Open access • Journal Article • DOI:10.1109/TPAMI.2012.69

A Novel Bayesian Framework for Discriminative Feature Extraction in Brain-Computer Interfaces — Source link

Heung-Il Suk, Seong-Whan Lee

Institutions: Korea University

Published on: 01 Feb 2013 - IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE Computer Society)

Topics: Discriminative model, Feature extraction, Probabilistic logic and Decision rule

A Novel Bayesian Framework for Discriminative Feature Extraction in Brain-Computer Interfaces

Heung-Il Suk, *Student Member*, *IEEE*, and Seong-Whan Lee, *Fellow*, *IEEE*

Abstract—As there has been a paradigm shift in the learning load from a human subject to a computer, machine learning has been considered as a useful tool for Brain-Computer Interfaces (BCIs). In this paper, we propose a novel Bayesian framework for discriminative feature extraction for motor imagery classification in an EEG-based BCI in which the class-discriminative frequency bands and the corresponding spatial filters are optimized by means of the probabilistic and information-theoretic approaches. In our framework, the problem of simultaneous spatiospectral filter optimization is formulated as the estimation of an unknown posterior probability density function *(pdf)* that represents the probability that a single-trial EEG of predefined mental tasks can be discriminated in a state. In order to estimate the posterior *pdf*, we propose a particle-based approximation method by extending a factored-sampling technique with a diffusion process. An information-theoretic observation model is also devised to measure discriminative power of features between classes. From the viewpoint of classifier design, the proposed method naturally allows us to construct a spectrally weighted label decision rule by linearly combining the outputs from multiple classifiers. We demonstrate the feasibility and effectiveness of the proposed method by analyzing the results and its success on three public databases.

Index Terms—Discriminative feature extraction, spatiospectral filter optimization, Brain-Computer Interface (BCI), ElectroEncephaloGraphy (EEG), motor imagery classification

1 Introduction

A Brain-Computer Interface (BCI), also called a Brain-Machine Interface (BMI), is a state-of-the-art technology that translates neuronal activities into user commands, and thereby can establish a direct communication pathway between the brain and an external device. Due to the huge potential of BCIs in medical and industrial applications for both disabled and normal people, it has been of great interest to many research groups [1], [2], [3], [4].

One of the revolutionary changes in the history of BCI is that there has been a paradigm shift in the realization of the BCI system; the learning burden has shifted from the subject to the computer [5]. Consequently, machine learning has been considered as the main tool for the classification or analysis of brain signals. A comprehensive review of machine learning algorithms for BCI is available in [6].

Despite the successful application of machine learning techniques to BCI, the high complexity of the human brain and the low signal-to-noise ratio in EEG signals prevent EEG-based BCI systems from decoding every human mental state or intention. Among a small subset of brain states widely considered in BCI, increasing attention has been devoted to the analysis of EEG signals induced by imagined body-part movement, called motor imagery [1]. This is because of its active and voluntary strategy for generating a specific regulation of an EEG pattern in frequency band(s), known as Event-Related Desynchonization (ERD) or Event-Related Synchronization (ERS). The neurophysiological ERD or ERS phenomenon during motor imagery is, respectively, the suppression or augmentation of the signal power in the motor and somatosensory cortex due to the loss of synchrony, in particular, in the frequency ranges: μ-rhythm (8-14 Hz) and β-rhythm (14-30 Hz).

However, the ERD/ERS patterns exhibit high variability across subjects and even trials for the same subject [7]. Therefore, finding the ERD/ERS-related frequency band(s) and the class-discriminative spatial patterns have been the main issues in the BCI community. Considerable efforts have been devoted to these problems in terms of signal processing, computational modeling and recognition, and analysis of brain signals [6], [8].

In this paper, we propose a novel framework of spatiospectral filter optimization for discriminative feature extraction in BCI. The main contributions of the paper are threefold. First, we build a Bayesian framework in which the class-discriminative frequency bands are probabilistically selected and the corresponding spatial filters are optimized. In our framework, a frequency band is represented as a random variable. The problem of optimizing the spatiospectral filter is formulated as the estimation of the posterior probability density function (*pdf*) that represents the relative probabilities of different states in discriminating single-trial EEGs. Second, we propose a particle-based

- *H.-I. Suk is with the Department of Computer Science and Engineering, Woo-Jung College of Information and Communications Building, Korea University, Room 408, Anam-dong, Seongbuk-ku, Seoul 136-713, Korea. E-mail: hisuk@image.korea.ac.kr.*
- *S.-W. Lee is with the Department of Brain and Cognitive Engineering, Woo-Jung College of Information and Communications Building, Korea University, Room 410, Anam-dong, Seongbuk-ku, Seoul 136-713, Korea. E-mail: swlee@image.korea.ac.kr.*

Manuscript received 19 Sept. 2011; revised 22 Feb. 2012; accepted 7 Mar. 2012; published online 13 Mar. 2012.
Recommended for acceptance by T. Cootes.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2011-09-0661.
Digital Object Identifier no. 10.1109/TPAMI.2012.69.

0162-8828/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

approximation method for the estimation of the posterior *pdf* by extending a factored-sampling technique with a diffusion process, and an observation model, which measures discriminative power of features between classes, by means of an information-theoretic approach. Third, thanks to the features of the proposed *pdf* estimation method, we can construct a spectrally weighted label decision rule by linearly combining the outputs from multiple classifiers. A preliminary partial version of this work was presented in [9].

The rest of the paper is organized as follows: We start by reviewing work on motor imagery classification in the literature in Section 2. A novel Bayesian framework for discriminative feature extraction by means of spatiospectral filter optimization, particle-based posterior *pdf* estimation, information-theoretic likelihood computation, and a spectrally weighted classification rule are described in Section 3. The experimental results are compared with competing methods in the literature and the discussions on this are detailed in Section 4. We conclude this paper by summarizing the proposed method and providing directions for future work in Section 5.

2 Related Work

Utilizing ERD/ERS patterns in a single-trial EEG, many works in the literature perform the following three prevalent steps for motor imagery classification:

1. spectral filtering: $\mathbf{z} = h \otimes \mathbf{x}$,
2. spatial filtering: $\mathbf{y} = \mathbf{W}^{\dagger}\mathbf{z}$,
3. feature extraction: $\mathbf{f} = log(var(\mathbf{y}))$,

where $\mathbf{x}$ denotes a single-trial EEG, h and $\mathbf{W}$ denote, respectively, a spectral filter and a spatial filter, $\mathbf{z}$ denotes the spectrally filtered EEG, and $\mathbf{y}$ denotes the spatially filtered signal of $\mathbf{z}$. Hereafter, the superscript $\dagger$ denotes a matrix transpose operation throughout this paper. The feature vector $\mathbf{f}$ composed of the logarithmic (log) values of the second-order statistics (var) of the signal $\mathbf{y}$ is fed into a classifier for training or evaluation.

Unfortunately, as stated above, the motor imagery responsive frequency band(s) varies across subjects and trials even for the same subject, and there is no general analytic method for finding the optimal frequency band(s), accordingly the spectral filter. Hereafter, we use the terms spectral filter and frequency band interchangeably. While some groups have considered subject-dependent frequency band selection [10], [11], [12], the frequency bands on which the rest of the processes in a BCI system operate are either selected manually based on visual inspection or unspecifically set to a broadband in many researches [6], [7], [13], [14], [15]. Contrary to the subject-dependent approach, there have also been the efforts of developing subject-independent methods for more practical applications shortening time-consuming calibration recordings [16], [17], [18].

For the optimization of a spatial filter $\mathbf{W}$, in contrast, the CSP algorithm [7] that utilizes a generalized eigenvalue problem is widely used [14], [15], [19]. The algorithm finds spatial patterns that transform signals of two classes to be maximally discriminative based on the ratio of the variance of the data conditioned on one class and the variance of the data conditioned on the other class. The learned spatial pattern is, however, highly dependent on the operating frequency band, thereby a spectral filter h. Therefore, from an optimization point of view, the spatiospectral filter should be simultaneously optimized from data in a unified framework for each subject.

Lemm et al. proposed a common spatiospectral pattern method that puts it into a CSP algorithm for spatially augmented signals with embedding of the time-delayed signals as new channels, resulting in doubling of the number of channels [20]. But the flexibility of the method is very limited due to the first-order FIR filter optimization approach. Furthermore, the time delay parameter should be tuned empirically. Later, Dornhege et al. extended Lemm et al.'s work, increasing the flexibility of FIR filters by optimizing an arbitrary FIR filter within the CSP analysis and enforcing sparse filter coefficients with an introduction of a regularization term [21].

Tomioka et al. devised a method that alternatively optimizes the spatial projection matrix and nonhomogeneous weighting coefficients of the cross-spectrum matrices [22]. Wu et al. tried to directly optimize spectral filters, FIR in the spectral domain, to achieve maximal classification accuracy. They proposed an iterative spatiospectral pattern learning method that employs a statistical learning theory for automatic learning of spatiospectral filter [23].

