample, $\mathbf{W}_1 \in \mathbb{R}^{D_1\times L}$ and $\mathbf{W}_2 \in \mathbb{R}^{D_2\times L}$ are the projection matrices in the spatial and temporal ways, respectively. Here, L is a lower dimensionality than the D_1 and D_2 (typically, $L^2 \ll D_1 D_2$).

The STDA method performs alternatingly optimization in the spatial and temporal ways of samples (see Fig. 3). For the

kth ($k = 1, 2$) way optimization, we first project each spatial-temporal sample $\mathbf{X}_i \in \mathbb{R}^{M\times P}$ at the $(3-k)$th way as:

$$\begin{cases} \mathbf{Y}_i^1 = \mathbf{X}_i \times_2 \mathbf{W}_2^T & \text{for } k = 1, \\ \mathbf{Y}_i^2 = \left(\mathbf{X}_i \times_1 \mathbf{W}_1^T\right)^T & \text{for } k = 2. \end{cases} \tag{6}$$

The between-class scatter matrix $\mathbf{S}_B^k$ and within-class scatter matrix $\mathbf{S}_W^k$ of the projected samples are then computed by:

$$\begin{aligned} \mathbf{S}_B^k &= \sum_{c=1}^{2} N_c(\bar{\mathbf{Y}}_c^k - \bar{\mathbf{Y}}^k)(\bar{\mathbf{Y}}_c^k - \bar{\mathbf{Y}}^k)^T, \\ \mathbf{S}_W^k &= \sum_{c=1}^{2} \sum_{i\in\mathcal{I}_c} (\mathbf{Y}_{c,i}^k - \bar{\mathbf{Y}}_c^k)(\mathbf{Y}_{c,i}^k - \bar{\mathbf{Y}}_c^k)^T, \end{aligned} \tag{7}$$

where $\mathbf{Y}_{c,i}^k$ is the ith projected sample in class c, $\bar{\mathbf{Y}}_c^k$ and $\bar{\mathbf{Y}}^k$ denote mean of the projected samples in class c and mean of all the projected samples, respectively, and are computed as:

$$\bar{\mathbf{Y}}_c^k = \frac{1}{N_c} \sum_{i\in\mathcal{I}_c} \mathbf{Y}_i^k \quad \text{and} \quad \bar{\mathbf{Y}}^k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_i^k. \tag{8}$$

Following the traditional LDA algorithm, the learned projection matrix in the kth way is given by:

$$\tilde{\mathbf{W}}_k = \arg\max_{\mathbf{W}_k} \frac{tr(\mathbf{W}_k^T \mathbf{S}_B^k \mathbf{W}_k)}{tr(\mathbf{W}_k^T \mathbf{S}_W^k \mathbf{W}_k)}, \tag{9}$$

which can be solved by generalized eigenvalue problem:

$$\mathbf{S}_B^k \mathbf{W}_k = \mathbf{S}_W^k \mathbf{W}_k \mathbf{\Lambda}_k, \tag{10}$$

where the eigenvectors in $\mathbf{W}_k$ corresponding to the largest L eigenvalues in $\mathbf{\Lambda}_k$ are retained to form the learned projection matrix $\tilde{\mathbf{W}}_k$. To solve the optimized projection matrices in the spatial and temporal ways, an iteration procedure is implemented for alternating learning of projection matrices in the two ways.

After the optimized projection matrices $\tilde{\mathbf{W}}_1 \in \mathbb{R}^{M\times L}$ and $\tilde{\mathbf{W}}_2 \in \mathbb{R}^{P\times L}$ are obtained, each spatial-temporal sample $\mathbf{X}_i \in \mathbb{R}^{M\times P}$ is transformed to a new one-way sample $\mathbf{f}_i \in \mathbb{R}^{L^2}$ ($L^2 \ll MP$) as:

$$\mathbf{f}_i = \text{vec}(\mathbf{X}_i \times_1 \tilde{\mathbf{W}}_1^T \times_2 \tilde{\mathbf{W}}_2^T), \quad i = 1, 2, \ldots, N, \tag{11}$$

where $\text{vec}(\cdot)$ denotes the vectorization operator. With the transformed one-way samples $\mathbf{f}_i$ and corresponding class labels $l_i \in \{1, 2\}$ ($i = 1, 2, \ldots, N$), a discriminant projection vector $\mathbf{w}_\mathbf{f} \in \mathbb{R}^{L^2}$ can be solved by simply using the LDA formulated in Eq. (1) and (2). Regarding a new test sample $\hat{\mathbf{X}} \in \mathbb{R}^{M\times P}$, we first obtain the one-way sample $\hat{\mathbf{f}} \in \mathbb{R}^{L^2}$ transformed from $\hat{\mathbf{X}}$ by Eq. (11), and then calculate its classification score as:

$$H = \mathbf{w}_\mathbf{f}^T \hat{\mathbf{f}}. \tag{12}$$

Note that the bias term is ignored here, since it is invariant for all test samples. The larger classification score represents more strongly the characteristic P300 as defined by the training data [21]. The bias term is also not considered for the other methods compared in this study.

Algorithm 1 depicts the pseudo-code about the alternating optimization procedure of the STDA method, extraction for features with much lower dimensionality, and subsequent classification for ERP.

Algorithm 1: Spatial-temporal discriminant analysis algorithm for ERP classification

Input: A set of EEG spatial-temporal samples $\mathbf{X}_i \in \mathbb{R}^{D_1\times D_2}$, the corresponding class labels $l_i \in \{1, 2\}$, $i \in \{1, 2, \ldots, N\}$, the number L of eigenvectors retained for projection matrices learning, and a test EEG sample $\hat{\mathbf{X}} \in \mathbb{R}^{D_1\times D_2}$.

