Optimum Spatio-Spectral Filtering Network for Brain–Computer Interface

Haihong Zhang, *Member, IEEE*, Zheng Yang Chin, *Member, IEEE*, Kai Keng Ang, *Member, IEEE*,
Cuntai Guan, *Senior Member, IEEE*, and Chuanchu Wang, *Member, IEEE*

***Abstract*—This paper proposes a feature extraction method for motor imagery brain–computer interface (BCI) using electroencephalogram. We consider the primary neurophysiologic phenomenon of motor imagery, termed event-related desynchronization, and formulate the learning task for feature extraction as maximizing the mutual information between the spatio-spectral filtering parameters and the class labels. After introducing a nonparametric estimate of mutual information, a gradient-based learning algorithm is devised to efficiently optimize the spatial filters in conjunction with a band-pass filter. The proposed method is compared with two existing methods on real data: a BCI Competition IV dataset as well as our data collected from seven human subjects. The results indicate the superior performance of the method for motor imagery classification, as it produced higher classification accuracy with statistical significance ($\geq$95% confidence level) in most cases.**

***Index Terms*—Brain–computer interface, motor imagery electroencephalography, spatio-spectral filtering.**

I. Introduction

THE necessity of developing high-performance brain–computer interface (BCI) is rapidly growing alongside advances in neural devices and demands from rehabilitation, assistive technology, and beyond [1], [2]. Among the various useful signals for electroencephalogram (EEG) based BCI [3], *motor imagery* [4] is probably the most common one. It refers to the imagination or mental rehearsal of a motor action without any real motor output.

The primary phenomenon of motor imagery electroencephalography (EEG) is event-related desynchronization (ERD) [4], [5], which is the attenuation of the rhythmic activity over the sensorimotor cortex in the μ (8–14 Hz) and β (14–30 Hz) rhythms. ERD can be induced by both imagined movements in healthy people or intended movements in paralyzed patients [6]. Previous studies have demonstrated that, based on ERD analysis, it is feasible to classify imagined movements of left hand, right hand, feet, and tongue [4], [7], [8]. A complementary phenomenon called Bereitschafts potential is a nonoscillatory characteristic of motor imagery EEG, and can be also used for BCI [9]. This paper will focus on the ERD.

For decoding different motor imaginations from EEG, the essential task is to distinguish the respective ERD signals. Neurologically, the spatial pattern of the ERD provides a clue. For instance, movements of left hand/right hand are associated with activities in the contralateral (right/left) motor cortex areas [4].

However, localization of the ERD sources is impeded by the EEG's poor spatial specificity caused by volume conduction and coherency [10], [11]. Furthermore, the ERD is sensitive to artifacts cased by muscle activities or by visual cortex activities, since their frequency ranges highly overlap while the ERD signal is rather weak [12]. Besides, both the spatial pattern and the particular rhythm vary among people, requiring subject-specific learning [5].

Therefore, from a signal processing point of view, it is important to design a feature extraction mechanism that can learn to capture effective spatial and spectral features associated with the ERD, for each particular person. As a recent survey [13] indicates, considerable efforts have been devoted to this topic by the signal processing, machine learning, and artificial neural networks communities.

Particularly, spatial filtering techniques are widely used to extract discriminative spatial features of ERD in multichannel EEG. Techniques such as independent component analysis [14] and beam-forming [15] were introduced, while the most commonly used technique thus far is the common spatial pattern (CSP) [4], [16], [17]. As [18] shows, CSP can yield significantly higher accuracy in motor imagery classification than various independent component analysis methods.

CSP consists of a linear projection of time samples of multichannel EEG onto a few vectors that correspond to individual *spatial filters*. Mathematically, the projection matrix is constructed by maximizing the separability, in terms of the Rayleigh coefficient [17], between motor imagery EEG classes. The coefficient is determined by the intraclass covariance matrices of EEG time samples, while its maximization can be readily solved by generalized eigenvalue decomposition.

Usually, CSP works together with a subject-specific band-pass filter to select the particular rhythm of the ERD. To learn the band-pass filter and the spatial filters in a unified framework, several extensions of CSP have been devised. In [19], the authors embedded a first-order finite impulse

Manuscript received April 3, 2010; revised August 1, 2010 and September 2, 2010; accepted September 28, 2010. Date of publication November 9, 2010; date of current version January 4, 2011. This work was supported by the Science and Engineering Research Council of the Agency for Science, Technology and Research, Singapore.

The authors are with the Institute for Infocomm Research, Agency for Science, Technology and Research, 138632, Singapore (e-mail: hhzhang@i2r.a-star.edu.sg; zychin@i2r.a-star.edu.sg; kkang@i2r.a-star.edu.sg; ctguan@i2r.a-star.edu.sg; ccwangg@i2r.a-star.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2010.2084099

1045–9227/$26.00 © 2010 IEEE

response filter into CSP. In view of the limited capability of first-order filters to choose frequency bands, a higher order finite impulse response (FIR) filter was proposed in [20], while a sophisticated regularization method was necessary to make the solution robust. More recently, Wu *et al.* [21] proposed an iterative learning method, in which an FIR filter and a classifier were simultaneously parameterized and optimized in the spectral domain, alternately with optimization of spatial filters using CSP in the spectral domain.

More recently, another method called filter bank common spatial pattern (FBCSP) [22] introduced a feature selection algorithm to combine a filter bank framework with CSP. It decomposed EEG data into an array of passbands, performed CSP in each band, and selected a reduced set of features from all the bands. An offline study [23] suggested its higher performance over the above-mentioned iterative learning method. Furthermore, its efficacy was demonstrated in the latest BCI Competition [24], where it served as the basis of all the winning algorithms in the EEG categories. FBCSP was further improved in [25] by employing a robust maximum mutual information criterion for feature selection. (Another method [8] used the maximum mutual information principle but in a different formulation to select spatial components from independent component analysis).

However, learning optimum spatio-spectral filters is still an open issue. Extensions of CSP often inherit its limitation in exploring spatial patterns. Specifically, as shown in the Appendix and in [26, Sec. 10.2], CSP is equivalent to minimizing a classification error bound for two unimodal multivariate Gaussian distributions only. As [13, p. R43] puts it, it can also be sensitive to artifacts in the training data, as a single trial contaminated with artifacts can unfortunately cause extreme changes to the filters.

In this paper, we present an information-theoretic approach to learning the spatio-spectral filters. Particularly, the approach constructs an optimum spatio-spectral filtering network (OSSFN) that optimizes the filters by maximizing the mutual information between the feature vectors and the corresponding class labels. As mentioned earlier, the maximum mutual information criterion was employed in [25] for *feature selection*, where numerical optimization of spatial filters was not considered. By contrast, this paper addresses the more challenging and interesting issue of *feature extraction*, which involves numerical optimization of spatial filters together with selection of a band-pass filter.

Therefore, one of the major contributions of this paper is the introduction of a nonparametric mutual information estimate to formulate the objective for spatio-spectral feature extraction. Importantly, based on this new formulation, we devise a gradient-based method for optimization of spatial filters jointly with a band-pass filter.

We conduct an experimental study to assess the proposed method while comparing with existing methods including CSP and FBCSP. The study collects motor imagery data from seven human subjects in our lab. The publicly available BCI Competition IV Dataset I is also used. The study performs randomized cross-validation to assess the classification accuracy with a linear support vector machine, and runs t-test to verify the statistical significance of the results between different methods.

TABLE I

LIST OF SYMBOLS

Symbol	Description
$\mathbf{z}(t)$	A block of raw n_c-channel EEG signal; $t \in [0\ L]$
$\mathbf{x}(t)$	Signal after spectral filtering using a bandpass filter h
$\mathbf{y}(t)$	Signal after spatial filtering
$\mathbf{W}$	$\in \mathbb{R}^{(n_c \times n_l)}$, spatial filtering matrix with spatial filter vectors in columns
$\mathbf{w}_l$	$\in \mathbb{R}^{(n_c \times 1)}$, the lth spatial filter vector in $\mathbf{W}$
$\mathbf{a}$; $\mathcal{A}$	A particular feature vector $\in \mathbb{R}^{(n_l \times 1)}$ for $\mathbf{z}(t)$; feature vector variable symbol
ω; Ω	A particular class label; class label variable symbol
p; P	Probability density function and probability function of a random variable
H	Entropy of a random variable
$I(\mathcal{A}, \Omega)$	Mutual information between $\mathcal{A}$ and Ω
n_a	Number of samples ($\mathbf{z}(t)$) in training data
n_ω	Number of class-ω samples in training data

The rest of this paper is organized as follows. Section II describes the proposed method, and formulates the maximum mutual information based learning problem. Section III derives a numerical solution. Section IV describes the experimental study and the results, followed by discussions in Section V. Section VI finally concludes this paper.

II. OSSFN

For the convenience of readers, Table I describes a list of essential mathematical symbols.