More recently, some groups have focused on finding an optimal spatiospectral filter by constructing a filter bank and applying a CSP algorithm in each frequency band independently. Ang et al. proposed a Filter Bank CSP (FBCSP) that first dissects a broadband frequency range of interest into predefined small nonoverlapping frequency bands with a fixed bandwidth [19]. The method applies a CSP algorithm to each band independently and uses a maximal mutual information criterion to select a discriminative feature set. Thomas et al. proposed a method of selecting subject-specific discriminative filter bank CSP using a coefficient decimation technique [24]. Last but not least, Zhang et al. proposed an Optimal Spatiospectral Filter Network (OSSFN) for which a gradient-based learning algorithm was devised to find optimal spatial filters in conjunction with a filter bank and mutual information [25]. Although the application of a filter bank covering the μ- and β-rhythms and the optimization of spatial patterns in each band improved the classification performance in their work, the frequency bands were still predetermined and fixed based on prior neurophysiological knowledge.

In this paper, we propose a novel framework of optimizing spatiospectral filter in probabilistic and information-theoretic approaches, which extracts discriminative features to improve classification performance. Table 1 gives a brief comparison of various methods in the literature for feature extraction and recognition of motor imagery tasks in an EEG-based BCI. To the best of our knowledge, this is the first time a method of finding the optimal spatiospectral filter in a probabilistic Bayesian approach has been proposed.

3 Bayesian Spatiospectral Filter Optimization (BSSFO)

In order for discriminative feature extraction, we consider the following problems: 1) How should we compose a filter bank—how many frequency bands should there be and

TABLE 1
A Summary of Spatiospectral Filter Optimization Methods in the Literature

Authors	Methods and Limitations	Database
Lemm *et al.* [20]	CSP on spatially augmented signals with embedded time-delayed signals, empirically determined time delay parameter, first-order FIR filter optimization	Self-collected dataset
Dornhege *et al.* [21]	Extended version of Lemm *et al.*'s work, enforcement of sparse filter coefficients with a regularization technique, computationally inefficient in optimization	Self-collected dataset
Tomioka *et al.* [22]	Alternative optimization of a spatial filter and weighting coefficients of the cross-spectrum matrices, requirement of an extensive number of parameters to be tuned empirically	Self-collected dataset
Ang *et al.* [19]	Dissection of a broadband into predefined small non-overlapping frequency bands with a fixed-bandwidth	BCI Competition III-IVa, self-collected dataset
Wu *et al.* [23]	Iterative optimization of spectral filters related to the classification of different classes	BCI Competition III-IVa, self-collected dataset
Thomas *et al.* [24]	A subject-dependent filter bank with a coefficient decimation technique, non-overlapping frequency bands and fixed-bandwidth	BCI Competition III-IVa, BCI Competition IV-IIb
Zhang *et al.* [25]	Gradient-based learning for spatial filter optimization in a filter bank, predefined non-overlapping frequency bands and fixed-bandwidth	BCI Competition IV-I, self-collected dataset
Proposed method	Data-driven discriminative filter bank construction and bandwidth selection in a Bayesian framework, particle-based posterior *pdf* estimation, information-theoretic observation model, spectrally weighted decision rule	Technische Universität Berlin dataset, BCI Competition III-IVa, BCI Competition IV-IIa

how wide should each frequency band be? and 2) how can we measure the discriminative power of features between classes?

We use a discriminative approach and propose a novel Bayesian framework for simultaneous optimization of the spectral and spatial filters, which we call *Bayesian Spatiospectral Filter Optimization*. A particle-based posterior *pdf* estimation method and an information-theoretic observation model are also devised. Taking advantage of the particle-based *pdf* estimation technique, a spectrally weighted classifier construction is also described. Refer to Table 2 for the description of the notations used throughout this paper.

3.1 Problem Formulation

Let us denote $\mathbf{B} = [b^s, b^e]^\dagger$ as a continuous random vector for a frequency band, where b^s and b^e are, respectively, the start and the end frequency of the band range with the constraint of $b^s < b^e$. We define the probability of a frequency band $\mathbf{b}$, $p(\mathbf{b})$, as the chance that the $\mathbf{b}$ bandpass-filtered signals can be correctly classified between two classes.

Since we are presumably uncertain about the discriminative frequency band, we encode this uncertainty as a prior distribution $p(\mathbf{B})$ over a random variable $\mathbf{B}$. Given a set of single-trial EEGs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{D}$ and the corresponding class labels $\mathbf{\Omega} = \{\omega_i\}_{i=1}^{D}$, where D is the number of trials, we can compute the posterior *pdf*, $p(\mathbf{B}|\mathbf{X}, \mathbf{\Omega})$, by the Bayes rule as follows:

$$p(\mathbf{B}|\mathbf{X}, \mathbf{\Omega}) = \frac{p(\mathbf{X}, \mathbf{\Omega}|\mathbf{B})p(\mathbf{B})}{p(\mathbf{X}, \mathbf{\Omega})}. \tag{1}$$

The prior, $p(\mathbf{B})$, describes the relative probabilities of different states, i.e., frequency bands, in which single-trial EEGs pertinent to motor imageries are correctly discriminated. The term $p(\mathbf{X}, \mathbf{\Omega}|\mathbf{B})$ is called the likelihood function. If the hypothesis $\mathbf{B}$, i.e., the frequency band, were true, this term indicates the probability that the single-trial EEGs $\mathbf{X}$ in conjunction with the class labels $\mathbf{\Omega}$ would have been available to support it. The posterior distribution $p(\mathbf{B}|\mathbf{X}, \mathbf{\Omega})$ defines the probability of frequency band $\mathbf{B}$ being true, given the observations of $\mathbf{X}$ and $\mathbf{\Omega}$. Thus, it indicates the relative likelihood of the single-trial EEGs $\mathbf{X}$ being correctly classified into $\mathbf{\Omega}$ by $\mathbf{B}$ bandpass filtering along with the ensuing computational processes. Note that in this paper we do not make any functional assumption about the

TABLE 2
Description of the Notations Used in This Paper

Notation	Description	Notation	Description
$\mathbf{x}_n$	A single-trial EEG	$\mathbf{Y}_k$	Spatially filtered signals of $\mathbf{Z}_k$
$\mathbf{X}$	$\{\mathbf{x}_n\}_{n=1}^N$, a set of single-trial EEGs	$\mathbf{Y}$	$\{\mathbf{Y}_k\}_{k=1}^K$, spatially filtered signals
h	Bandpass filter	$\mathbf{F}_k$	A feature vector set extracted from $\mathbf{Y}_k$
$\mathbf{W}$	A spatial transformation matrix	$\mathbf{F}$	$\{\mathbf{F}_k\}_{k=1}^K$, feature vector sets
$\mathbf{B}$	A random vector for a frequency band	ω	A particular class label
$\mathcal{B}$	$\{\mathbf{b}_k\}_{k=1}^K$, a set of random variates of $\mathbf{B}$	$\boldsymbol{\Omega}$	A set of the class labels
$\mathbf{Z}_k$	$\mathbf{b}_k$ bandpass-filtered signals	$H(\mathbf{a})$	Shannon's entropy of a random vector $\mathbf{a}$
$\mathbf{Z}$	$\{\mathbf{Z}_k\}_{k=1}^K$, filter bank-filtered signals	$I(\mathbf{F};\boldsymbol{\Omega})$	Mutual information between $\mathbf{F}$ and $\boldsymbol{\Omega}$

densities $p(\mathbf{B})$ and $p(\mathbf{B}|\mathbf{X},\boldsymbol{\Omega})$, such as linearity, Gaussianity, unimodality, etc.

Given a frequency band $\mathbf{B}$ and raw EEG signals $\mathbf{X}$, the bandpass-filtered signals $\mathbf{Z}$ are deterministically obtained. Hence, the likelihood $p(\mathbf{X},\boldsymbol{\Omega}|\mathbf{B})$ and the evidence $p(\mathbf{X},\boldsymbol{\Omega})$ are equal to $p(\mathbf{Z},\boldsymbol{\Omega}|\mathbf{B})$ and $p(\mathbf{Z},\boldsymbol{\Omega})$, respectively. We can then rewrite (1), replacing the raw EEGs $\mathbf{X}$ with the bandpass-filtered signals $\mathbf{Z}$ as follows:

$$p(\mathbf{B}|\mathbf{Z},\boldsymbol{\Omega}) = \frac{p(\mathbf{Z},\boldsymbol{\Omega}|\mathbf{B})p(\mathbf{B})}{p(\mathbf{Z},\boldsymbol{\Omega})}. \tag{2}$$

The posterior $p(\mathbf{B}|\mathbf{Z},\boldsymbol{\Omega})$ represents all the knowledge about $\mathbf{B}$ that is deducible from the bandpass-filtered single-trials $\mathbf{Z}$ and the corresponding class labels $\boldsymbol{\Omega}$.