Output: Classification score H of $\hat{\mathbf{X}}$.

Initialization for $\mathbf{W}_2 = \mathbf{I}_{D_2}$;

repeat

  for $k = 1$ *to* 2 **do**

    $\mathbf{Y}_i^k = \mathbf{X}_i \times_{3-k} \mathbf{W}_{3-k}^T$, $i = 1, 2, \ldots, N$

    if $k == 2$ **do** $\mathbf{Y}_i^k = \left(\mathbf{Y}_i^k\right)^T$ **end**

    $\bar{\mathbf{Y}}_c^k = \frac{1}{N_c} \sum_{i\in\mathcal{I}_c} \mathbf{Y}_i^k, \bar{\mathbf{Y}}^k = \frac{1}{N} \sum_{i=1}^{N} \mathbf{Y}_i^k$

    $\mathbf{S}_B^k = \sum_{c=1}^{2} N_c(\bar{\mathbf{Y}}_c^k - \bar{\mathbf{Y}}^k)(\bar{\mathbf{Y}}_c^k - \bar{\mathbf{Y}}^k)^T$

    $\mathbf{S}_W^k = \sum_{c=1}^{2} \sum_{i\in\mathcal{I}_c} (\mathbf{Y}_{c,i}^k - \bar{\mathbf{Y}}_c^k)(\mathbf{Y}_{c,i}^k - \bar{\mathbf{Y}}_c^k)^T$

    Solve generalized eigenvalue problem: $\mathbf{S}_B^k \mathbf{W}_k = \mathbf{S}_W^k \mathbf{W}_k \mathbf{\Lambda}_k$, obtain $\mathbf{W}_k \in \mathbb{R}^{D_k\times L}$

  end

until *Stop criterion is met*;

$\tilde{\mathbf{W}}_1 = \mathbf{W}_1$, $\tilde{\mathbf{W}}_2 = \mathbf{W}_2$

$\mathbf{f}_i = \text{vec}(\mathbf{X}_i \times_1 \tilde{\mathbf{W}}_1^T \times_2 \tilde{\mathbf{W}}_2^T)$, $i = 1, 2, \ldots, N$

$\mathbf{w}_\mathbf{f} = \text{LDA}([\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N], [l_1, l_2, \ldots, l_N])$

$\hat{\mathbf{f}} = \text{vec}(\hat{\mathbf{X}} \times_1 \tilde{\mathbf{W}}_1^T \times_2 \tilde{\mathbf{W}}_2^T)$

$H = \mathbf{w}_\mathbf{f}^T \hat{\mathbf{f}}$.

It is worth noting that a Fisher's criterion (FC)-based spatial filtering method for ERP classification introduced in [36] could be regarded as a special case of the dimensionality reduction procedure described in the proposed STDA method. If we execute the spatial way optimization in the STDA method only once instead of alternating iteration, the same result as that of the FC-based spatial filtering method could be straightway obtained. The FC-based spatial filtering has shown stronger denoising capability than the common spatial pattern (CSP) for ERP classification [36]. In addition to the function of denoising, FC-based spatial filtering could also be interpreted as a supervised dimensionality reduction method, and hence may alleviate the effect of small training sample size on ERP classification to some extent. In this study, we also compared the proposed STDA method to the classification method with FC-based spatial filtering which was referred as FC+LDA method and the classification method with CSP-based spatial filtering which was referred as CSP+LDA.

III. Results

A. Single-trial classification performance

With the dataset-1, we investigated performance of the proposed STDA algorithm for single-trial ERP classification.

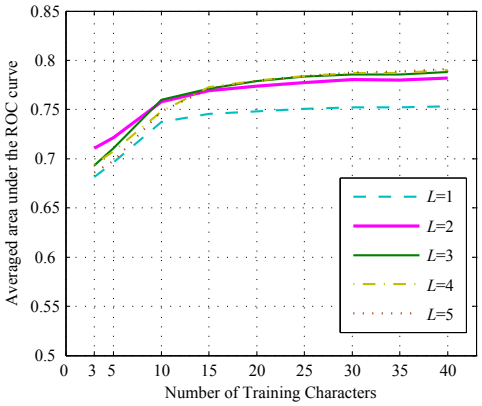


Fig. 4. Averaged areas under the ROC curves when using various numbers of character spellings for classifier training with the STDA algorithm, and retaining one to five eigenvectors for projection matrices learning. Note that each character spelling contains 180 samples.

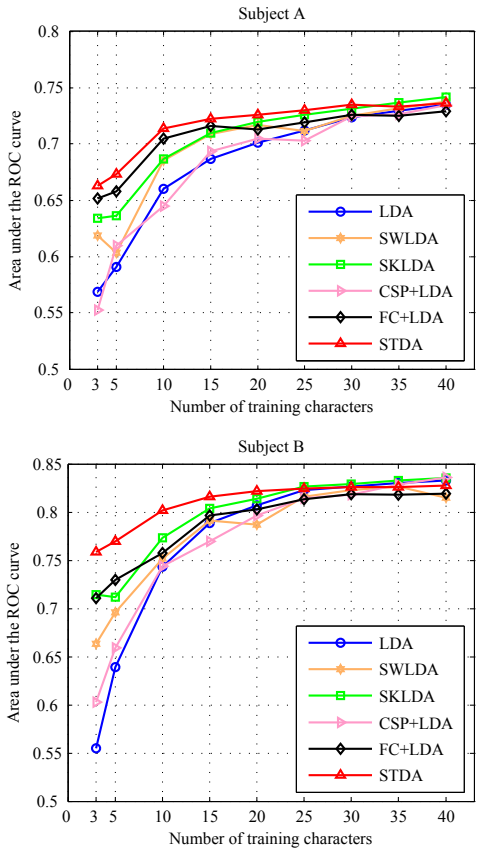


Fig. 5. Areas under the ROC curves obtained by the LDA, SWLDA, SKLDA, CSP+LDA, FC+LDA, and STDA with various numbers of training character spellings, for the subjects A and B. Note that each character spelling contains 180 samples.