The architecture of the proposed filtering network OSSFN is illustrated in Fig. 1. It learns and performs consecutive band-pass filtering, spatial filtering, and log power integral to extract discriminative features for motor imagery classification. The input of the network is a time window of n_c-channel EEG waveforms $\mathbf{z}(t)$ (without loss of generality, we assume $t \in [0\ L]$ in the time window), and the output is a feature vector $\mathbf{a}$ that represents the mean power of spatio-spectral components of $\mathbf{z}(t)$. The procedure of transforming the EEG *block* of $\mathbf{z}(t)$ into the feature vector $\mathbf{a}$ comprises the following steps.

1) *Spectral filtering*: A band-pass filter that extracts a specific rhythmic activity of the ERD, it produces the band-pass-filtered signal $\mathbf{x}$.

2) *Spatial filtering*: A linear projection of $\mathbf{z}$ that transforms $\mathbf{x}$ into a lower dimensional signal $\mathbf{y}$

$$\mathbf{y}(t) = \mathbf{W}^T \mathbf{x}(t). \tag{1}$$

Here, the superscript T denotes the transpose operator. Each column in the transformation matrix $\mathbf{W} \in \mathbb{R}^{(n_c \times n_l)}$ determines one of the n_l spatial filters. Therefore, each element in $\mathbf{y}$ describes the activity of a particular spatial component.

3) *Log power integral*: A process that computes ERD features as the mean power of $\mathbf{y}$ in the time window

$$\mathbf{a} = \log \left[\frac{1}{L} \int_0^L \left[\mathbf{y}(t)\right]^2 dt \right]. \tag{2}$$

Each of the element in $\mathbf{a}$ represents the mean band power of a particular spatial component in $\mathbf{W}$.

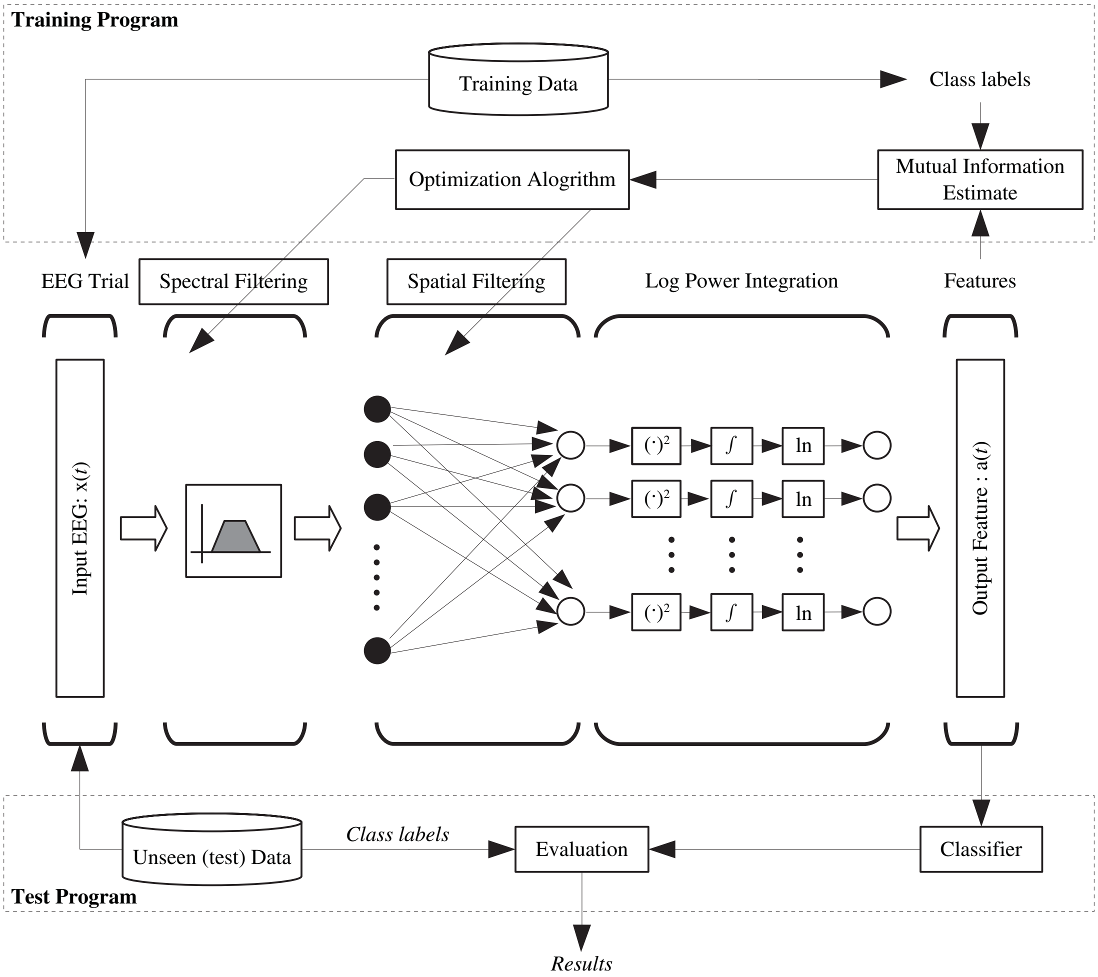

Fig. 1. Diagram of the proposed network for extracting motor imagery EEG features. A motor imagery EEG block, in the form of time-windowed multivariate waveforms $\mathbf{x}(t)$, is processed firstly by a spectral (band-pass) filter to pick up subject-specific responsive rhythm activity, and subsequently by a linear transformation (in the form of spatial filters) and log power integration. The output feature vector describes the mean power of particular spatio-spectral components associated with motor imagery. The network takes a maximum mutual information approach to optimizing the spectral filter and the spatial filters.

> The logarithm operation has been widely used since the introduction of CSP in [16], which describes its purpose as "to approximate normal distribution of the data." We would like to note that another positive effect of the logarithm operation is the reduced dynamic range, which facilitates the subsequent processing, e.g., by a classifier. In addition, extreme feature values (suspects of artifacts) in some EEG blocks can be largely reduced before the corrupted information (such as intraclass variance) is fed into the learning machine. Our BCI experience suggests that the logarithm operation can improve classification accuracy.

This paper introduces mutual information [27] to formulate the objective function for the learning machine. Consider the mutual information between the feature vector variable $\mathcal{A}$ and the class label variable Ω

$$\begin{aligned} I(\mathcal{A}, \Omega) &= H(\Omega) - H(\Omega|\mathcal{A}) \\ &= H(\mathcal{A}) - H(\mathcal{A}|\Omega) \\ &= H(\mathcal{A}) - \sum_{\omega \in \Omega} H(\mathcal{A}|\omega) P(\omega) \end{aligned} \tag{3}$$

where $H(\Omega)$ (or $H(\mathcal{A})$) is the entropy of the class label (or the feature vector). ω is a particular class label (e.g., $\omega = 1$ or $\omega = 2$ represents left- or right-hand motor imagination). $H(\mathcal{A}|\Omega)$ is the conditional entropy of the obtained feature vector for a particular class. $H(\Omega|\mathcal{A})$ is then the conditional entropy of the class label given the obtained feature vector.

Now we define the objective function for learning. Since the feature vector $\mathbf{a}$ is determined by the band-pass filter h and the spatial filters $\mathbf{W}$, the objective is to maximize $I(\mathcal{A}, \Omega)$ with respect to h and $\mathbf{W}$

$$\{h_{opt}, \mathbf{W}_{opt}\} = \max_{\{h, \mathbf{W}\}} I(\mathcal{A}, \Omega). \tag{4}$$

Let us discuss the relevance of mutual information to objective function for discriminative learning. The mutual information $I(\mathcal{A}, \Omega)$ is the reduction of uncertainty by the feature vector [27], the entropy $H(\Omega)$ is the uncertainty about class label, while after observing the feature vector, the uncertainty reduces to the conditional entropy $H(\Omega|\mathcal{A})$.

An earlier paper [28] has connected the maximum mutual information criterion to minimum Bayes error via lower and upper bounds. A recent paper [29] further studied the relationship between maximum mutual information and other criteria for feature extraction, though in the context of linear feature extraction rather than in the present nonlinear context (see the processing steps above). Importantly, that paper concludes that maximum mutual information is Bayesian optimum under more general conditions than others. Coincidently, recent years have seen attempts [30], [31] to address linear feature extraction problems through using the maximum mutual information principle.

III. Learning Algorithm

The technical challenge to achieve the objective in (4) primarily lies in the fact that the objective function (mutual information) is a function of probability density functionals and cannot be expressed in explicit form generally. To address this problem, we propose a learning method below that first introduces a mutual information estimation method and then derives a gradient-based optimization algorithm.

A. Mutual Information Estimate

Since the mutual information in (3) is dependent on the entropies, we approximate it by first estimating the entropies.