Note that a spatial filter $\mathbf{W}$ is found from $\mathbf{Z}$ via a standard CSP algorithm [13] or its variants [14], [15], [16], [26], in which $\mathbf{W}$ is analytically obtained by computing a generalized eigenvector problem. As described above, a feature vector is extracted by computing simple matrix multiplication between $\mathbf{Z}$ and $\mathbf{W}$ and the second-order statistics followed by a monotonically increasing logarithmic function. It means that the posterior $p(\mathbf{B}|\mathbf{Z},\boldsymbol{\Omega})$ can be indirectly estimated from $p(\mathbf{B}|\mathbf{F},\boldsymbol{\Omega})$, where $\mathbf{F} = log[var(\mathbf{W}^{\dagger}\mathbf{Z})]$, without losing information in the data. Therefore, we can rewrite (2) as follows with the feature vector set $\mathbf{F}$ extracted from the spatially filtered signals of $\mathbf{Z}$:

$$\begin{aligned} p(\mathbf{B}|\mathbf{Z},\boldsymbol{\Omega}) &\triangleq p(\mathbf{B}|\mathbf{F},\boldsymbol{\Omega}) \\ &= \frac{p(\mathbf{F},\boldsymbol{\Omega}|\mathbf{B})p(\mathbf{B})}{p(\mathbf{F},\boldsymbol{\Omega})}, \end{aligned} \tag{3}$$

where $p(\mathbf{F},\boldsymbol{\Omega}) = \int_{\mathbf{B}} p(\mathbf{F},\boldsymbol{\Omega}|\mathbf{B})p(\mathbf{B})d\mathbf{B}$. Thus, our goal of finding the optimal spatiospectral filter for discriminative feature extraction, ultimately improving classification accuracy, can be defined as estimation of the posterior *pdf* $p(\mathbf{B}|\mathbf{F},\boldsymbol{\Omega})$ in (3).

3.2 Posterior Estimation

Although there exists informative neurophysiological knowledge about the rhythmic activity involved in motor imagery, the functional form of the *pdf* $p(\mathbf{B})$ is unknown. Furthermore, in this case, where $p(\mathbf{F},\boldsymbol{\Omega}|\mathbf{B})$ in (3) is sufficiently complex that $p(\mathbf{B}|\mathbf{F},\boldsymbol{\Omega})$ cannot be directly evaluated in a closed form, a particle(sample)-based approximation technique can be used. Hereafter, we use the term "*particle*" instead of "*sample*" in order to avoid the terminological confusion with the sample in an EEG.

We utilize the sophisticated factored-sampling algorithm [27] for the estimation of the unknown-formed *pdf*. Assume that a particle-set $\mathcal{B} = \{\mathbf{b}_k\}_{k=1}^K$ is generated from the prior density $p(\mathbf{B})$, where $\mathbf{b}_k$ denotes a particle representing a single frequency band. Based on these particles we can compute the weight of each particle as follows:

$$\pi_k = \frac{p(\mathbf{F}_k,\boldsymbol{\Omega}|\mathbf{b}_k)}{\sum_{i=1}^K p(\mathbf{F}_i,\boldsymbol{\Omega}|\mathbf{b}_k)}, \tag{4}$$

where $\mathbf{F}_k$ denotes a feature vector set extracted from the spectrally ($\mathbf{b}_k$) and spatially ($\mathbf{W}_k$) filtered signals and $p(\mathbf{F}_k,\boldsymbol{\Omega}|\mathbf{b}_k)$ denotes the conditional observation density. As a result, the weighted particle-set $\mathcal{B} = \{\mathbf{b}_k, \pi_k\}_{k=1}^K$ approximates the distribution of the posterior $p(\mathbf{B}|\mathbf{F}_k,\boldsymbol{\Omega})$ and it converges to the true density as the number of particles increases.

We iteratively apply the factored-sampling algorithm until it converges in the estimation of the posterior $p(\mathbf{B}|\mathbf{F}_k,\boldsymbol{\Omega})$ in order to circumvent the effect of the biased prior density; this is similar to the burn-in step in the Markov chain Monte Carlo algorithm [28]. We believe that iteration helps obtain a stabilized approximation of the *pdf* by considering the potential states around a given state, i.e., frequency band, as described below.

It is clear that the algorithm needs to begin with a prior density $p(\mathbf{B})$ from which we can generate particles at each iteration. From a Bayesian point of view, the effective prior for the tth iteration should be derived from the output of the previous iteration, i.e., the weighted particle set representation $\mathcal{B}^{(t-1)} = \{\mathbf{b}_k^{t-1}, \pi_k^{t-1}\}_{k=1}^K$.

In the tth iteration, our first task is to choose a particle $\mathbf{b}_k^t$ with a probability π_k^{t-1} with replacement from $\mathcal{B}^{(t-1)}$, and we repeat the operation K times, resulting in a new particle-set $\mathcal{B}^{(t)}$. In this manner, some particles, especially those with high weights, can be chosen multiple times, leading to identical copies in the new particle set. We therefore apply a diffusion method to the new particle set by adding a Gaussian noise as follows:

$$\mathbf{b}'_k = \begin{cases} \mathbf{b}_k + \sigma & \text{if, } \psi(k) > 1, \\ \mathbf{b}_k, & \text{otherwise,} \end{cases} \tag{5}$$

where $\psi(k)$ denotes the number of times that the kth particle in $\mathcal{B}^{(t-1)}$ is chosen while composing the new particle set $\mathcal{B}^{(t)}$ and σ is a normally distributed diffusion noise. By applying the diffusion process only to the particles that have already been in the new particle set $\mathcal{B}^{(t)}$, we can avoid losing the current optimal state. The rationale for the application of the diffusion method in *pdf* estimation is that it first allows us to consider the states around the current optimal frequency band, which might prevent us from falling into a local optimum, and it also prevents the particles from converging to a single local optimal solution as the method iterates. The sequential application of the factored-sampling and diffusion method is similar to the Sequential Monte Carlo algorithm [29], also known as the particle-filter algorithm, which is developed for object tracking in computer vision [30].

The main advantage of this particle-based approximation of the posterior density $p(\mathbf{B}|\mathbf{F},\mathbf{\Omega})$ is that we can naturally obtain a data-driven filter bank that is composed of multiple particles, each of which may have a different weight and bandwidth and can possibly be overlapped. This overcomes the limitation of the previous work of [19], [24], [25] in which the authors constructed a filter bank by dissecting a broadband into predefined and fixed nonoverlapping frequency bands with a uniform bandwidth. Another important feature of the proposed method is that the set of the particles $\{\mathbf{b}_k\}_{k=1}^K$ and the corresponding weights $\{\pi\}_{k=1}^K$ allows us to design a spectrally weighted classification rule as described below.

3.3 Likelihood Estimation with Mutual Information

The likelihood computation in (3) is another big challenge. In order to meet this challenge, this paper introduces mutual information, which measures the mutual dependence of two random variables or reduction in uncertainty of random variables. The information-theoretic approach has recently received considerable attention in both the BCI [25], [31], [32] and the machine learning communities [33], [34], [35] for the selection of an informative subset from original features. Unlike that use of mutual information, in this paper, we consider it for the likelihood computation in the proposed Bayesian framework. We use it to measure the discriminative power of features in terms of classifying single-trial EEGs from the probabilistic viewpoint.

We define the likelihood $p(\mathbf{F}_k,\mathbf{\Omega}|\mathbf{b}_k)$ in (4) as follows:

$$p(\mathbf{F}_k,\mathbf{\Omega}|\mathbf{b}_k) \equiv exp[I(\mathbf{F}_k;\mathbf{\Omega})], \tag{6}$$

where $I(\mathbf{F}_k;\mathbf{\Omega})$ denotes the mutual information between the feature vector set $\mathbf{F}_k$ and the class label set $\mathbf{\Omega}$. This clearly reflects our intention of computing the discriminative power between classes based on the features extracted from the $\mathbf{b}_k$ bandpass-filtered and $\mathbf{W}_k$ spatially transformed signals.

The mutual information $I(\mathbf{F}_k;\mathbf{\Omega})$ is defined as follows:

$$I(\mathbf{F}_k;\mathbf{\Omega}) = H(\mathbf{F}_k) - H(\mathbf{F}_k|\mathbf{\Omega}), \tag{7}$$

where $H(\cdot)$ and $H(\cdot|\cdot)$ denote, respectively, the entropy and the conditional entropy. For continuous random variables, the entropy and mutual information are defined as

$$H(\mathbf{F}_k) = -\int p(\mathbf{F}_k) log(p(\mathbf{F}_k)) d\mathbf{F}_k, \tag{8}$$

$$H(\mathbf{F}_k|\mathbf{\Omega}) = -\sum_{\omega\in\mathbf{\Omega}} \int p(\mathbf{F}_k|\omega) log(p(\mathbf{F}_k|\omega)) d\mathbf{F}_k, \tag{9}$$

where ω is a particular class label, which is either positive $(+)$ or negative $(-)$ in this paper. Given a feature vector set $\mathbf{F}_k = \{\mathbf{f}_k^i\}_{i=1}^D$, the entropy and the conditional entropy can be approximated by

$$H(\mathbf{F}_k) \cong -\frac{1}{D}\sum_{i=1}^{D} log\big(p(\mathbf{f}_k^i)\big), \tag{10}$$

$$H(\mathbf{F}_k|\omega = c) \cong -\frac{1}{D_c}\sum_{j \text{ s.t.} \omega_j = c} log\big(p(\mathbf{f}_k^j|\omega = c)\big), \tag{11}$$

where $c \in \{+,-\}$, D denotes the total number of trials, and D_c denotes the number of trials of the class c.