Instead of classification accuracy, the area under the receiver operating characteristic (ROC) curve was adopted to evaluate the classification performance, since the numbers of target and non-target segments were not equal. The ROC curve [37] is generated by plotting the fraction of true positive rate (TPR) against the fraction of false positive rate (FPR) as the threshold for discrimination between two classes is varied. The area under the ROC curve is calculated by using trapezoidal integration [38].

We first analyzed how changes for the number L of retained eigenvectors during projection matrices learning in the STDA method could affect the subsequent classification performance, especially when using few training samples. Fig. 4 depicts the averaged areas under the ROC curves obtained by using various numbers of character spellings (each character spelling contains 180 samples) for classifier training when changing L from one to five. With sufficient training samples, the classification performance had no big difference when choosing L from 2 to 5, while the classification performance was obviously improved by choosing larger L compared to 1. With few training samples (e.g., the number of training character spellings is less than 10), the classification performance achieved the best by choosing $L = 2$. Therefore, we adopted $L = 2$ for subsequent analysis since the aim of this study is to decrease the required number of training samples for classifier training with minimal degradation of classification accuracy.

To validate the single-trial classification performance, the proposed STDA was compared with the traditional LDA, SWLDA, SKLDA, CSP+LDA, and FC+LDA. Fig. 5 shows the areas under the ROC curves obtained by the aforementioned methods with various numbers of training samples. With few training samples, the STDA outperformed all the other classification methods, while all the modified classification methods, except for the CSP+LDA, yielded obviously better classification performances than that of the traditional LDA. With increasing of the number of training samples, classification performance differences among the five classification methods trended to be smaller.

B. Target detection accuracy with few training samples

For the dataset-2, the number of training samples is insufficient (320) compared to the features dimensionality of 240 (16 channels$\times$15 points), since the required number of parameters for the covariance matrix estimation is 240$\times$239/2 = 28680 when using the LDA. The estimated covariance matrix is consequently ill-conditioned. With this dataset, we investigated and compared the target detection accuracies of the LDA, SWLDA, SKLDA, CSP+LDA, and FC+LDA methods by using various numbers of trials average (see Fig. 6). The proposed STDA method achieved better average target detection accuracy than those of the other methods across different numbers of trials average, and such superiority was more prominent for the S3 and S5 who had poorer target detection accuracies compared to other subjects.

The paired t-tests with Bonferroni correction were adopted for further analysis of accuracy difference between the STDA method and each of the other methods. Table I shows the results of statistical analysis. The STDA method yielded significantly higher target detection accuracy than most of the other methods when using one to four trials average. Especially, the STDA outperformed significantly all the other methods using two trials average.

C. Online target detection accuracy

Further online experiments were implemented to validated the proposed STDA method. Five healthy subjects (all males,

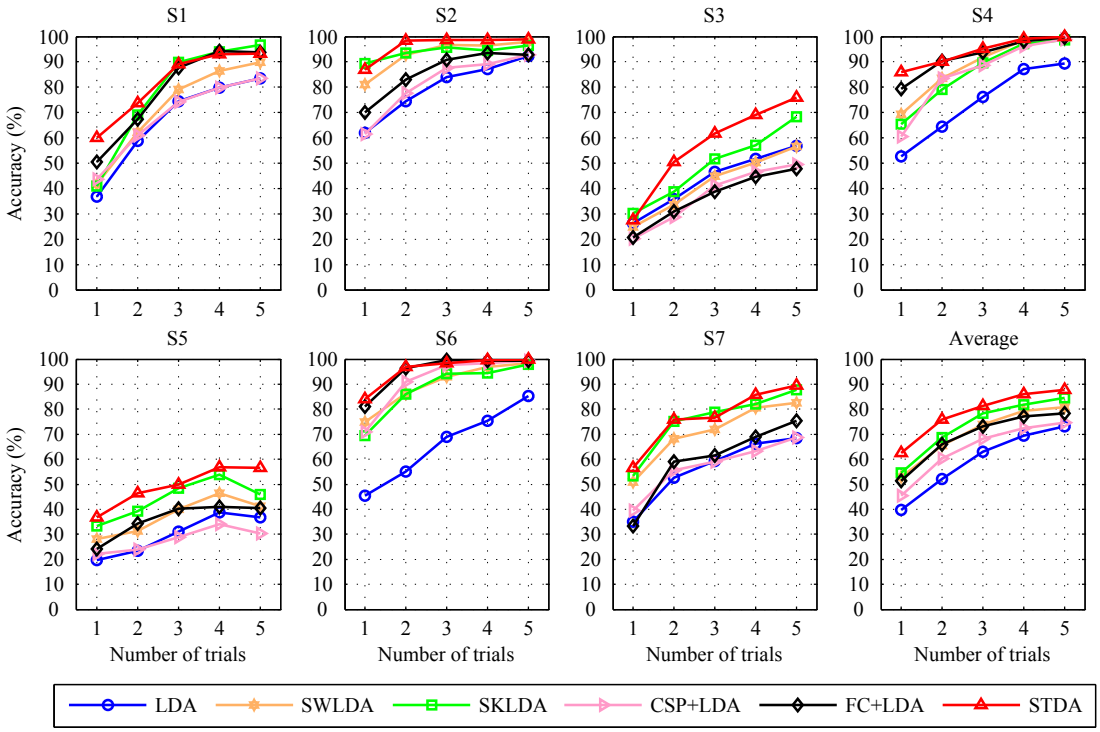


Fig. 6. Target detection accuracies obtained by the LDA, SWLDA, SKLDA, CSP+LDA, FC+LDA and STDA methods using one to five trials average for the seven subjects. Superior average accuracy was achieved by the STDA over those of the other methods.