The entropy of the feature vector variable and the conditional entropy are, respectively, given by

$$H(\mathcal{A}) = -\int_{\mathbf{a}} p(\mathbf{a})\log\left(p(\mathbf{a})\right) d\mathbf{a} \tag{5}$$

and

$$H(\mathcal{A}|\omega) = -\int_{\mathbf{a}} p(\mathbf{a}|\omega)\log\left[p(\mathbf{a}|\omega)\right] d\mathbf{a}. \tag{6}$$

The entropy $\mathcal{A}$ can be viewed as an expectation of the function $\log(p(\mathbf{a}))$ [32, Sec. 5]. Suppose a set of n_a empirical samples of feature vector $\mathbf{a}$ is available: $\mathbf{a}_i$, $i = 1, \ldots, n_a$. The entropy can be estimated by

$$\begin{aligned} H(\mathcal{A}) &= -E[\log(p(\mathbf{a}))] \\ &\cong -\frac{1}{n_a}\sum_{i=1}^{n_a}\log(p(\mathbf{a}_i)). \end{aligned} \tag{7}$$

Similarly

$$\begin{aligned} H(\mathcal{A}|\omega) &= -E[\log(p(\mathbf{a}|\omega))] \\ &\cong -\frac{1}{n_\omega}\sum_{\mathbf{a}_i\in\omega}\log(p(\mathbf{a}_i)). \end{aligned} \tag{8}$$

The underlying probability density function can also be estimated from the samples using kernel density estimation [33]

$$\hat{p}(\mathbf{a}) = \frac{1}{n_a}\sum_{i=1}^{n_a}\varphi(\mathbf{a}-\mathbf{a}_i). \tag{9}$$

Using Gaussian for the kernel function φ, the Gaussian kernel density estimation is well known for its capability for general data analysis [33], [34]. A multivariate Gaussian function is given by

$$\varphi(\mathbf{r}) = (2\pi)^{-\frac{n_l}{2}}|\psi|^{-\frac{1}{2}}e^{\left(-\frac{1}{2}\mathbf{r}^T\psi^{-1}\mathbf{r}\right)} \tag{10}$$

where $\mathbf{r}$ denotes the term $\mathbf{a}-\mathbf{a}_i$, ψ usually takes a diagonal matrix form called the bandwidth matrix. The diagonal elements in the bandwidth matrix determine the smoothness of the kernel.

We choose the following bandwidth for the kernel:

$$\psi_{k,k} = \frac{\zeta}{n_a-1}\sum_{i=1}^{n_a}(a_{ik}-\bar{a}_k)^2 \tag{11}$$

where $\bar{a}_k$ is the empirical mean of $\{a_{ik}\}$, i.e., the kth element in the feature vector samples. We use the normal optimal smoothing strategy [34] to set the coefficient, i.e., $\zeta = (4/3n_a)^{0.1}$.

By introducing (9) into (7), the entropy $H(\mathcal{A})$ is approximated by

$$H(\mathcal{A}) \cong \hat{H}(\mathcal{A}) = -\frac{1}{n_a}\sum_{i=1}^{n_a}\log\left\{\frac{1}{n_a}\sum_{j=1}^{n_a}\varphi(\mathbf{a}_i-\mathbf{a}_j)\right\}. \tag{12}$$

The conditional intraclass entropy $\hat{H}(\mathcal{A}|\omega)$ is estimated similarly.

We replace the entropies in (3) by the estimates $\hat{H}(\mathcal{A})$ and $\hat{H}(\mathcal{A}|\omega)$. This results in a sample-based estimate of mutual information. The full expression of the estimate is omitted since it is straightforward from the above.

B. Subspace Gradient Descent Learning

In this section, we derive a numerical solution to maximizing the mutual information estimate with respect to spatial filters in $\mathbf{W}$ in conjunction with a band-pass filter.

For simultaneous optimization of all the spatial filter vectors in $\mathbf{W}$, we consider a joint vector by concatenating all the spatial filters

$$\hat{\mathbf{w}} = \left[\mathbf{w}_1^T \ \ldots \mathbf{w}_l^T \ \ldots \ \mathbf{w}_{n_l}^T\right]^T. \tag{13}$$

As described earlier, the mutual information $I(\mathcal{A},\Omega)$ is estimated from all the feature vector samples $\{\mathbf{a}_i\}$. Since each of the samples in turn is a function of $\hat{\mathbf{w}}$, we have

$$\boxed{\frac{\partial I(\mathcal{A},\Omega)}{\partial\hat{\mathbf{w}}} = \sum_{i=1}^{n_a}\frac{\partial I(\mathcal{A},\Omega)}{\partial\mathbf{a}_i}\frac{\partial\mathbf{a}_i}{\partial\hat{\mathbf{w}}}.} \tag{14}$$

The partial derivative $\partial I(\mathcal{A},\Omega)/\partial\mathbf{a}_i$ can be computed by differentiating (3) to give

$$\frac{\partial I(\mathcal{A},\Omega)}{\partial\mathbf{a}_i} = \frac{\partial H(\mathcal{A})}{\partial\mathbf{a}_i} - P(\omega)\frac{\partial H(\mathcal{A}|\omega)}{\partial\mathbf{a}_i} \tag{15}$$

where ω is the class label of the sample $\mathbf{a}_i$.

To compute $\partial I/\partial\mathbf{a}_i$, the partial derivatives $\partial H(\mathcal{A})/\partial\mathbf{a}_i$ and $\partial H(\mathcal{A}|\omega)/\partial\mathbf{a}_i$ are required. To compute $\partial H(\mathcal{A})/\partial\mathbf{a}_i$, differentiate (12) with respect to $\mathbf{a}_i$, which gives

$$\frac{\partial H(\mathcal{A})}{\partial\mathbf{a}_i} = -\frac{1}{n_a}\sum_{j=1}^{n_a}\beta_j\frac{1}{n_a}\sum_{k=1}^{n_a}\frac{\partial\varphi[\mathbf{a}_j-\mathbf{a}_k]}{\partial\mathbf{a}_i} \tag{16}$$

where

$$\beta_j = \left\{\frac{1}{n_a}\sum_{k=1}^{n_a}\varphi[\mathbf{a}_j-\mathbf{a}_k]\right\}^{-1} \tag{17}$$

and

$$\frac{\partial \varphi(\mathbf{a}_j - \mathbf{a}_k)}{\partial \mathbf{a}_i} = \begin{cases} -\varphi(\mathbf{a}_i - \mathbf{a}_k)\psi^{-1}(\mathbf{a}_i - \mathbf{a}_k), & \text{if } i = j \\ -\varphi(\mathbf{a}_i - \mathbf{a}_j)\psi^{-1}(\mathbf{a}_i - \mathbf{a}_j), & \text{if } i = k \\ 0, & \text{otherwise} \end{cases} \tag{18}$$

the computation of the partial derivative $\partial H(\mathcal{A}|\omega)/\partial \mathbf{a}_i$ is performed similarly.

To compute the partial derivative $\partial \mathbf{a}_i / \partial \hat{\mathbf{w}}$, we first consider a particular element, say, the lth element a_{il} in $\mathbf{a}_i$. From (13), the partial derivative of this element with respect to $\mathbf{w}_l$ is

$$\begin{aligned} \frac{\partial a_{il}}{\partial \mathbf{w}_l} &= \frac{\partial \log\left[\frac{1}{L}\int_0^L \left[\mathbf{w}_l^T \mathbf{x}_i(t)\right]^2 dt\right]}{\partial \mathbf{w}_l} \\ &= \frac{1}{e^{a_{il}}} \cdot \frac{\partial \left\{\mathbf{w}_l^T \left[\frac{1}{L}\int_0^L \mathbf{x}_i(t)\mathbf{x}_i^T(t)dt\right]\mathbf{w}_l\right\}}{\partial \mathbf{w}_l} \\ &= \frac{2}{Le^{a_{il}}}\mathbf{w}_l^T R_{xi} \end{aligned} \tag{19}$$

where $\mathbf{x}_i(t)$ denotes the EEG sequence in the ith trial, and

$$R_{xi} = \int_0^L \mathbf{x}_i(t)\mathbf{x}_i^T(t)dt. \tag{20}$$

Since a_{ij} is dependent on $\mathbf{w}_l$ only

$$\frac{\partial a_{ij}}{\partial \mathbf{w}_l} = 0 \text{ if } j \neq l \tag{21}$$

the partial derivative of $\mathbf{a}$ with respect to $\hat{\mathbf{w}}$ is thus

$$\frac{\partial \mathbf{a}_i}{\partial \hat{\mathbf{w}}} = \begin{bmatrix} \frac{2}{Le^{a_{i1}}}\mathbf{w}_1^T R_{xi} & 0 & \cdots & 0 \\ 0 & \frac{2}{Le^{a_{i2}}}\mathbf{w}_2^T R_{xi} & 0 & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2}{Le^{a_{in_l}}}\mathbf{w}_{n_l}^T R_{xi} \end{bmatrix}. \tag{22}$$

Now we can compute the gradient by introducing the above equation to (14).