However, it is still very difficult for the continuous value $\mathbf{f}_k^i$ to estimate the *pdf*s, $p(\mathbf{f}_k^i)$ and $p(\mathbf{f}_k^i|\omega)$. In this paper, we estimate the underlying *pdf*s using a Parzen window density estimator [33], which involves the superposition of a normalized window function centered on a set of training data, i.e., single-trial EEGs. Given the feature vector set $\mathbf{F}_k$, the density function $p(\mathbf{f}_k)$ is estimated by

$$\hat{p}(\mathbf{f}_k) = \frac{1}{D}\sum_{i=1}^{D} \varphi\big(\mathbf{f}_k - \mathbf{f}_k^i, \nu\big), \tag{12}$$

where $\varphi(\cdot)$ is a window function and ν is a window width parameter determining the smoothness of the window function. With the appropriate selection of $\varphi(\cdot)$ and ν, the estimated density function $\hat{p}(\mathbf{f}_k)$ converges to the true density [36]. Here, we use a multivariate Gaussian window function defined as

$$\varphi(\mathbf{a},\nu) = \frac{1}{(2\pi)^{d/2}\nu^d|\Sigma|^{1/2}} exp\left[-\frac{\mathbf{a}^{\dagger}\Sigma^{-1}\mathbf{a}}{2\nu^2}\right], \tag{13}$$

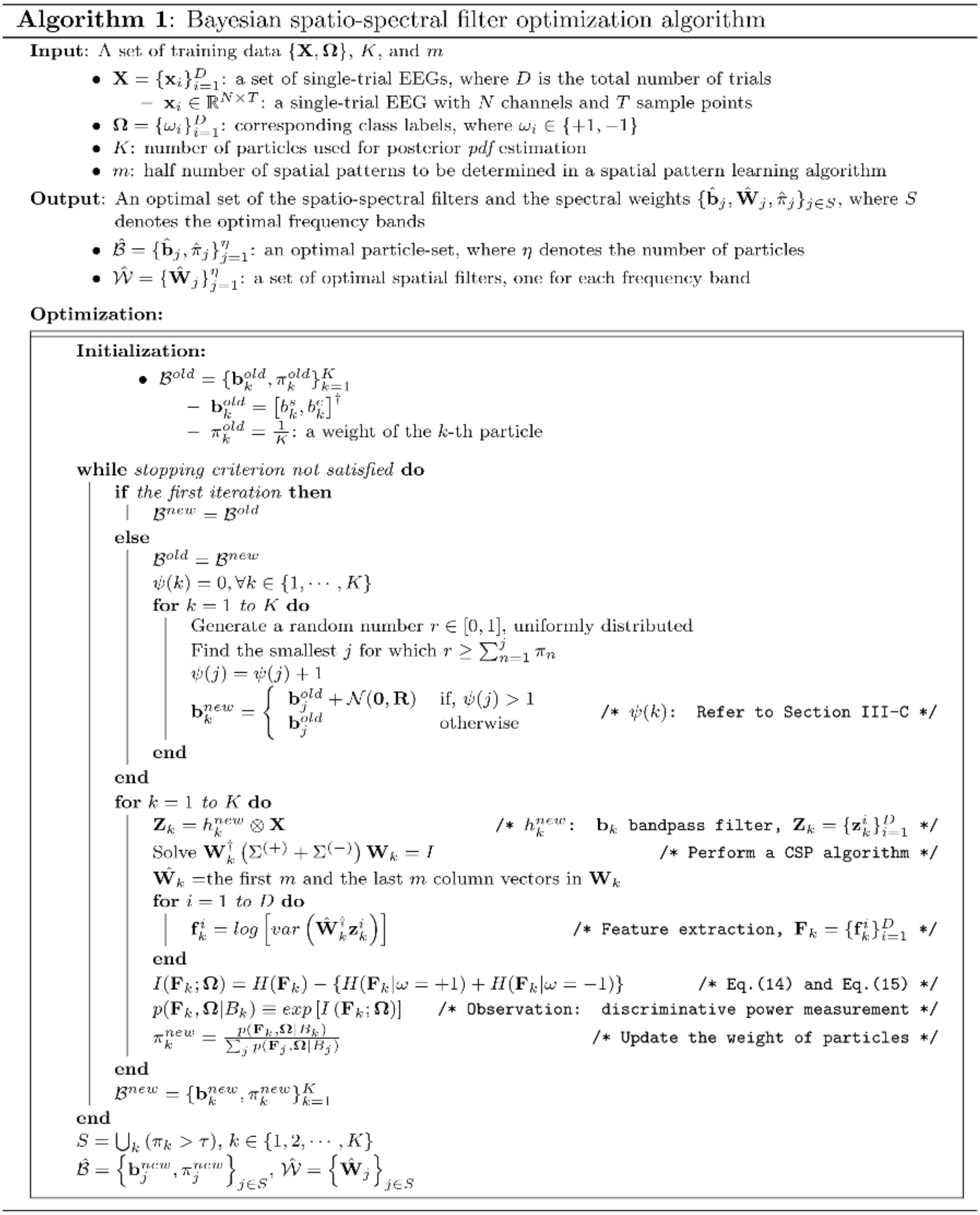
where Σ denotes a covariance matrix, $\mathbf{a}$ denotes a d-dimensional random vector, and ν denotes a width of the window function.

We can rewrite (10) and (11) with the introduction of (12) and (13) as follows:

$$H(\mathbf{F}_k) \cong -\frac{1}{D}\sum_{i=1}^{D} log\left[\frac{1}{D}\sum_{j=1}^{D}\varphi\big(\mathbf{f}_k^i - \mathbf{f}_k^j, \nu\big)\right], \tag{14}$$

$$H(\mathbf{F}_k|\omega = c) \cong -\frac{1}{D_c}\sum_{(i \text{ s.t.} \omega_i = c)} log\left[\frac{1}{D_c}\sum_{(j \text{ s.t.} \omega_j = c)}\varphi\big(\mathbf{f}_k^i - \mathbf{f}_k^j, \nu\big)\right], \tag{15}$$

where D and D_c denote, respectively, the total number of trials in a training dataset and the number of trials of the class c. Based on the values obtained from (14) and (15), we



```
Algorithm 1: Bayesian spatio-spectral filter optimization algorithm
Input: A set of training data $\{\mathbf{X}, \mathbf{\Omega}\}$, $K$, and $m$
    • $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{D}$: a set of single-trial EEGs, where $D$ is the total number of trials
        – $\mathbf{x}_i \in \mathbb{R}^{N \times T}$: a single-trial EEG with $N$ channels and $T$ sample points
    • $\mathbf{\Omega} = \{\omega_i\}_{i=1}^{D}$: corresponding class labels, where $\omega_i \in \{+1, -1\}$
    • $K$: number of particles used for posterior pdf estimation
    • $m$: half number of spatial patterns to be determined in a spatial pattern learning algorithm
Output: An optimal set of the spatio-spectral filters and the spectral weights $\{\hat{\mathbf{b}}_j, \hat{\mathbf{W}}_j, \hat{\pi}_j\}_{j \in S}$, where $S$
        denotes the optimal frequency bands
    • $\hat{\mathcal{B}} = \{\hat{\mathbf{b}}_j, \hat{\pi}_j\}_{j=1}^{\eta}$: an optimal particle-set, where $\eta$ denotes the number of particles
    • $\hat{\mathcal{W}} = \{\hat{\mathbf{W}}_j\}_{j=1}^{\eta}$: a set of optimal spatial filters, one for each frequency band
Optimization:
    Initialization:
        • $\mathcal{B}^{old} = \{\mathbf{b}_k^{old}, \pi_k^{old}\}_{k=1}^{K}$
            – $\mathbf{b}_k^{old} = [b_k^s, b_k^e]^\dagger$
            – $\pi_k^{old} = \frac{1}{K}$: a weight of the $k$-th particle
    while stopping criterion not satisfied do
        if the first iteration then
            $\mathcal{B}^{new} = \mathcal{B}^{old}$
        else
            $\mathcal{B}^{old} = \mathcal{B}^{new}$
            $\psi(k) = 0, \forall k \in \{1, \cdots, K\}$
            for $k = 1$ to $K$ do
                Generate a random number $r \in [0, 1]$, uniformly distributed
                Find the smallest $j$ for which $r \geq \sum_{n=1}^{j} \pi_n$
                $\psi(j) = \psi(j) + 1$
                $\mathbf{b}_k^{new} = \begin{cases} \mathbf{b}_j^{old} + \mathcal{N}(\mathbf{0}, \mathbf{R}) & \text{if, } \psi(j) > 1 \\ \mathbf{b}_j^{old} & \text{otherwise} \end{cases}$    /* $\psi(k)$: Refer to Section III-C */
            end
        end
        for $k = 1$ to $K$ do
            $\mathbf{Z}_k = h_k^{new} \otimes \mathbf{X}$    /* $h_k^{new}$: $\mathbf{b}_k$ bandpass filter, $\mathbf{Z}_k = \{\mathbf{z}_k^i\}_{i=1}^{D}$ */
            Solve $\mathbf{W}_k^\dagger \left(\Sigma^{(+)} + \Sigma^{(-)}\right) \mathbf{W}_k = I$    /* Perform a CSP algorithm */
            $\hat{\mathbf{W}}_k$ =the first $m$ and the last $m$ column vectors in $\mathbf{W}_k$
            for $i = 1$ to $D$ do
                $\mathbf{f}_k^i = log\left[var\left(\hat{\mathbf{W}}_k^\dagger \mathbf{z}_k^i\right)\right]$    /* Feature extraction, $\mathbf{F}_k = \{\mathbf{f}_k^i\}_{i=1}^{D}$ */
            end
            $I(\mathbf{F}_k; \mathbf{\Omega}) = H(\mathbf{F}_k) - \{H(\mathbf{F}_k|\omega = +1) + H(\mathbf{F}_k|\omega = -1)\}$    /* Eq.(14) and Eq.(15) */
            $p(\mathbf{F}_k, \mathbf{\Omega}|B_k) = exp\left[I(\mathbf{F}_k; \mathbf{\Omega})\right]$    /* Observation: discriminative power measurement */
            $\pi_k^{new} = \frac{p(\mathbf{F}_k, \mathbf{\Omega}|B_k)}{\sum_j p(\mathbf{F}_j, \mathbf{\Omega}|B_j)}$    /* Update the weight of particles */
        end
        $\mathcal{B}^{new} = \{\mathbf{b}_k^{new}, \pi_k^{new}\}_{k=1}^{K}$
    end
    $S = \bigcup_k (\pi_k > \tau)$, $k \in \{1, 2, \cdots, K\}$
    $\hat{\mathcal{B}} = \left\{\mathbf{b}_j^{new}, \pi_j^{new}\right\}_{j \in S}$, $\hat{\mathcal{W}} = \left\{\hat{\mathbf{W}}_j\right\}_{j \in S}$
```