TABLE I
STATISTICAL ANALYSIS FOR ACCURACY DIFFERENCE BETWEEN THE STDA AND EACH OF THE OTHER METHODS BY THE PAIRED T-TEST WITH BONFERRONI CORRECTION.

Method Comparison	Number of trials				
	1	2	3	4	5
STDA vs. LDA	†	††	††	††	††
STDA vs. SWLDA	†	††	**	*	∼
STDA vs. SKLDA	∼	†	∼	*	∼
STDA vs. CSP+LDA	††	†	†	**	*
STDA vs. FC+LDA	**	*	*	∼	∼

Note: ∼ nonsignificant, $*p < 0.05$, $**p < 0.01$, $\dagger p < 0.005$, $\dagger\dagger p < 0.001$

TABLE II
ONLINE TARGET DETECTION ACCURACY (%) OBTAINED BY THE LDA, SWLDA, SKLDA, FC+LDA AND STDA USING TWO TRIALS AVERAGE FOR THE FIVE SUBJECTS

Subject	Method				
	LDA	SWLDA	SKLDA	FC+LDA	STDA
ZY	54.2	66.7	70.8	58.3	83.3
MJ	50.0	58.3	66.7	58.3	79.2
ZQ	54.2	79.2	75.0	70.8	87.5
GX	54.2	70.8	70.8	75.0	83.3
NY	37.5	50.0	66.7	50.0	70.8
Average	50.0±7.23	65.0±11.3	70.0±3.47	62.5±10.2	80.8±6.32

aged from 26 to 34) participated in the online experiments. The same experimental settings as those for dataset-2 acquisition were adopted to record the training data from eight arrow command selections. For each subject, a total of 320 segments (40 targets and 280 non-targets) were derived as training data which were insufficient for covariance matrix estimation in the LDA. Five classifiers were then trained with the training data using the LDA, SWLDA, SKLDA, FC+LDA and STDA, respectively, and applied to the subsequent online tests. In the online tests, each classification method was tested with three experimental sessions and a total of 15 sessions were completed for each subject. The 15 sessions were randomly conducted with a five minutes break between each two sessions. For each subject, eight arrow commands were selected in each session and 24 arrow commands were selected during each classification method test. Two trials average was adopted for online target detection. Table II shows the online target detection accuracy. The paired t-tests revealed that the proposed STDA achieved significantly higher online accuracy for target detection than those of the other methods with insufficient training samples (STDA>LDA: p<0.001, STDA>SWLDA: p<0.005, STDA>SKLDA: p<0.005, STDA>FC+LDA: p<0.005).

IV. DISCUSSION

A. System performance

From classification performance of single-trial ERP and target detection accuracies with various numbers of trials average, we investigated the effects of the proposed STDA method on calibration time reduction and classification accuracy improvement for the ERP-based BCI.

For the dataset-1, the STDA method yielded smaller degradation for the classification performance of single-trial ERP compared with the LDA, SWLDA, SKLDA, CSP+LDA and FC+LDA, when using fewer training character spellings (see Fig. 5). This suggests that the STDA is effective to reduce

calibration time for the single-trial ERP classification. For instance, the STDA decreases about 10 training characters (i.e., more than 5 minutes) to achieve the same single-trial ERP classification performance as the traditional LDA. Under the computation environment of Matlab R2009a on a laptop with 2.80 GHz CPU, the computational time (s) is 1.80 for the LDA, 2.84 for the SWLDA, 4.49 for the SKLDA, 0.263 for the CSP+LDA, 0.121 for the FC+LDA, and 0.702 for the STDA to train the classifiers with training data recorded from 10 character spellings. While the FC+LDA, the CSP+LDA and the STDA improve the computational efficiency in contrast to the LDA, SWLDA and SKLDA, all of the computational time are not taken into account for evaluating the system calibration time, since they are very short compared with the whole calibration time and thus could be ignored.

For the dataset-2, the calibration procedure requires eight training command selections with five trials for each of them, which takes about 73.6 s. With the system calibration time little longer than one minute, the STDA significantly outperformed all the other methods (see Fig. 6). Also, a significantly higher average accuracy 80.8 % for online target detection was achieved by the STDA compared to 50.0 % of the LDA, 65.0 % of the SWLDA, 70.0 % of the SKLDA, and 62.5 % of the FC+LDA when using two trials average (see Table II). The outstanding target detection accuracy with such short system calibration time suggests that the STDA is effective to reduce system calibration time of the ERP-based BCI, and hence improve the system practicability and encourage the users to use the BCI system. The aforementioned results also imply that the STDA algorithm enhances the classification performance using average of fewer trials when the number of training samples is limited. This could be of much practical value for the ERP-based BCI, since ERP is relatively weak and usually difficult to be accurately detected by using average of only a few trials.