However, a practical issue arises for multichannel EEG and multiple spatial filters. Consider an example in which EEG has $n_c = 59$ channels and $\mathbf{W}$ contains $n_l = 2$ filters. The number of free parameters would be $2 \times 59 = 118$. Gradient-based optimization in this high-dimensional space would be difficult.

To address this issue, we propose a subspace optimization approach in below.

Consider a n_u-dimensional ($n_u \ll n_c$) subspace $\mathbb{U}$, linearly spanned by the n_c-dimensional column vectors as bases in a matrix $\mathbf{U} \in \mathbb{R}^{(n_c \times n_u)}$

$$\mathbf{U} = \left[\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_k, \ldots, \mathbf{u}_{n_u}\right] \tag{23}$$

where $\mathbf{u}_k$ denotes the kth basis vector.

A spatial filter vector $\mathbf{w}_l$ in the subspace can be expressed by

$$\mathbf{w}_l = \sum_{k=1}^{n_u} b_{lk}\mathbf{u}_k = \mathbf{U}\mathbf{b}_l \tag{24}$$

where $\mathbf{b}_l$ is a coefficient vector that determines $\mathbf{w}_l$

$$\mathbf{b}_l = \left[b_1, b_2, \ldots, b_{n_u}\right]. \tag{25}$$

Hence, $\mathbf{b}_l$ is the low-dimensional representation of the spatial filter $\mathbf{w}_l$.

In the subspace $\mathbb{U}$, simultaneous optimization of the spatial filters is equivalent to simultaneous optimization of the concatenated coefficient vectors

$$\hat{\mathbf{b}} = \left[\mathbf{b}_1^T \; \mathbf{b}_2^T \ldots \mathbf{b}_{n_l}^T\right]^T. \tag{26}$$

Now consider the partial derivatives of $I(\mathcal{A}, \Omega)$ with respect to $\hat{\mathbf{b}}$

$$\frac{\partial I(\mathcal{A}, \Omega)}{\partial \hat{\mathbf{b}}} = \sum_{i=1}^{n_a} \frac{\partial I}{\partial \mathbf{a}_i}\frac{\partial \mathbf{a}}{\partial \hat{\mathbf{b}}}. \tag{27}$$

Substitution of (24) into (2) gives

$$a_{il} = \log\left[\frac{1}{L}\int_0^L \left[(\mathbf{U}\mathbf{b}_l)^T \mathbf{x}_i(t)\right]^2 dt\right]. \tag{28}$$

Similar to (19), differentiating (28) gives

$$\begin{aligned} \frac{\partial a_{il}}{\partial \mathbf{b}_l} &= \frac{\partial \log\left[\frac{1}{L}\int_0^L \left[(\mathbf{U}\mathbf{b}_l)^T \mathbf{x}_i(t)\right]^2 dt\right]}{\partial \mathbf{b}_l} \\ &= \frac{1}{e^{a_{kl}}} \cdot \frac{\partial \left\{(\mathbf{U}\mathbf{b}_l)^T \left[\frac{1}{L}\int_0^L \mathbf{x}_i(t)\mathbf{x}_i(t)^T dt\right](\mathbf{U}\mathbf{b}_l)\right\}}{\partial b_{kl}} \\ &= \frac{2}{Le^{a_{kl}}}(\mathbf{U}\mathbf{b}_l)^T R_{xi}\mathbf{U}. \end{aligned} \tag{29}$$

Therefore

$$\frac{\partial \mathbf{a}_i}{\partial \hat{\mathbf{b}}} = \begin{bmatrix} \frac{2(\mathbf{U}\mathbf{b}_1)^T}{Le^{a_{i1}}} R_{xi}\mathbf{U} & \cdots & 0 \\ 0 & \frac{2(\mathbf{U}\mathbf{b}_2)^T}{Le^{a_{i2}}} R_{xi}\mathbf{U} & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{2\left(\mathbf{U}\mathbf{b}_{n_l}\right)^T}{Le^{a_{in_u}}} R_{xi}\mathbf{U} \end{bmatrix}. \tag{30}$$

Now, introducing the above equation to $\partial I(\mathcal{A}, \Omega)/\partial \hat{\mathbf{b}}$ (expressed in a similar form as (14) by substituting $\hat{\mathbf{b}}$ for $\hat{\mathbf{w}}$), we can compute the gradient of the mutual information estimate with respect to $\hat{\mathbf{b}}$ of low dimensionality. This effectively reduces the number of free parameters for learning. In the earlier example, the number of free parameters will reduce from 118 to 8 in a $n_u = 4$-D subspace.

How to optimally construct the subspace $\mathbb{U}$ is, however, beyond the scope of this paper. Tentatively, we simply use spatial filters by CSP (band-pass filter selected by FBCSP) as the subspace bases. We would like to stress that the proposed optimization procedure, as a general approach, is neither tailored nor dedicated to the CSP or FBCSP subspace. We expect that more effective subspace construction methods will be devised.

As mentioned earlier, the subject-specific sensorimotor rhythm of the ERD must be selected for effective extraction of spatial patterns associated with the ERD. To this end, we need to maximize the mutual information estimate with respect to spatial filters in conjunction with a band-pass filter. Inspired by previous works [22], [35] that choose the optimum band-pass filter from an array of filter banks, we propose a joint spatio-spectral filter learning algorithm below (Fig. 2) in a filter bank framework. Briefly, the algorithm first decomposes the EEG data into an array of frequency bands that cover the range of possible ERD rhythm, performs spatial filter optimization in each band, and then selects the band with maximum mutual information estimate.

IV. Experiments and Results

This section reports an offline analysis of the proposed method for extracting the ERD features.

A. Materials: Motor Imagery EEG Datasets

1) **BCI Competition IV Dataset I**: The dataset [24] consists of both human and artificially generated motor imagery data. We consider human EEG data only, which were collected from four healthy subjects using the EEG amplifier of BrainAmp MR plus with 59 channels sampled at 1000 Hz. Each subject participated in two data collection sessions with different protocols as described below.

   In the calibration session, a visual cue was displayed on a computer screen to the subjects who then started to perform motor imagery tasks according to the cue. The cue represented specific motor imagery tasks: each subject chose two classes of motor imagery tasks from left hand, right hand, or foot. Specifically, subject “a” chose {*left*, *foot*}, “b” chose {*left*, *right*}, “f” chose {*left*, *foot*}, “g” chose {*left*, *right*}. Each subject performed a total of 200 motor imagery tasks (balanced between the two tasks) each in the [0 4]-s window after the cue. Consecutive motor imagery tasks were interleaved with a 4-s break.

   In the evaluation session, the subjects followed the soft voice commands from an instructor to perform motor imagery tasks of varying length between 1.5 and 8 s. Consecutive tasks were also interleaved with a varying length interval from 1.5 to 8 s. This session was meant for offline validation of motor imagery classification algorithms for self-paced BCI (see [36]).

   Our study uses the down-sampled data (provided by the organizer) at 10-Hz sampling rate, with all the 59 channels employed for spatio-spectral feature extraction. The 59 channels are AF3, AF4, F5, F3, F1, Fz, F2, F4, F6, FC5, FC3, FC1, FCz, FC2, FC4, FC6, CFC7, CFC5, CFC3, CFC1, CFC2, CFC4, CFC6, CFC8, T7, C5, C3, C1, Cz, C2, C4, C6, T8, CCP7, CCP5, CCP3, CCP1, CCP2, CCP4, CCP6, CCP8, CP5, CP3, CP1, CPz, CP2, CP4, CP6, P5, P3, P1, Pz, P2, P4, P6, PO1, PO2, O1, and O2.

2) **Our Motor Imagery Data Set**: The data were recorded in our laboratory from seven healthy male subjects. Each subject performed 160 tasks of motor imagery (including 80 left-hand and 80 right-hand tasks). Similar to the calibration session of the BCI Competition dataset, the data collection procedure used visual cues to prompt the subjects to perform motor imagery tasks for 4 s each. Consecutive motor imagery tasks were interleaved with a 6-s break. The EEG data were recorded using a NuAmps amplifier with 25 channels sampled at 250 Hz. The 25 channels, including F7, F3, Fz, F4, F8, FT7, FC3, FCz, FC4, FT8, T7, C3, Cz, C4, T8, TP7, CP3, CPz, CP4, TP8, P7, P3, Pz, P4, and P8, cover the full scalp. The data collection and study was approved by the National University of Singapore Institutional Review Board with reference code 08-036.

Given the considerable difference in data collection setup in terms of EEG amplifiers and motor imagery task protocols, effective unification of the two datasets is difficult. Instead, this paper validates the proposed method on the two datasets separately. Furthermore, this allows validation of the proposed method in two different conditions (or effectively three conditions, since the calibration session and the evaluation session in the BCI Competition data were of different protocols), which is an important consideration for studying generalization performance.