Fig. 1. The proposed Bayesian spatiospectral filter optimization algorithm for class-discriminative feature extraction.

can obtain an estimate of the mutual information between the feature vectors and the class labels in (7). Thus, we can eventually compute the likelihood of (6). Refer to Fig. 1 for the complete algorithm of the proposed Bayesian spatiospectral filter optimization method for class-discriminative feature extraction.

3.4 Spectrally Weighted Classification

As explained above, the output from the proposed Bayesian framework is the particle set $\{\mathbf{b}_k, \pi_k\}_{k=1}^{K}$ and the spatial filter set $\{\mathbf{W}_k\}_{k=1}^{K}$. That is, it finds the class-discriminative frequency bands, represented by the particles comprising the data-driven filter bank, and the spatial patterns, one for each band. Consequently, a set of spatiospectral filters $\{\mathbf{b}_k, \mathbf{W}_k\}_{k=1}^{K}$ optimized in probabilistic and information-theoretic manners is obtained. An important point here is that we also obtain the weights $\{\pi_k\}_{k=1}^{K}$ for the class-discriminative frequency bands besides the spatiospectral filter. We utilize the informative spectral weights in constructing a classifier.

The classifier training is preceded by frequency bands selection based on the weights of particles. The motivation for this step is the possibility of drawing a low-probability particle because of the nature of particle-based posterior *pdf* estimation. We compose an optimal filter bank S with the set of class-discriminative frequency bands selected by the following rule:

$$S = \bigcup_k (\pi_k > \tau), \tag{16}$$

where $k \in \{1, 2, \ldots, K\}$ and τ denotes a threshold parameter that is determined empirically.

Then a Support Vector Machine (SVM) [37], which has been proven to have strong performance in many applications, including BCIs [6], is trained for each frequency band of the optimal filter bank with the feature vectors extracted from the spectrally and spatially transformed data. In evaluation, we linearly combine the outputs from multiple classifiers with the weights assigned to each frequency band. That is, given a single-trial EEG $\mathbf{x}^*$, the class label is determined by the following rule:

$$\hat{c} = \underset{c \in \{+,-\}}{\operatorname{argmax}} \left\{ \sum_{k=1}^{|S|} \pi_k \cdot \Phi_k^c(\mathbf{f}_k^*) \right\}, \tag{17}$$

where $|S|$ denotes the size of the optimal filter bank S, $\mathbf{f}_k^*$ denotes the feature vector from the input single-trial EEG $\mathbf{x}^*$, and $\Phi_k^c(\mathbf{f}_k^*)$ is the score of an SVM which classifies the EEG into the class c, in the kth frequency band.

4 Experimental Results and Analysis

In this section, we describe experiments on three public datasets available on the web: the Technische Universität Berlin Dataset [38], BCI Competition III Dataset-IVa [39], and BCI Competition IV Dataset-IIa [40]. We present the effectiveness of the proposed Bayesian Spatiopectral Filter Optimization by comparing its classification results with those of the other competing methods, namely, the standard Common Spatial Pattern (CSP) [7], Common Spatial Spectral Pattern (CSSP) [20], Filter Bank CSP (FBCSP) [19], Discriminative FBCSP (DCSP) [24], and Optimal SpatioSpectral Filter Network (OSSFN) [25]. For the OSSFN, we consider two approaches, each of which utilizes FBCSP (OSSFNwFBCSP) or DCSP (OSSFNwDCSP) in composing a filter bank. In our experiments, the proposed methods as well as the competing methods are trained in a subject-dependent way with trials of a single subject. For all the datasets, the performance of the methods is evaluated by measuring the ratio of trials correctly classified to the total number of test trials for each subject. The source codes for the proposed method is available at http://image.korea.ac.kr/sources.

4.1 Preprocessing and Hyperparameters Setting

The EEG signals of all the datasets were bandpass-filtered between 4 Hz and 40 Hz covering both the μ-rhythm (8-14 Hz) and β-rhythm (14-30 Hz). We then applied the small Laplacian derivation calculated by subtracting four surrounding channels with weights equal to the central one in order to reduce artifacts and noise. All the single-trials were baseline corrected by subtracting the mean of the samples before cue-onset because, during recording, an EEG undergoes slow shifts over time such that the zero level might differ considerably across trials.

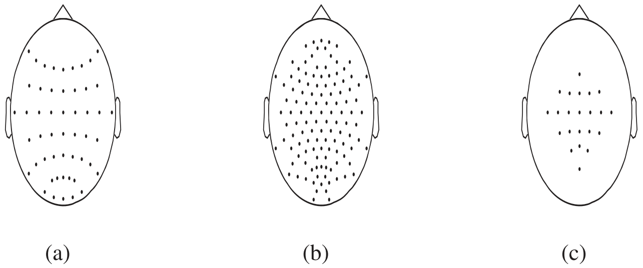


Fig. 2. Electrode montages for the three datasets considered in this paper. (a) Technische Universität Berlin Dataset (51), (b) BCI Competition III Dataset-IV (118). (c) BCI Competition IV Dataset-IIa (22). The numbers in the parentheses denote the number of electrodes.

As stated in Section 3.1, since there is no functional assumption of the prior density we are free to generate particles for the initialization of the BSSFO. In our experiments, we started by generating particles from a mixture of Gaussians $p(\mathbf{B})$ defined as follows based on prior neurophysiological knowledge:

$$p(\mathbf{B}) = \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ denote, respectively, the μ-rhythm and β-rhythm, and the covariances $\boldsymbol{\Sigma}$ are set to be diagonal. Due to the nature of the particle-based *pdf* approximation embedded in the proposed BSSFO, we repeated the experiments 10 times to obtain more statistically robust results. The threshold τ in (16) used for the selection of optimal frequency bands is set to the mean weight of the particles. We use the standard CSP algorithm [7] to learn an optimal spatial filter in each frequency band.

The competing CSP method is applied for the signals bandpass-filtered between 8-30 Hz [7], which is the most commonly selected frequency range in the literature. For the FBCSP and DCSP algorithms, we employed a filter bank composed of nine frequency bands, as described in the original papers [19], [24]. The time delay in CSSP is varied from 1 to 30. For all the competing methods including the proposed BSSFO, a fifth-order Butterworth filter was used to bandpass-filter the signals for any given frequency band. With respect to spatial filter learning, we consider two spatial patterns obtained from the standard CSP algorithm, i.e., the first and last column vectors in the projection matrix.

Although some studies have shown that linear filters are sufficient for sensory motor rhythm-based BCI in their own experiments, there has been also other research that presented better performance with a nonlinear Gaussian SVM than a linear SVM [6], [41], [42], [43]. In our preparatory experiment, we considered both types of classifiers and obtained a slightly better classification accuracy from a Gaussian kernel-based SVM [37]. Therefore, we use a Gaussian kernel-based SVM for all the three datasets.

4.2 Technische Universität Berlin Dataset

4.2.1 Description

During the experiment, a subject was instructed to perform either left-hand or right-hand motor imagery according to the visual cue. The EEGs were recorded using 51 electrodes

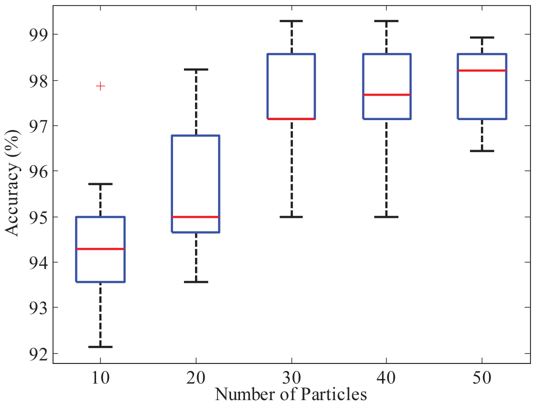


Fig. 3. Performance variation over number of particles—experiment on the Technische Universität Berlin Dataset.