B. Convergence and robustness

Since the proposed STDA algorithm involves an iterative optimization procedure, a stop criterion should be appropriately chosen to guarantee convergence of the iteration process. The stop criterion is defined as:

$$Error = \|\mathbf{W}(n) - \mathbf{W}(n-1)\|_2 < 10^{-5}, \qquad (13)$$

where n denotes the number of iteration steps. To confirm the validity and stability of the proposed stop criterion, we repeatedly execute the STDA algorithm (100 iteration steps for each execution) on different ten training data sets randomly selected from the EEG data of S2 in the dataset-2. Ten convergence curves are then plotted for each of the two projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ (see Fig. 7). The results show that all the ten executions of STDA algorithm converged ($Error$<10^{-10}) within 100 iteration steps for both the $\mathbf{W}_1$ and $\mathbf{W}_2$, and the convergence is quite stable. To achieve a trade-off between efficiency and accuracy, our experiment results show that $Error$<10^{-5} is an appropriate choice.

Another important issue is whether the STDA algorithm can be robust to the inter-trial variability of ERP features. From the

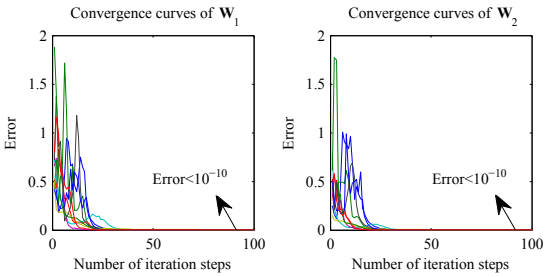


Fig. 7. Convergence curves of projection matrices $\mathbf{W}_1$ and $\mathbf{W}_2$ obtained by repeatedly executing the STDA algorithm on different ten training data sets randomly selected from the EEG data of S2 in the dataset-2. Each convergence curve was formed by running the collaborative optimization procedure in the STDA algorithm for 100 iterations.

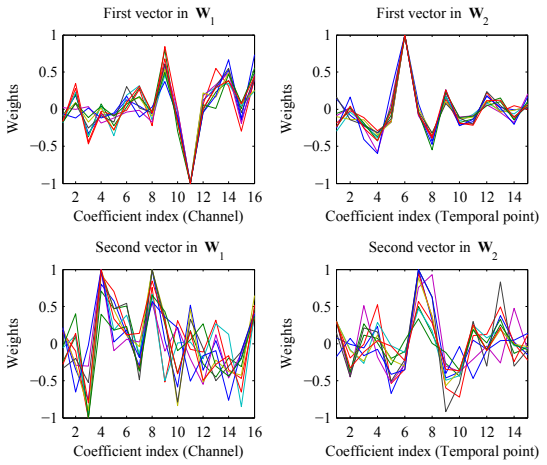


Fig. 8. Projection matrices estimated by repeatedly executing the STDA algorithm on different ten training data sets randomly selected from the EEG data of S2 in the dataset-2.

aforementioned repeat executions of the STDA algorithm on different ten training data sets, we learned ten results for each of the two projection matrices (see Fig. 8). The weights of $\mathbf{W}_1$ and $\mathbf{W}_2$ obtained from the ten training data sets are almost similar, which indicates that the STDA algorithm provides good generalization capacity without overtraining.

C. Estimation of covariance matrices

Instead of vectorized one-way samples, the STDA method employs spatial-temporal two-way samples to collaboratively learn two projection matrices from the spatial and temporal feature subspaces, which reduces significantly feature dimensionalities and improves estimation of covariance matrices in the spatial and temporal discriminant analysis. With the learned projection matrices, the spatial-temporal two-way samples are transformed to new one-way samples with much lower dimensionality which improves further the covariance matrix estimation for subsequent ERP classification. As an example, we show how the STDA method improves estimation of covariance matrices compared to the traditional LDA with the dataset-2. The number of training samples (320) is insufficient and much lower than the required number of parameters (240×239/2 = 28680) for the covariance matrix estimation when using the LDA to train classifier. Consequently, the esti-

mated covariance matrix is ill-conditioned and results in poor generalization capability of the classifier. However, with the STDA method, both of the required numbers of parameters for estimation of the covariance matrices in the spatial (16×15/2 = 120) and temporal (15×14/2 = 105) discriminant analysis are less than the number of training samples. After the two projection matrices are estimated, the dimensionality of each new one-way sample transformed by Eq. (11) is only 4, since the number of eigenvectors retained for the projection matrices learning is selected to be 2. The number of training samples is fully sufficient for the parameters estimation (4×3/2 = 6) of covariance matrix in the subsequent ERP classification. Therefore, estimation of covariance matrices in the STDA is much improved over that of the traditional LDA through collaborative discriminant analysis in the spatial and temporal feature subspaces. The well-estimated covariance matrices by the STDA enhance generalization capability of the classifier, thereby assisting to yield better ERP classification accuracy.

D. Multiway optimization

Both the STDA and the FC+LDA methods have retained spatial and temporal dimensions which provide more natural representation of the original EEG data structure. However, the STDA implements alternatingly spatial and temporal dimensions optimization (see Fig. 3 and Algorithm 1) while the FC+LDA optimizes spatial dimension only. Collaborative optimization in multi-dimension has been suggested to be probably more promising for EEG data analysis compared to one-dimension optimization [39], [40]. Squared pointwise biserial correlation coefficients (r^2-values) [22] were adopted to evaluate the discriminative information derived from the STDA and FC+LDA, respectively. As a discrimination index, the pointwise biserial correlation coefficient is defined as:

$$r(x) = \frac{\sqrt{N_1 N_2}}{N_1 + N_2} \frac{\text{mean}\{x_i | l_i = 1\} - \text{mean}\{x_i | l_i = 2\}}{\text{std}\{x_i | l_i = 1, 2\}}, \quad (14)$$