B. Selection of Hyperparameters

The following describes how we set the hyperparameters for feature extraction and classification.

First, selection of a time interval in the motor imagery tasks is almost a common practice in learning motor imagery EEG. This paper selects the time interval [1 4] s after the cue. The first 1-s period after the cue is excluded since it contains the spontaneous responses (evoked potentials) to the cue stimulus [37, Sec. V]. As the BCI Competition evaluation set has varying duration of motor imagery tasks, we consider the same time interval and remove those motor imagery tasks of less than 4 s long. Consequently, the number of remaining motor imagery tasks in the evaluation data ranges from 111 to 126 in the four subjects.

Second, the filter banks (an array of band-pass filters) are constructed to continuously cover a wide frequency range. Specifically, a total of eight Chebyshev Type II filters (though other type of filters can also be used instead) are built with center frequencies spanning from 8 to 32 Hz at a constant interval in the logarithm domain. Consequently, the center frequencies are respectively 8, 9.75, 11.89, 14.49, 17.67, 21.53, 26.25, and 32 Hz. All the filters all have a uniform Q-factor (bandwidth-to-center frequency) of 0.33 as well as an order of 4. The filter banks process each of the EEG blocks separately after they are extracted from the selected time interval mentioned above.

The number of spatial filters to be constructed is also an important hyperparameter. This paper considers the learning of two spatial filters only, corresponding to a transformation matrix $\mathbf{W}$ (1) of two column vectors. Consequently, the feature vector is a bivariate.

C. Mutual Information Surface and Selected Spatial Patterns

Here we use the calibration data from the BCI Competition dataset to investigate the surface of the mutual information versus spatial filters. We visualize the mutual information estimate in a low-dimensional space $\mathbb{U}$ (see Section III-B), in which each point defines a particular spatial filter. As more

```
Input:Training EEG data that comprises $N$ sample blocks of $\{z(t)\}$, each block has a specific class label;
Output:The filtering network as depicted in Fig.1, with optimum parameters for spatial filters and the selection of the optimum band-pass filter;
Step 1: Construct an array of $n_s$ band-pass filters that covers the EEG rhythms of motor imagery, then filter $\{z(t)\}$ to yield $\{x_m(t)\}$ for $m = 1, \ldots, n_s$;
Step 2: For each band-pass filter's output $\{x_m(t)\}$:
  1) Construct a discriminative spatial filter subspace:
      a) Compute the empirical covariance matrices of the two classes: $\Psi_{x0}$ and $\Psi_{x1}$;
      b) Compute the eigenvectors and eigenvalues of $\Psi_{x0}^{-1}\Psi_{x0}$ (refer to equation(37));
      c) Select $n_u$ eigenvectors that correspond to the largest and smallest eigenvalues $\lambda$, sort the eigenvectors from large to small eigenvalues, use these eigenvectors as the bases U for the low dimensional subspace for parameterization of spatial filters;
  2) Set the initial parameters of the spatial filters $b_1^0 = [1, 0, \ldots, 0]$, $b_2^0 = [0,\ldots, 0, 1]$, $b_3^0 = [0, 1, 0, \ldots, 0]$, and so on. This setting effectively chooses the top and bottom spatial filters generated by CSP or FBCSP.
  3) Set iteration count $k = 0$, repeat the following steps until convergence whereby the criterion is defined as the change of the mutual information estimate being smaller than a small threshold $\zeta$:
      a) Compute the spatial filters W using $b^{Iter}$ inequation(24): $W = U\hat{b}^k$ where $\hat{b}^k = [b_1^k, b_2^k, \ldots, b_{n_l}^k]$,
      b) Use W to update the feature vector according to equations (1) and (2);
      c) Compute the gradient $\Delta\hat{b} = \frac{\partial I}{\partial \hat{b}}$ using equation (27);
      d) Perform a linear search with a stepfactor s, alternately selected from the range [−1 1] with an interval of 0.01:
          i) Set
                 $$\hat{b}(s) = \hat{b}^k + s\frac{\Delta\hat{b}}{\|\Delta\hat{b}\|_2^{\frac{1}{2}}} \quad (31)$$
             where $\|.\|_2$ represents the $l-2$ norm of the gradient vector.
          ii) Compute the mutual information estimate $I(s)$ with the spatial filters defined by $\hat{b}(s)$;
      e) Update the parameter vectors for spatial filters using the optimum update step $s_{opt} = \mathrm{argmax}_s I(s)$
      f) Update the mutual information $I^k = I(s_{opt})$ and $b^k = \hat{\hat{b}}(s_{opt})$;
      g) Compute the change in mutual information by $\delta = I^k - I^{k-1}$ ($I^{k-1} = I(s = 0)$ifunassigned); if $\delta < \zeta$ or the iteration count $k$ is larger than a preset number, continue to next step; otherwise go back to step a);
      h) Set the optimum spatial filters for the frequency bandas $W_m = U\hat{b}^k$ , and set the corresponding mutual information $I_m = I^k$;
Step 3:Select the optimum frequency band $m_{opt}$by $m_{opt} = \mathrm{argmin}_m I_m$, and finally set the spectral filter to be the $m_{opt}$ band-pass filter, and the spatial filter to be W$m_{opt}$.
```

Fig. 2. Learning algorithm for the spatio-spectral filtering network.

than 2-D surfaces would be difficult to visualize, we consider a 2-D subspace spanned by the first pair (i.e., the top and the bottom) of CSP filters from the frequency band selected by FBCSP [22].

Fig. 3 uses color image presentations to illustrate the result for each of the four subjects in the BCI Competition data. Each pixel represents a spatial filter, while the value of the corresponding mutual information estimate is denoted by the color. In three of the subjects, including "a," "b," and "g," a peak mutual information estimate appears near the point [0 1], which represents the bottom spatial filter from FBCSP. However, in "f," there is no such peak found near [0 1]. Instead, a peak is prominent near the point [1 0], which corresponds to the top filter from FBCSP. Hence, we use the top filter for the FBCSP mark in "f," while using the bottom one in the others.

The result suggests favorable conditions for the proposed method. First, the surface is smooth that facilitates gradient-based optimization. Second, target peaks on the mutual information surface often have an FBCSP filter in the vicinity, which validates the use of FBCSP spatial filters for optimization initialization.

Furthermore, Fig. 4 shows the top three spatial patterns (each with a particular frequency band) which together maximize the mutual information measure for each of the subjects in the competition dataset. The patterns are consistent with neurophysiological principles on motor imagery except for subject "b." For example, the spatial patterns for subject "g" show that the two most discriminative patterns correspond to EEG sources that originate from the motor cortex of the right and left hemisphere. Furthermore, the frequency bands of the selected spatial patterns are mostly from the Beta rhythm except for the second spatial pattern of subject "a."

D. Classification Results

This paper compares the proposed method in comparison with CSP and FBCSP, using five rounds of fivefold cross-validation study. FBCSP shares the same band-pass filter array as described in Section IV-B. Literally, it is the proposed method before optimization. Thus, it selects only one frequency band and two spatial filters. CSP is implemented by following [16] and using a [8 30]-Hz Chebyshev Type II band-pass filter. The top and the bottom filters by CSP are selected.