(Fig. 2a) and sampled at 100 Hz. There were 140 trials for each task from a single subject. We considered the samples between 0.75s and 4s after onset for both training and evaluation. In our 10-fold cross-validation, we randomly selected 70 trials per task for training and used the remaining trials for evaluation. It should be noted that in order to ensure a valid comparison among the competing methods, the same training and test partitions were used for evaluation in the cross-validation.

4.2.2 Results

Since the number of particles that is used to estimate the posterior *pdf* in the proposed method can be an important factor affecting the classification performance, we first performed an experiment that involved varying the number of particles. Fig. 3 presents a box plot of the classification results according to the number of particles. We can see that the classification performance increases in accordance with the number of particles. However, the improvement of classification accuracy slows down after 30 particles. In fact, repeated experiments showed that the maximum accuracy was obtained with 30 particles. Therefore, we decided to use 30 particles for the rest of the experiments on not only this dataset but also the other two datasets.

The classification performances are presented in Table 3, from which we can see that the proposed BSSFO outperforms the other six competing methods with a smaller standard deviation. We also illustrate the optimal spatial patterns in Fig. 4 for each competing method. In Fig. 4, the spatial patterns of the proposed method, which are obtained in the frequency band of the largest weight, exhibit the most localized ERD/ERS patterns.

In order to see the effectiveness of the proposed method we computed the distribution of the discriminative frequency band in (3) by calculating probabilities between 4 Hz and 40 Hz at an interval of 0.5 Hz. The distribution is visualized in Fig. 5. In the figure, we also marked the optimized 10 frequency bands, which are selected by (16), with black squares that are distributed around the μ-rhythm.

4.3 BCI Competition III Dataset-IVa

4.3.1 Description

The EEG signals were collected from five healthy subjects ("aa," "al," "av," "aw," "ay"), who were asked to perform left-hand, right-hand, or right-foot motor imaginary, but cues for only the classes of right-hand and right-foot were provided for the competition [39]. The EEG data was acquired using 118 electrodes (Fig. 2b) at the positions given by the extended international 10/20 system and sampled at 250 Hz. We used signals downsampled at 100 Hz. There were 280 trials in total, 140 trials per task, for each subject. The number of trials for training and test varied across subjects as follows: subject(right-hand, right-foot)—aa(80, 88), al(112, 112), av(42, 42), aw(30, 26), and ay(18, 10) were for training and the rest were for evaluation.

4.3.2 Results

The sample points in a time period of 2s, between 0.5s and 2.5s, were considered for both calibration and evaluation. The samples of the first 0.5s period after cue-onset were excluded since these might contain the spontaneous responses to visual stimulus. The classification accuracy of all the competing methods is presented in Table 4, in which the best performance for each subject is highlighted in boldface. The proposed method resulted in the highest classification performance over all subjects with a mean accuracy of 75.46 percent and a standard deviation of 19.06 percent.

The small number of training EEGs for subjects av, aw, and ay consistently resulted in low performance across the methods, except for the proposed BSSFO and CSSP for the subject aw. Based on the web announcement about the results of the competition, the top three winners showed better performance than all the methods considered in this paper, including the proposed BSSFO, did. We believe that the high performance of the winners resulted from the fact that they applied an adaptation or retraining method with an extended training set using data from other subjects to tackle the problem of the small training sets. However, it is out of scope of this paper for methods to combine information from other subjects' training data and would be our forthcoming research issue, extending our method for the problems of small training sets and subject-independent BCIs.

4.4 BCI Competition IV Dataset-IIa

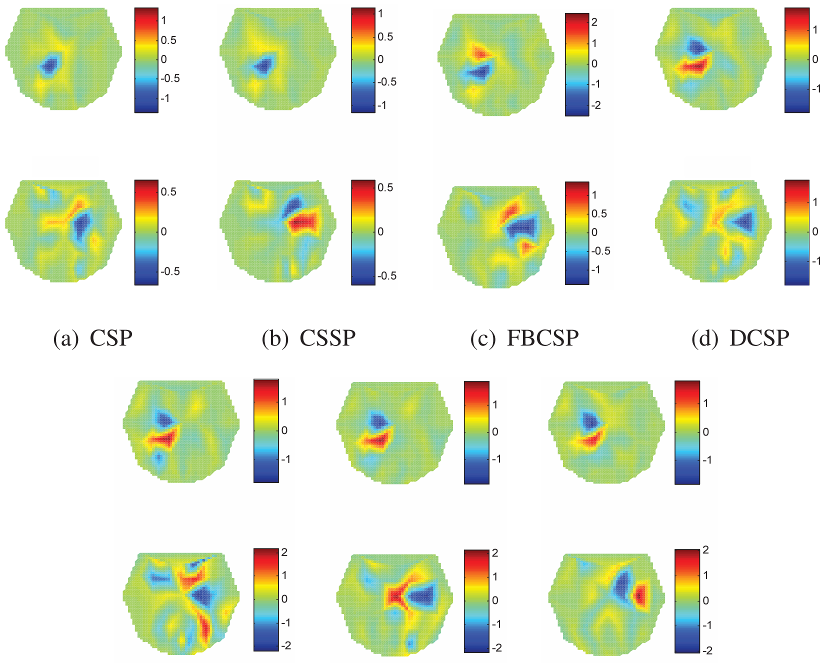
4.4.1 Description

The EEG signals were recorded from nine subjects performing four different motor imagery tasks, i.e., left-hand, right-hand, foot, and tongue, comprised of two sessions conducted

TABLE 3
Classification Performances on the Technische Universität Berlin Dataset

	CSP	CSSP	FBCSP	DCSP	OSSFNwFBCSP	OSSFNwDCSP	BSSFO
Mean (%)	76.79	93.57	75.07	79.36	84.36	84.37	97.57
SD (%)	11.75	2.21	4.86	5.52	5.05	5.06	1.22

SD: Standard Deviation.



(a) CSP (b) CSSP (c) FBCSP (d) DCSP

(e) OSSFNwFBCSP (f) OSSFNwDCSP (g) BSSFO

Fig. 4. The most significant spatial patterns learned from the Technische Universität Berlin dataset in each method. For the BSSFO, these were obtained from the frequency band of the maximum weight.

TABLE 4
Classification Performance on BCI Competition III Dataset IVa

	aa	al	av	aw	ay	Mean (%)	SD (%)
CSP	66.96	89.29	52.55	47.77	52.38	61.79	16.98
CSSP	**79.46** (τ=10)	92.86 (τ=1)	52.55 (τ=1)	91.52 (τ=5)	51.59 (τ=3)	73.60	20.33
FBCSP	69.64	80.36	47.96	55.36	48.41	60.35	14.21
DCSP	69.64	82.14	54.08	50.89	48.41	61.03	14.41
OSSFNwFBCSP	75.00	83.93	53.06	74.11	48.81	66.98	15.22
OSSFNwDCSP	75.00	83.93	52.05	74.11	47.22	66.46	15.93
BSSFO	**79.46**	**94.64**	**57.65**	**91.96**	**53.57**	**75.46**	19.06

τ: a time delay in the sample points, SD: Standard Deviation.

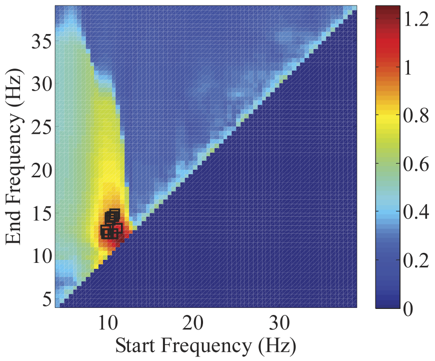


Fig. 5. Visualization of the distribution of the discriminative frequency band for the Technische Universität Berlin dataset and the optimized 10 frequency bands, marked with black squares, obtained in the proposed method. The online color version provides a clearer view.

on different days. Each session includes six runs separated by short breaks, and a run is further composed of 48 trials; there are 12 trials per motor imagery task and 288 trials in total per session [40]. The EEG data were acquired using 22 AG/AgCl electrodes (Fig. 2c) and sampled at 250 Hz. The signals were bandpass-filtered between 0.5 Hz and 100 Hz and an additional 50 Hz notch filter was also applied to suppress line noise.