where N_1 and N_2 are the numbers of variables belong to the class 1 (target) and class 2 (non-target), respectively, x_i and l_i are the value and class label of the ith variable, respectively, and the r^2-value is equal to the square of $r(x)$. Larger r^2-value indicates higher separability of distributions. Fig. 9 shows that the discriminative information of features extracted by the STDA is much improved over that by the FC+LDA. The most discriminative two features were adopted to depict the feature distributions derived by the FC+LDA and STDA methods (see Fig. 10). The target and non-target classes present larger between-class scatter but smaller within-class scatter for the STDA in contrast to the FC+LDA, which provides an evidence for the better ERP classification performance achieved by the STDA over the FC+LDA. It is worth noting that we only consider the spatial and temporal dimensions of EEG signals in this study, since P300 characteristics are relatively more prominent in these two dimensions. The spatial-temporal analysis could be further extended to higher-order analysis by constructing higher-dimensional (i.e., tensor) samples which provide multiway array presentation for the EEG data structure and include more neurophysiological meanings [39], [41], [42], [43].

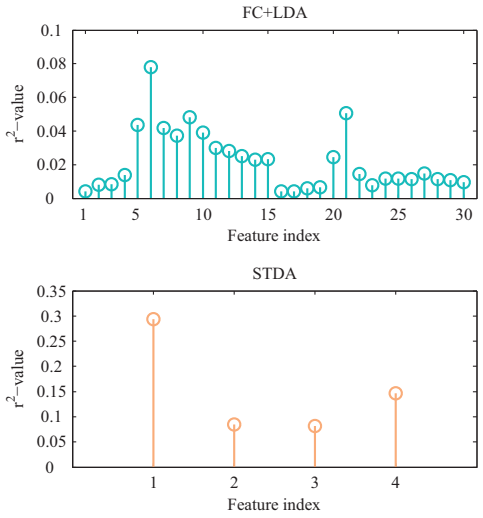


Fig. 9. Discriminative information of features extracted by the FC+LDA and STDA methods, evaluated by squared pointwise biserial correlation coefficients (r^2-values).

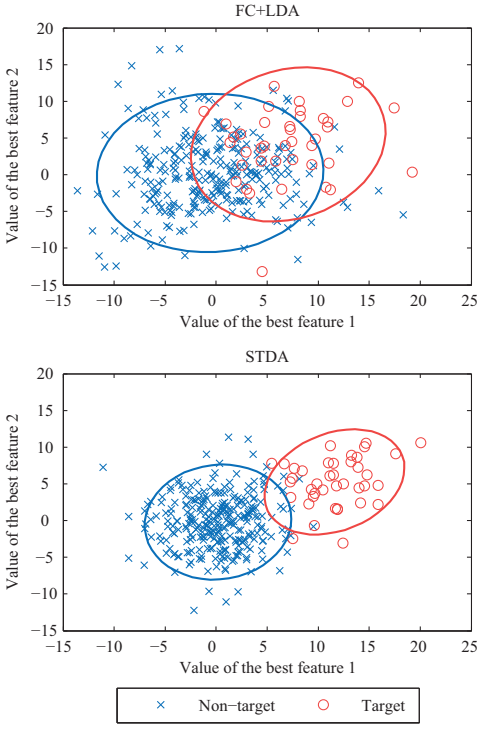


Fig. 10. Distributions depicted with the most discriminative two features extracted by the FC+LDA (6th and 21th features) and STDA (1th and 4th features) (see Fig. 9), respectively, from the EEG data of S2 in the dataset-2.

V. Conclusions

In this study, we introduced a method of spatial-temporal discriminant analysis (STDA) to ERP classification in the BCI application. As a multiway extension of the LDA, the STDA method implements collaboratively discriminant analysis in the spatial and temporal dimensions of constructed spatial-temporal two-way EEG samples, which reduces significantly feature dimensionalities, and hence improves estimation of covariance matrices in the discriminant analysis from limited

number of training samples. This assists to enhance generalization capability of the trained classifier. Consequently, the system calibration time of the ERP-based BCI is effectively reduced with improvement in classification accuracy by using the proposed STDA method in contrast to the state-of-the-art methods for ERP classification. Future studies will try to optimize further the STDA method and investigate its application in other types of BCIs.

REFERENCES

[1] J. Wolpaw, N. Birbaumer, G. Pfutscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," *Proc. IEEE*, vol. 89, no. 7, pp. 1123–1134, 2001.

[3] B. Blankertz, F. Losch, M. Krauledat, G. Dornhege, G. Curio, and K. Müller, "The Berlin brain–computer interface: Accurate performace from first-session in BCI-naïve subjects," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 10, pp. 2452–2462, 2008.

[4] C. Neuper, G. Müller-Putz, A. Kübler, N. Birbaumer, and G. Pfurtscheller, "Clinical application of an EEG-based brain-computer interface: a case study in a patient with severe motor impairment," *Clin. Neurophysiol.*, vol. 114, no. 3, pp. 399–409, 2003.

[5] R. Leeb, F. Lee, C. Keinrath, R. Scherer, H. Bischof, and G. Pfurtscheller, "Brain-computer communication: Motivation, aim, and impact of exploring a virtual apartemt," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 473–482, 2007.

[6] X. Gao, D. Xu, M. Cheng, and S. Gao, "A BCI-based environmental controller for the motion-disabled," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 137–140, 2003.

[7] B. Allison, D. McFarland, G. Schalk, S. Zheng, M. Jackson, and J. Wolpaw, "Towards an independent brain-computer interface using steady state visual evoked potentials," *Clin. Neurophysiol.*, vol. 119, no. 2, pp. 399–408, 2008.