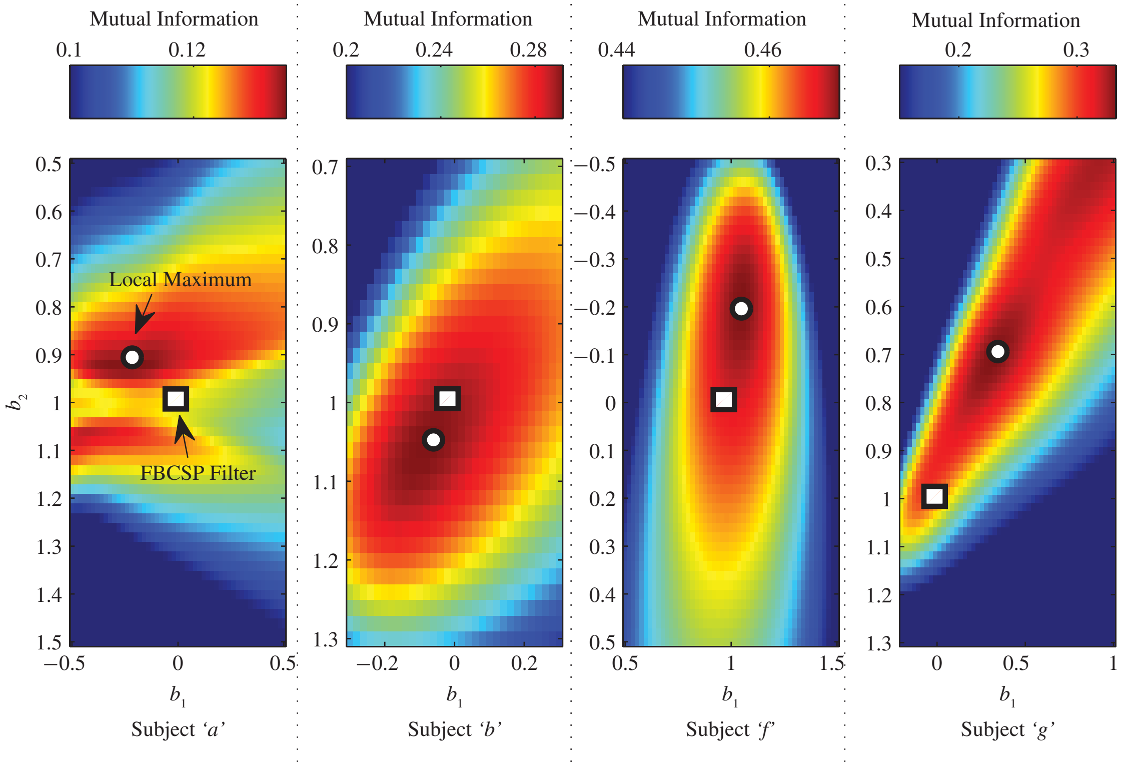

Fig. 3. Surface of mutual information estimate over a bivariate **b** (see 24), i.e., the coefficient vector that defines a spatial filter. Notes: see Section IV-C for details; the four graphs are for each of the four subjects in the BCI Competition IV Dataset I. The axes b_1 and b_2 denote the first and the second elements of the coefficient vector **b** in (24) that defines a spatial filter. The value of mutual information estimate is indicated by the color according to the overhead color bar. See Section. III-B for the description of parameterization of spatial filters in the subspace of spatial filters. Here the subspace is spanned by two spatial filters by FBCSP. Therefore, e.g., the point [1 0] represents to the first spatial filter by FBCSP. The previous FBCSP filter and the local optimum filter are annotated, respectively, by a square and a circle in each graph.

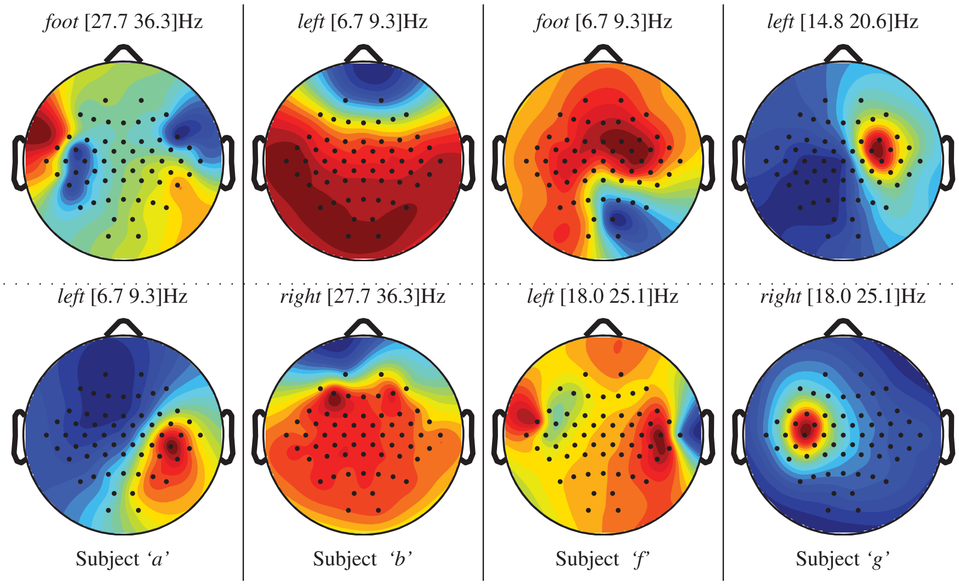

Fig. 4. Spatial patterns of motor imagery EEG. In each column, the two spatial features selected according to maximal mutual information between classes, are plotted in the form of spatial patterns (see [16]). The positions of the electrodes are superimposed as black dots. The view of the spatial pattern is from the top whereby the nose is facing upward. The motor imagery class and the frequency band of the feature are also given above each plot.

TABLE II

CLASSIFICATION ACCURACY (MEAN AND SD) IN BCI COMPETITION IV DATASET I

Sub	Train–Test	Feature extraction method			p-value	
		CSP	FBCSP	OSSFN	OSSFN=CSP	OSSFN=FBCSP
a	Calib.	67.1(11.4)	66.5(13.3)	**89.8(9.0)**	<**0.01**	<**0.01**
	Calib.–Eval.	65.9(4.7)	79.2(7.2)	**93.6(3.4)**	<**0.01**	<**0.01**
b	Calib.	76.6(10.6)	**87.3(5.5)**	86.8(5.6)	<**0.01**	0.31
	Calib.–Eval.	66.2(8.3)	90.5(0.9)	**90.8(0.9)**	<**0.01**	0.07
f	Calib.	65.3(10.7)	87.2(14.4)	**93.0(4.7)**	<**0.01**	**0.05**
	Calib.–Eval.	60.8(5.3)	92.4(7.6)	**96.0(0.8)**	<**0.01**	**0.03**
g	Calib.	74.6(6.9)	82.5(9.8)	**92.4(3.9)**	<**0.01**	<**0.01**
	Calib.–Eval.	68.2(1.3)	84.4(7.0)	**95.3(0.9)**	<**0.01**	<**0.01**

Notes: the first column denotes the subjects; the second column denotes two types of cross-validation study (see Section IV-D): “Calib.” stands for five rounds of fivefold cross-validation test in the calibration session of the data; “Calib.–Eval.” stands for the test that uses the evaluation session to assess the generalization performance of the models built in “Calib.” The two columns to the right side summarize the statistical significance test (paired t-test) results, and the p-values represent the probabilities of null hypotheses: the proposed method produces the same mean accuracy as CSP (OSSFN = CSP), and it produces the same mean as FBCSP (OSSFN = FBCSP). The p-value is in **BOLD** style if ≤ 0.05, meaning that the null hypotheses is rejected at >95% confidence level.

TABLE III

CLASSIFICATION ACCURACY (MEAN AND SD) IN OUR DATASET

Sub	Feature extraction methods			p-Value	
	CSP	FBCSP	OSSFN	OSSFN = CSP	OSSFN = FBCSP
1	82.0(9.8)	86.7(5.2)	**87.7(5.5)**	<**0.01**	0.25
2	84.7(10.9)	88.4(5.6)	**90.7(5.1)**	**0.01**	<**0.01**
3	70.6(10.9)	79.1(7.0)	**80.0(7.2)**	<**0.01**	0.57
4	87.3(5.4)	87.6(4.9)	**89.8(5.5)**	**0.02**	**0.05**
5	64.0(11.5)	**68.2(6.8)**	67.8(8.2)	0.20	0.85
6	63.1(13.2)	66.2(7.4)	**71.1(7.2)**	<**0.01**	<**0.01**
7	78.1(7.5)	86.0(5.7)	**89.3(7.2)**	<**0.01**	<**0.01**

Note: Refer to the notes under Table II or Section IV-D for explanation.

The cross-validation technique assesses how the results generated by the methods will generalize to an independent dataset. Each round of the fivefold cross-validation involves partitioning a sample of data into five subsets, alternately performing the learning on one subset (called the training set) and validating the learned model on the others (aggregated as the test set). Five rounds of cross-validation are performed using different partitions of data in order to reduce variability. The partitions are randomly generated using the cross-validation function “crossvalind” in the MATLAB Bioinformatics toolbox.

The cross-validation study creates a total of 25 pairs of training and testing tasks. Depending on the size of the total data for each subject, the number of EEG blocks is 160 (or 128) in each training set and 40 (or 32) in each test set for the BCI Competition data (or our data). To ensure a valid comparison between different methods, they all use the same data partitions in cross-validation.

Particularly on the BCI Competition dataset, a special cross-validation is performed to evaluate the method’s generalization performance. It begins with the cross-validation on the calibration set. Then the trained models are applied to both the cross-validation test set (part of the calibration set) and the whole evaluation set.

A classifier is used to assess the performance of the feature extraction methods for classifying motor imagery classes. The classifier learns and predicts the class labels from the features generated by CSP, FBCSP, and the proposed method, respectively. The accuracy rate is taken as the performance measure. For the classifier, we consider the linear support vector machine (SVM), since it is widely used in the field. Particularly, we use the default implementation in the library of support vector machine toolbox [38] (tuning of the regularization parameter in SVM is not performed according to our experience).

Tables II and III summarize the result. In a total of 15 cases (each case is a particular subject session), the proposed method significantly outperformed (at 95% confidence level) CSP in 14 cases or FBCSP in 10 cases. Compared to FBCSP, the largest boost in classification accuracy was in subject “a,” with the mean accuracy rate increased from 66.5% to 89.9% in the calibration test and from 79.2% to 93.6% in the calibration–evaluation test.