4.4.2 Results

We consider the signals between 0.5s and 2.5s after onset of the stimulus for both calibration and evaluation. Although there are four different motor-imagery tasks, here we consider motor imagery binary classification: left-hand versus right-hand, left-hand versus foot, left-hand versus tongue, right-hand versus foot, right-hand versus tongue, and foot versus tongue, since the main concern of

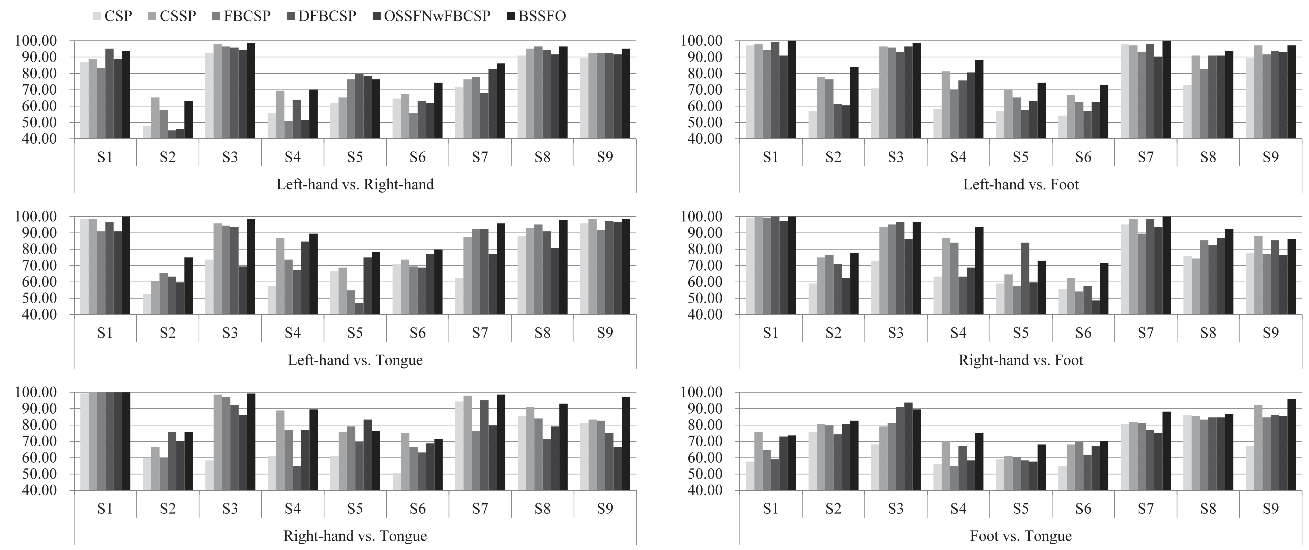


Fig. 6. Binary classification performances on the BCI Competition IV Dataset-IIa.

this paper is discriminative feature extraction by means of spatiospectral filter optimization, not multiclass motor imagery classification.

The results of the six types of motor imagery binary classification are presented in Fig. 6. Although there is high variability in classification performance over subjects, overall the proposed method clearly outperforms the other methods. The mean and standard deviation for each type of binary classification are also presented in Fig. 7. An interesting result from the paired motor imagery classification on this dataset is that the motor imageries of left-hand and foot are the most discriminative, as shown in Table 5. Left-hand versus tongue, right-hand versus tongue, right-hand versus foot, left-hand versus right-hand, and foot versus tongue follow in that order. This result conflicts with the use of the stimuli exploited in the motor imagery-based BCI literature, which mostly considered the left-hand and right-hand tasks.

Technische Universität Berlin Dataset

	Mean (SD)	*p*-value
CSP	76.79 (11.75)	**0.00034**
CSSP	93.57 (2.21)	**0.00082**
FBCSP	75.07 (4.86)	**0.000001**
DCSP	79.36 (5.52)	**0.00001**
OSSFNwFBCSP	84.36 (5.05)	**0.00003**
OSSFNwDCSP	84.37 (5.06)	**0.00003**
Proposed method	97.57 (1.22)	-

BCI Competition III Dataset-IVa

	Mean (SD)	*p*-value
CSP	61.79 (16.98)	0.15652
CSSP	73.60 (20.33)	0.10603
FBCSP	60.35 (14.21)	**0.05319**
DCSP	61.03 (14.41)	0.10300
OSSFNwFBCSP	66.98 (15.22)	**0.03206**
OSSFNwDCSP	66.46 (15.93)	**0.02145**
Proposed method	75.46 (19.06)	-

BCI Competition IV Dataset-IIa (Left-Right)

	Mean (SD)	*p*-value
CSP	73.46 (16.93)	**0.00010**
CSSP	79.78 (13.69)	**0.02741**
FBCSP	76.31 (17.90)	**0.01693**
DCSP	77.55 (18.33)	**0.04450**
OSSFNwFBCSP	76.31 (18.59)	**0.01294**
OSSFNwDCSP	75.62 (18.84)	**0.02071**
Proposed method	83.80 (13.09)	-

BCI Competition IV Dataset-IIa (Left-Foot)

	Mean (SD)	*p*-value
CSP	72.76 (17.93)	**0.00134**
CSSP	86.19 (12.47)	**0.00151**
FBCSP	81.33 (13.18)	**0.00039**
DCSP	80.71 (17.95)	**0.00875**
OSSFNwFBCSP	80.94 (14.18)	**0.00314**
OSSFNwDCSP	84.11 (12.15)	**0.00070**
Proposed method	89.89 (10.71)	-

BCI Competition IV Dataset-IIa (Left-Tongue)

	Mean (SD)	*p*-value
CSP	74.07 (16.57)	**0.00362**
CSSP	84.80 (13.96)	**0.00646**
FBCSP	80.86 (15.17)	**0.00266**
DCSP	79.71 (18.28)	**0.01213**
OSSFNwFBCSP	79.01 (11.02)	**0.00621**
OSSFNwDCSP	78.32 (10.81)	**0.00371**
Proposed method	90.43 (10.03)	-

BCI Competition IV Dataset-IIa (Right-Foot)

	Mean (SD)	*p*-value
CSP	73.07 (15.82)	**0.00136**
CSSP	82.64 (14.09)	**0.03285**
FBCSP	79.86 (15.52)	**0.00405**
DCSP	82.10 (15.41)	0.17589
OSSFNwFBCSP	75.54 (16.71)	**0.00127**
OSSFNwDCSP	78.01 (15.38)	**0.00981**
Proposed method	87.89 (11.29)	-

BCI Competition IV Dataset-IIa (Right-Tongue)

	Mean (SD)	*p*-value
CSP	72.38 (17.83)	**0.00389**
CSSP	86.34 (11.92)	0.16667
FBCSP	80.32 (12.89)	**0.01354**
DCSP	77.47 (15.27)	**0.01943**
OSSFNwFBCSP	79.01 (10.31)	**0.02682**
OSSFNwDCSP	79.55 (12.29)	**0.03136**
Proposed method	89.04 (11.43)	-

BCI Competition IV Dataset-IIa (Foot-Tongue)

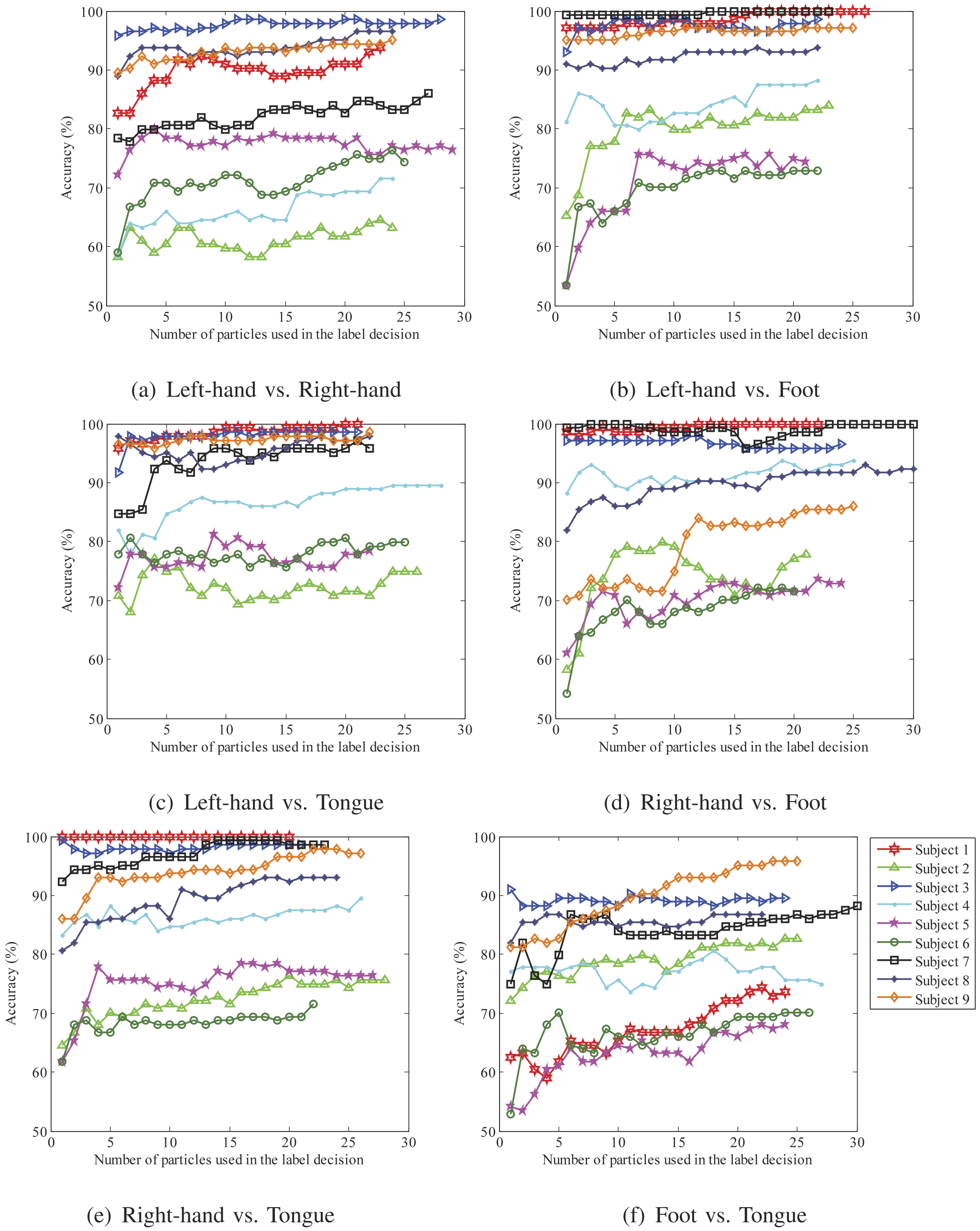
	Mean (SD)	*p*-value
CSP	67.28 (11.40)	**0.00128**
CSSP	77.16 (9.55)	**0.01210**
FBCSP	73.30 (11.17)	**0.00343**
DCSP	73.30 (12.33)	**0.00119**
OSSFNwFBCSP	75.08 (12.39)	**0.03021**
OSSFNwDCSP	76.08 (10.11)	**0.05087**
Proposed method	81.10 (9.73)	-