[8] G. Bin, X. Gao, Z. Yan, B. Hong, and S. Gao, "An online multi-channel SSVEP-based brain-computer interface using a canonical correlation anaylsis method," *J. Neural Eng.*, vol. 6, no. 4, p. 046002, 2009.

[9] B. Allison, T. Lüth, D. Valbuena, A. Teymourian, I. Volosyak, and A. Gräser, "BCI Demographics: How many (and what kinds of) people can use an SSVEP BCI?" *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 18, no. 2, pp. 107–116, 2010.

[10] C. Brunner, B. Allison, C. Altstätter, and C. Neuper, "A comparison of three brain-computer interfaces based on event-related desynchronization, steady state visual evoked potentials, or a hybrid approach using both signals," *J. Neural Eng.*, vol. 8, no. 2, p. 025010, 2011.

[11] R. Fisher, G. Harding, G. Erba, G. Barkley, and A. Wilkins, "Photic- and pattern-induced seizures: A review for the epilepsy foundation of america working group," *Epilepsia*, vol. 46, no. 9, pp. 1426–1441, 2005.

[12] L. Farwell and E. Donchin, "Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, no. 6, pp. 510–523, 1988.

[13] S. Sutton, M. Braren, J. Zubin, and E. John, "Evoked-potential correlates of stimulus uncertainty," *Science*, vol. 150, no. 3700, pp. 1187–1188, 1965.

[14] E. Sellers and E. Donchin, "A P300-based brain-computer interface: Initial tests by ALS patients," *Clin. Neurophysiol.*, vol. 117, no. 3, pp. 538–548, 2006.

[15] J. Jin, B. Allison, E. Sellers, C. Brunner, P. Horki, X. Wang, and C. Neuper, "Optimized stimulus presentation patterns for an event-related potential EEG-based brain-computer interface," *Med. Biol. Eng. Comput.*, vol. 49, no. 2, pp. 181–191, 2010.

[16] B. Hong, F. Guo, T. Liu, X. Gao, and S. Gao, "N200-speller using motion-onset visual response," *Clin. Neurophysiol.*, vol. 120, no. 9, pp. 1658–1666, 2009.

[17] Y. Zhang, Q. Zhao, J. Jin, X. Wang, and A. Cichocki, "A novel BCI based on ERP components sensitive to configural processing of human faces," *J. Neural Eng.*, vol. 9, no. 2, p. 026018, 2012.

[18] T. Kaufmann, S. Schulz, C. Grünziger, and A. Kübler, "Flashing characters with famous faces improves ERP-based brain-computer interface performance," *J. Neural Eng.*, vol. 8, no. 5, p. 056016, 2011.

[19] J. Jin, B. Allison, X. Wang, and C. Neuper, "A combined brain-computer interface based P300 potentials and motion-onset visual evoked potentials," *J. Neurosci. Meth.*, vol. 205, no. 2, pp. 265–276, 2012.

[20] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, 2007.

[21] D. Krusienski, E. Sellers, F. Cabestaing, S. Bayoudh, D. McFarland, T. Vaughan, and J. Wolpaw, "A comparison of classification techniques for the P300 speller," *J. Neural Eng.*, vol. 3, no. 4, pp. 299–305, 2006.

[22] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K. Müller, "Single-trial analysis and classification of ERP components – A tutorial," *NeuroImage*, vol. 56, no. 2, pp. 814–825, 2011.

[23] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, pp. 252–264, 1991.

[24] Y. Li and C. Guan, "An extended EM algorithm for joint feature extraction and classification in brain-computer interfaces," *Neural Comput.*, vol. 18, no. 11, pp. 2730–2761, 2006.

[25] ——, "Joint feature re-extraction and classification using an iterative semi-supervised support vector machine algorithm," *Mach. Learn.*, vol. 71, no. 1, pp. 33–53, 2008.

[26] D. Krusienski, E. Sellers, D. McFarland, T. Vaughan, and J. Wolpaw, "Toward enhanced P300 speller performance," *J. Neurosci. Meth.*, vol. 167, no. 1, pp. 15–21, 2008.

[27] C. Vidaurre, N. Krämer, B. Blankertz, and S. A., "Time domain parameters as a feature for EEG-based brain computer interfaces," *Neural Netw.*, vol. 22, no. 9, pp. 1313–1319, 2009.

[28] S. Yan, D. Xu, Q. Yang, L. Zhang, X. Tang, and H. Zhang, "Multilinear discriminant analysis for face recognition," *IEEE Trans. Image Processing*, vol. 16, no. 1, pp. 212–220, 2007.

[29] B. Blankertz, K. Müller, D. Krusienski, G. Schalk, J. Wolpaw, A. Schloegl, G. Pfurtscheller, J. Millan, M. Schroeder, and N. Birbaumer, "The BCI competition III: Validating alternative approaches to actual BCI problem," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 153–159, 2006.

[30] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley and Sons, 2001.

[31] G. Townsend, B. LaPallo, C. Boulay, D. Krusuenski, G. Frye, C. Hauser, N. Schwartz, T. Vaughan, J. Wolpaw, and E. Sellers, "A novel P300-based brain-computer interface stimulus presentation paradigm: Moving beyond rows and columns," *Clin. Neurophysiol.*, vol. 121, no. 7, pp. 1109–1120, 2010.

[32] J. Hohne, M. Treder, B. Blankertz, and M. Tangermann, "Performance optimization of ERP-based BCIs using dynamic stopping," *In Conf. Proc. IEEE Eng. Med. Bio. Soc.*, pp. 4580–4583, 2011.