V. Discussion

The experimental results have demonstrated the efficacy of the proposed approach. Compared with the state-of-the-art FBCSP, the proposed method produced higher classification accuracy, in 10 cases with statistical significance, in 3 cases without statistical significance. Furthermore, it did not deteriorate classification accuracy with statistical significance in the remaining two cases.

Moreover, for the proposed method the classification accuracy in the calibration session and that in the calibration–evaluation test (calibration models applied to evaluation data) are similar. For example, the method yielded a mean classification accuracy of 89.9% for subject "a" in the cross-validation in the calibration data. The sample models applied to the evaluation set yielded a mean accuracy of 93.6%, which was even slightly higher. The slightly better performance in the evaluation set is interesting, and calls for future studies.

CSP was initially designed for two-class paradigms only [8]. Extensions to multiclass paradigms have been suggested, but are based on heuristics. In comparison, the mutual information formulation for learning naturally deals with multiclass problems. Hence, further work may look into the use of the present method for multiclass motor imagery classification. However, we need to note that it will be a challenge to construct effective subspace for low-dimensional representation of spatial filters in the multiclass contexts.

Besides, there is a connection between the current optimization procedure and FBCSP. The spatial filter learning algorithm runs in a low-dimensional representation subspace, instead of original space for multichannel EEG, in order to ease the optimization problem. In the subspace, any spatial filter can be expressed as a combination of the subspace bases, and tentatively we use FBCSP to form the bases. Nevertheless, the optimization procedure, as a general approach, is neither tailored nor dedicated to FBCSP-created subspaces. This means that one may devise more effective subspace construction methods and run the optimization procedure there, and expect improved performance.

There is a growing awareness of the importance of self-paced BCI that allows the user to operate BCI at any time at will, thus providing more natural and potentially faster interactions [39]. This paper, on the other hand, used a cue-based classification scheme to examine the efficacy of the optimum spatio-spectral filtering method. Nevertheless, subject-specific motor imagery models, learned through cue-based calibration, are still necessary in initialization of self-paced BCI systems [36], [40], [41] for each user. Thus, it is interesting to investigate the proposed method for self-paced BCI in future studies.

VI. Conclusion

In this paper, we have considered extracting spatio-spectral features of the ERD for motor imagery classification. We formulated the learning of optimum spatio-spectral filters as a maximum mutual information problem. We proposed a gradient-based optimization approach to solve the problem. To make the solution robust and efficient, we developed a subspace spatial filtering learning approach in which spatial filters were parameterized by lower dimensional vectors. The experimental results attest to the efficacy of the proposed method. Compared to CSP and FBCSP, the method produced significantly higher classification accuracy in most cases and it did not deteriorate classification accuracy with statistical significance in the rest few cases. We expect that more effective subspace construction methods will be devised to further improve the performance and extend the methods for multiclass motor imagery classification.

VII. Acknowledgment

The authors would like to thank the organizers of the BCI Competition IV [24] and the providers of the BCI Competition Dataset I [41].

Appendix A
Relationship of OSSFN with CSP and FBCSP

This section briefly reviews the CSP and the FBCSP algorithms and then discusses their relations to the proposed OSSFN.

1) CSP: CSP computes features whose variances are optimum for discriminating two classes of EEG measurements [16]. The method is based on simultaneous diagonalization of two covariance matrices of each class. In summary, the spatially filtered signal $\mathbf{y}(t)$ of a single-trial EEG $\mathbf{x}(t)$ is given in (1), where $\mathbf{W}^T$ is the CSP projection matrix. The spatial filtered signal $\mathbf{y}(t)$ maximizes the differences in the variance of the two classes of EEG by solving the eigenvalue decomposition problem [17]

$$\Psi_{x0}\mathbf{W} = \Lambda\Psi_{x1}\mathbf{W} \tag{32}$$

where Ψ_{x0} (or Ψ_{x1}) is the covariance matrix of the band-pass filtered EEG of motor imagery class 0 (or 1), and Λ is the diagonal matrix that contains the eigenvalues of $\mathbf{W}$.

Refer to [26, Sec. 10.2]. From a pattern classification perspective, CSP is equivalent to optimum linear transformation which minimizes the *Bhattacharyya bound* for two zero-mean unimodal Gaussian classes of EEG time samples. Below is a brief explanation.

The *Bhattacharyya bound* refers to an upper bound of Bayesian classification error given as

$$\epsilon_B(\mathbf{x}) = \sqrt{P(\omega_0)P(\omega_1)}\int\sqrt{p(\mathbf{x}|\omega_0)p(\mathbf{x}|\omega_1)}d\mathbf{x} \tag{33}$$

where $p(\mathbf{x}|\omega_0)$ and $p(\mathbf{x}|\omega_1)$ are the conditional probability density functions for the multichannel EEG $\mathbf{x}$ of the two classes ω_0 and ω_1.

After transforming $\mathbf{x}$ linearly into $\mathbf{y}$ using $\mathbf{W}$, the Bhattacharyya bound becomes

$$\epsilon_B(\mathbf{y}) = \sqrt{P(\omega_0)P(\omega_1)}\int\sqrt{p(\mathbf{y}|\omega_0)p(\mathbf{y}|\omega_1)}d\mathbf{y}. \tag{34}$$

If the conditional density functions are both zero-mean Gaussians with covariance matrix Ψ_{x0} and Ψ_{x1}, maximizing the Bhattacharyya bound is equivalent to solving the dual problem [26, Sec. 10.2, eqs. (10.42) and (10.43)]

$$\left(\Psi_{x_1}^{-1}\Psi_{x_0}\right)\mathbf{W} = \mathbf{W}\left(\Psi_{y_1}^{-1}\Psi_{y_0}\right) \tag{35}$$

$$\left(\Psi_{x_0}^{-1}\Psi_{x_1}\right)\mathbf{W} = \mathbf{W}\left(\Psi_{y_0}^{-1}\Psi_{y_1}\right) \tag{36}$$

where $\Psi_{y_0} = \mathbf{W}^T\Psi_{x_0}\mathbf{W}$ and $\Psi_{y_1} = \mathbf{W}^T\Psi_{x_1}\mathbf{W}$ are the covariance matrix of $\mathbf{y}$ in class ω_0 and ω_1. Note that in the implementation of [16], the covariance matrix was computed as the mean of trial-based covariance matrices after normalization (dividing each of the covariance matrices by its trace).

A solution to the above two equations exists for $\mathbf{W}$ being the eigenvectors of both $\Psi_{x_1}^{-1}\Psi_{x_0}$ and $\Psi_{x_0}^{-1}\Psi_{x_1}$. Since these two matrices are related by $\Psi_{x_1}^{-1}\Psi_{x_0} = (\Psi_{x_0}^{-1}\Psi_{x_1})^{-1}$, they share the same eigenvector matrix and the solution is given by

$$\Psi_{\mathbf{x0}}\mathbf{W} = \Lambda\Psi_{x1}\mathbf{W} \tag{37}$$

where Λ is the diagonal matrix of eigenvalues. The eigenvectors thus obtained are exactly the same as the ones obtained using CSP. Besides, as [26] puts it, the minimization of the Bhattacharyya bound is associated with the maximization of the following measure

$$\mathbf{J} = \sum_{i=1}^{n} \log\left(\lambda_i + \frac{1}{\lambda_i} + 2\right). \tag{38}$$

Therefore, in order to minimize the Bhattacharyya bound, the eigenvectors corresponding to the largest $(\lambda_i + \frac{1}{\lambda_i} + 2)$ terms are to be selected.

2) FBCSP algorithm: The FBCSP [22] processes input EEG with an array of band pass filters, and applies CSP to the EEG data after each band-pass filtering. It then concatenates the extracted CSP features from each filter band to form a joint feature vector. The FBCSP algorithm then selects from the joint feature vector a discriminative set of features, by employing a mutual information-based feature selection algorithm.

Therefore, we can summarize the relationship between OSSFN with CSP and FBCSP as below.

1) Spatial filtering: CSP and FBCSP produce optimum linear transformation for EEG *samples* of unimodal Gaussians, on the other hand, OSSFN employs a sample-based nonparametric mutual information estimate as the objective function, and can explore complex data structures.
2) Spectral filtering: CSP itself does not address selection of band-pass filter, FBCSP selects pairs of spatial features produced by CSP from each band-pass filter in a filter bank, where the selection is performed to maximize a mutual information estimate. OSSFN selects the optimum band-pass filter in conjunction with optimization of spatial filters in each band-pass filter, resulting in optimum spatio-spectral features for trial-by-trial EEG classification.

References

[1] J. R. Wolpaw, "Brain-computer interfaces as new brain output pathways," *J. Physiol.*, vol. 579, no. 3, pp. 613–619, Mar. 2007.

[2] A. Nijholt and D. Tan, "Brain-computer interfacing for intelligent systems," *IEEE Intell. Syst.*, vol. 23, no. 3, pp. 72–79, May–Jun. 2008.