[33] L. Acqualagna and B. Blankertz, "A gaze independent spelling based on rapid serial visual presentation," *In Conf. Proc. IEEE Eng. Med. Bio. Soc.*, pp. 4560–4563, 2011.

[34] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for function genomics," *Stat. Appl. Genet. Mol. Biol.*, vol. 4, no. 1, p. Article 32, 2005.

[35] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *J. Multivar. Anal.*, vol. 88, no. 2, pp. 365–411, 2004.

[36] G. Pires, U. Nunes, and M. Castelo-Branco, "Statistical spatial filtering for a P300-based BCI: Test in able-bodied, and patients with cerebral palsy and amyotrophic lateral sclerosis," *J. Neurosci. Meth.*, vol. 195, no. 2, pp. 270–281, 2011.

[37] M. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press, 2003.

[38] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recogn.*, vol. 30, no. 7, pp. 1145–1159, 1997.

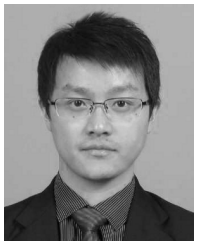
[39] A. Cichocki, Y. Washizawa, T. Rutkowski, H. Bakardjian, P. A.H., S. Choi, H. Lee, Q. Zhao, L. Zhang, and Y. Li, "Noninvasive BCIs: Multiway signal-processing array decompositions," *IEEE Computer*, vol. 41, no. 10, pp. 34–42, 2008.

[40] W. Wu, Z. Chen, S. Gao, and E. Brown, "A hierarchical Bayesian approach for learning sparse spatio-temporal decompositions of multichannel EEG," *NeuroImage*, vol. 56, no. 4, pp. 1929–1945, 2011.

[41] A. Cichocki, R. Zdunek, A. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorization: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley and Sons, 2009.

[42] D. Tao, X. Li, X. Wu, and S. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 10, pp. 1700–1714, 2007.

[43] J. Li and L. Zhang, "Regularized tensor discriminant analysis for single trial EEG classification in BCI," *Pattern Recogn. Lett.*, vol. 31, no. 7, pp. 619–628, 2010.

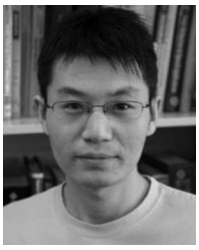


Yu Zhang received the B.Sc. degree in electrical engineering and automation from East China University of Science and Technology, Shanghai, China, in 2008, and is currently working toward the Ph.D. degree at the School of Information Science and Engineering at East China University of Science and Technology, and also as an International Program Associate in the Laboratory for Advanced Brain Signal Processing at RIKEN Brain Science Institute, Wako-shi, Japan.

His research interests include brain-computer interface, signal processing, tensor analysis, machine learning and pattern recognition.

Guoxu Zhou was born in Hubei Province, China, in 1977. He received his Ph.D degree in intelligent signal and information processing from South China University of Technology, Guangzhou, China, in 2010. He is currently a research scientist of the laboratory for Advanced Brain Signal Processing, at RIKEN Brain Science Institute (Japan).

His research interests include statistical signal processing, tensor analysis, intelligent information processing, and machine learning.



Qibin Zhao received the Ph.D. degree in engineering from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a Research Scientist at the Laboratory for Advanced Brain Signal Processing, Brain Science Institute, RIKEN, Wako-shi, Japan.

His current research interests include multiway data analysis, brain-computer interface, machine learning and pattern recognition.

Jing Jin Jing Jin received the Ph.D. degree in control theory and control engineering from the East China University of Science and Technology, Shanghai, China, in 2010. His Ph.D. advisors were Prof. Gert Pfurtscheller at Graz University of Technology from 2008 to 2010 and Prof. Xingyu Wang at East China University of Science and Technology from 2006 to 2008. He is currently an assistant Professor at East China University of Science and Technology.

His research interests include brain-computer interface, signal processing and pattern recognition.

Xingyu Wang was born in Sichuan, China, in 1944. He received the B.S. degree in mathematics from Fudan University, Shanghai, China, in 1967, and the M.S. in control theory from East China Normal University, Shanghai, China, in 1982, and Ph.D. degrees in industrial automation from East China University of Science and Technology, Shanghai, China, in 1984. He is currently a Professor at the School of Information Science and Engineering, East China University of Science and Technology, Shanghai, China.

His research interests include control theory, control techniques, the application to biomedical system, and brain control.

Andrzej Cichocki received the M.Sc. (with Hons.), Ph.D., and Dr.Sc. (Habilitation) degrees, all in electrical engineering, from Warsaw University of Technology, Warsaw, Poland. Since 1972, he has been with the Institute of Theory of Electrical Engineering, Measurement and Information Systems, Faculty of Electrical Engineering at the Warsaw University of Technology, where he received the title of a Full Professor in 1995. He spent several years at the University Erlangen-Nuerenberg, Germany, at the Chair of Applied and Theoretical Electrical Engineering directed by Prof. R. Unbehauen, as an Alexander-von-Humboldt Research Fellow and Guest Professor. From 1995 to 1997, he was a team leader of the laboratory for Artificial Brain Systems, at the Frontier Research Program RIKEN, Japan, in the Brain Information Processing Group. He is currently the head of the laboratory for Advanced Brain Signal Processing, at RIKEN Brain Science Institute, Wako-shi, Japan. He is author of more than 250 technical papers and four monographs (two of which have been translated to Chinese).

His research interests include signal processing, inverse problems, neural network and learning algorithms, tensor analysis, and brain-computer interface.