[3] J. R. Wolpaw, N. Birbaumer, D. J. MacFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophysiol.*, vol. 113, no. 6, pp. 767–791, Jun. 2002.

[4] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer, "EEG-based discrimination between imagination of right and left hand movement," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 6, pp. 642–651, Dec. 1997.

[5] J. Muller-Gerking, G. Pfurtscheller, and H. Flyvbjerg, "Designing optimal spatial filtering of single trial EEG classification in a movement task," *Clin. Neurophysiol.*, vol. 110, no. 5, pp. 787–798, May 1999.

[6] A. Kübler, F. Nijboer, J. Mellinger, T. M. Vaughan, H. Pawelzik, G. Schalk, D. J. McFarland, N. Birbaumer, and J. R. Wolpaw, "Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface," *Neurology*, vol, 64, no. 10, pp. 1775–1777, May 2005.

[7] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 993–1002, Jun. 2004.

[8] M. Grosse-Wentrup and M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 8, pp. 1991–2000, Aug. 2008.

[9] B. Blankertz, G. Dornhege, C. Schafer, R. Krepki, J. Kohlmorgen, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio, "Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG analysis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 127–131, Jun. 2003.

[10] P. L. Nunez, R. Srinivasan, A. F. Westdorp, R. S. Wijesinghe, D. M. Tucker, R. B. Silberstein, and P. J. Cadusch, "EEG coherency I: Statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 5, pp. 499–515, Nov. 1997.

[11] P. L. Nunez, R. B. Silberstein, Z. Shi, M. R. Carpenter, R. Srinivasan, D. M. Tucker, S. M. Doran, P. J. Cadusch, and R. S. Wijesinghe, "EEG coherency II: Experimental comparisons of multiple measures," *Clin. Neurophysiol.*, vol. 110, no. 3, pp. 469–486, Mar. 1999.

[12] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, "EMG contamination of EEG: Spectral and topographical characteristics," *Clin. Neurophysiol.*, vol. 114, no. 9, pp. 1580–1593, Sep. 2003.

[13] A. Bashashati, M. Fatourechi, R. K. Ward, and G. E. Birch, "A survey of signal processing algorithms in brain-computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, no. 2, pp. R32–R57, Jun. 2007.

[14] L. Qin, L. Ding, and B. He, "Motor imagery classification by means of source analysis for brain-computer interface applications," *J. Neural Eng.*, vol. 1, no. 3, pp. 135–141, 2004.

[15] M. Grosse-Wentrup, C. Liefhold, K. Gramann, and M. Buss, "Beamforming in noninvasive brain-computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1209–1219, Apr. 2009.

[16] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[17] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing spatial filters for robust EEG single-trial analysis," *IEEE Signal Process. Mag.*, vol. 25, no. 1, pp. 41–56, Jan. 2008.

[18] M. Naeem, C. Brunner, R. Leeb, B. Graimann, and G. Pfurtscheller, "Separability of four-class motor imagery data using independent components analysis," *J. Neural Eng.*, vol. 3, no. 3, pp. 208–216, Sep. 2006.

[19] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-spectral filters for improving the classification of single trial EEG," *IEEE Trans. Biomed. Eng.*, vol. 52, no. 9, pp. 1541–1548, Sep. 2005.

[20] G. Dornhege, B. Blankertz, M. Krauledat, F. Losch, G. Curio, and K.-R. Müller, "Combined optimization of spatial and temporal filters for improving brain-computer interfacing," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 11, pp. 2274–2281, Nov. 2006.

[21] W. Wu, X. R. Gao, B. Hong, and S. K. Gao, "Classifying single-trial EEG during motor imagery by iterative spatio-spectral patterns learning (ISSPL)," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 6, pp. 1733–1743, Jun. 2008.

[22] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter bank common spatial pattern (FBCSP) in brain-computer interface," in *Proc. Int. Joint Conf. Neural Netw.*, Hong Kong, China, 2008, pp. 2391–2398.

[23] K. K. Ang, C. Guan, K. S. G. Chua, B. T. Ang, C. W. K. Kuah, C. Wang, K. S. Phua, Z. Y. Chin, and H. H. Zhang, "A clinical evaluation of non-invasive motor imagery-based brain-computer interface in stroke," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Vancouver, BC, Canada, Aug. 2008, pp. 4178–4181.

[24] *BCI Competition IV* [Online]. Available: http://www.bbci.de/competition/

[25] H. Zhang, C. Guan, and C. Wang, "Spatio-spectral feature selection based on robust mutual information estimate for brain-computer interfaces," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Minneapolis, MN, Sep. 2009, pp. 2391–2398.

[26] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.

[27] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[28] M. Ben-Bassat, "User of distance measures, information measures and error bounds in feature evaluation," in *Handbook of Statistics*, vol. 2, P. Krishnaiah and L. Kanal, Eds. Amsterdam, The Netherlands: North Holland, 1982, ch. 35, pp. 773–791.

[29] S. Petridis and S. J. Perantonis, "On the relation between discriminant analysis and mutual information for supervised linear feature extraction," *Pattern Recognit.*, vol. 37, no. 5, pp. 857–874, May 2004.

[30] J. M. Sotoca and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based distances," *Pattern Recognit.*, vol. 43, no. 6, pp. 2068–2081, Jun. 2010.

[31] P. A. Estevez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[32] P. Viola and W. M. Wells, III, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, Sep. 1997.

[33] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley, 1992.

[34] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. New York: Oxford Univ. Press, 1997.

[35] N. Yamawaki, C. Wilke, Z. Liu, and B. He, "An enhanced time-frequency-spatial approach for motor imagery classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 250–254, Jun. 2006.

[36] R. Scherer, F. Lee, A. Schlögl, R. Leeb, H. Bischof, and G. Pfurtscheller, "Toward self-paced brain-computer communication: Navigation through virtual worlds," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 2, pp. 675–682, Feb. 2008.

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

[38] C.-C. Chang and C.-J. Lin. (2001). *LIBSVM: A Library for Support Vector Machines* [Online]. Available: http://www.csie.ntu.edu.tw/∼cjlin/libsvm

[39] B. Obermaier, G. R. Muller, and G. Pfurtscheller, "Virtual keyboard controlled by spontaneous EEG activity," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 4, pp. 422–426, Dec. 2003.

[40] M. Fatourechi, R. K. Ward, and G. E. Birch, "A self-paced brain-computer interface system with a low false positive rate," *J. Neural Eng.*, vol. 5, no. 1, pp. 9–23, Mar. 2008.

[41] B. Blankertz, G. Dornhege, M. Krauledat, K.-R. Müller, and G. Curio, "The non-invasive Berlin brain-computer interface: Fast acquisition of effective performance in untrained subjects," *NeuroImage*, vol. 37, no. 2, pp. 539–550, Aug. 2007.

Haihong Zhang (M'07) received the Ph.D. degree in computer science from the National University of Singapore, Singapore, in 2005.

He joined the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, and later became a Principal Investigator for multimodal neural decoding. He is currently a Senior Research Fellow at the same institute. His current research interests include machine learning, pattern recognition, and brain signal processing for high-performance brain–computer interfaces.

Zheng Yang Chin (M'08) received the Masters degree in electrical engineering from the National University of Singapore, Singapore, in 2008.

He has been working on the brain–computer interface project at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, since 2005. His current research interests include signal processing and machine learning techniques for biomedical signal analysis.

Kai Keng Ang (S'05–M'07) received the B.A.Sc. (1st Hons.), M.Phil., and Ph.D. degrees in computer engineering from Nanyang Technological University, Singapore, in 1997, 1999, and 2008, respectively.

He was a Senior Software Engineer with Delphi Automotive Systems Singapore Pte Ltd., Singapore, from 1999 to 2003, working on embedded software for automotive engine controllers. He is currently a Senior Research Fellow working on brain–computer interface at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His current research interests include computational intelligence, machine learning, pattern recognition, and signal processing.

Cuntai Guan (S'91–M'92–SM'03) received the Ph.D. degree in electrical and electronic engineering from Southeast University, Nanjing, China, in 1993.

He is a Principal Scientist & Program Manager at Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. From 1993 to 1999, he worked on speech recognition in universities, research institutes, and industries. In 2003, he established the Brain-Computer Interface Laboratory at Institute for Infocomm Research, Singapore. His current research interests include brain-computer interface, neural signal processing, machine learning, pattern classification, and statistical signal processing, with applications to neuro-rehabilitation, health monitoring, and cognitive training.

Dr. Guan is an Associate Editor of *Frontiers in Neuroprosthetics*.

Chuanchu Wang (M'07) received the B.Eng. and M.Eng. degrees in communication and electrical engineering from the University of Science and Technology of China, Hefei, China, in 1988 and 1991, respectively.

He is currently a Research Manager at the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His current research interests include electroencephalogram based brain–computer interfacing and related applications.