Fig. 7. Summary of the results for the statistical significance test (paired t-test). The p-values in boldface mean that the null hypotheses can be rejected beyond the 95 percent confidence level. (Left: Left-hand, Right: Right-hand).

TABLE 5
Comparison over Different Types of Motor Imagery Binary Classification

	Left-Right	Left-Foot	Left-Tongue	Right-Foot	Right-Tongue	Foot-Tongue
Mean (%)	77.55	**82.28**	81.03	79.87	80.59	74.76
SD (%)	3.36	5.36	5.23	4.92	5.56	4.24

The numbers are the average classification accuracies of all the competing methods.



(a) Left-hand vs. Right-hand

(b) Left-hand vs. Foot

(c) Left-hand vs. Tongue

(d) Right-hand vs. Foot

(e) Right-hand vs. Tongue

(f) Foot vs. Tongue

Fig. 8. Evolution of the classification accuracies according to the number of particles used in the label decision. The number of particles for each subject is considered up to the size of the filter bank that resulted in the performances presented above.

4.5 Discussion

From the experimental results on the three public datasets presented above, the following questions arise: 1) How statistically significant are the classification performances? and 2) how does the classification performance change according to the number of frequency bands considered in designing the classification system?

4.5.1 Statistical Significance Testing

The null hypothesis in this paper is that the proposed BSSFO method produces the same mean accuracy as the competing methods, e.g., BSSFO = CSP, BSSFO = CSSP, etc. We compute the p-values using a paired t-test to assess whether the differences in classification accuracies between two methods are at a significant level on each dataset. A paired t-test was performed among subjects for both BCI Competition III Dataset IVa and BCI Competition IV Dataset IIa. Meanwhile, for the Technische Universität Berlin Datadet, it was done among folds in cross-validation as if they were from multiple subjects since the dataset was from a single subject.

The summary of results for the statistical significance tests over the datasets is presented in Fig. 7. In the figure, we highlighted the cases in boldface that the null hypothesis can be rejected beyond the 95 percent confidence level. From the statistics shown in Fig. 7, it is clear that the proposed BSSFO significantly outperforms the competing methods: There are seven cases for CSP, six cases for CSSP, eight cases for FBCSP, six cases for DCSP, eight cases for OSSFNwFBCSP, and eight cases for OSSFNwDCSP among the eight binary classification experiments (cases).

4.5.2 Effect of the Number of Frequency Bands

Since we apply a simple threshold-based method in our framework for selection of the optimal number of particles, i.e., frequency bands, the selected particle set can be suboptimal in the sense of classification enhancement. In order to see the effect of the number of frequency bands used in the class label decision, we illustrate the evolution of the classification accuracies according to the number of particles up to the size of a filter bank chosen by the threshold-based method in Fig. 8, where the results are on the BCI Competition IV Dataset-IIa.

Although the classification accuracies vary slightly for some subjects along with the number of particles, our filter bank determined by the threshold-based method showed high performance, as illustrated by the last value for each line in the graph. However, for Subjects 2 and 5, we achieved the same or better performance with a smaller number of particles than the threshold-determined filter bank. That is, it is possible to reduce the computational cost for those subjects in our BCI system and would be an interesting issue for further research.

5 Conclusions and Future Research

In the history of BCI research, one of the revolutionary changes in BCIs may be a paradigm shift with respect to the learning load from the subject to the computer. Meanwhile, machine learning has emerged as the most useful tool for real-life BCIs, helping minimize the amount of subject training time and improve the classification performance. However, there are still two main problems that make it a challenge to classify a single-trial EEG of motor imagery and to prevent the application of BCIs in real-life. First, the frequency bands, in which ERD/ERS patterns reflect sensorimotor activation and deactivation, are highly variable across subjects and across event trials for the same subject. Second, EEG signals are generally contaminated with artifacts and noise that can cause performance degradation in classification.

In this paper, we proposed a novel Bayesian framework to simultaneously optimize spectral filters and spatial filters along with a modified factored-sampling method for *pdf* estimation, an information-theoretic observation model, and a spectrally-weighted decision fusion method. In our experiments on three public databases, the proposed method outperformed the state-of-the-art methods in terms of statistical significance, rejecting the null hypothesis beyond the 95 percent confidence level.

While, in this work we considered only ERD/ERS features for motor imagery classification, there is another well-known neurophysiological feature of slow shifts of the cortical DC potential revealed during imagined or intended movement, BereitschaftsPotential (BP) or readiness potential. Babiloni et al. [44] argued that ERD/ERS and BP represent different aspects of cortical processing and Dornhege et al. [39], [45] presented the improved classification performance with the combination of both aspects of the ERD/ERS and the BP. Inspired by their work, we believe that it would be a meaningful issue to adapt the proposed framework for feature combination.

We would like to note that the proposed method is also applicable to other kinds of single-trial EEG classification problems that are based on modulations of brain rhythms, so it is by no means limited to motor imagery-based BCIs. In order to further increase the discriminability of brain signals, it is necessary to incorporate the problem of task-related electrodes selection into the proposed Bayesian framework. It is also an important issue to combine information from trials of other subjects to overcome the problems of small training sets and subject-independent BCIs.

Acknowledgments

This work was supported by the World Class University (WCU) Program through the National Research Foundation of Korea funded by the Ministry of Education, Science, and Technology (R31-10008). This work was also supported by the National Research Foundation grant funded by the Korea government (MEST) (No. 2012-005741). All correspondence should be directed to S.-W. Lee.

References

[1] G. Pfurtscheller and C. Neuper, "Motor Imagery and Direct Brain-Computer Communication," *Proc. IEEE,* vol. 89, no. 7, pp. 1123-1134, July 2001.

[2] R. Palaniappan and D. Mandic, "Biometric from the Brain Electrical Activity: A Machine Learning Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 4, pp. 738-742, Apr. 2007.

[3] H. Cecotti, "A Self-Paced and Calibration-Less SSVEP-Based Brain-Computer Interface Speller," *IEEE Trans. Neural Systems and Rehabilitation Eng.,* vol. 18, no. 2, pp. 127-133, Apr. 2010.

[4] H. Cecotti and A. Gräser, "Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 3, pp. 433-445, Mar. 2011.

[5] G. Dornhege, J. Millán, T. Hinterberger, D. McFarland, and K.-R. Müller, *Toward Brain-Computer Interfacing.* MIT Press, 2007.

[6] F. Lotte, M. Cngedo, A. Lecuyer, F. Lamarche, and B. Armaldi, "A Review of Classification Algorithm for EEG-Based Brain-Computer Interfaces," *J. Neural Eng.,* vol. 4, no. 2, pp. R1-R13, June 2007.

[7] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal Spatial Filtering of Single Trial EEG during Imagined Hand Movement," *IEEE Trans. Rehabilitation Eng.,* vol. 8, no. 4, pp. 441-446, Dec. 2000.

[8] A. Bashashati, M. Fatourechi, R. Ward, and G. Birch, "A Survey of Signal Processing Algorithms in Brain-Computer Interfaces Based on Electrical Brain Signals," *J. Neural Eng.,* vol. 4, no. 2, pp. R32-57, June 2007.

[9] H.-I. Suk and S.-W. Lee, "A Probabilistic Approach to Spatio-Spectral Filters Optimization in Brain-Computer Interface," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics,* pp. 19-24, Oct. 2011.

[10] M. Dalponte, F. Bovolo, and L. Bruzzone, "Automatic Selection of Frequency and Time Intervals for Classification of EEG Signals," *Electronics Letters,* vol. 43, no. 25, pp. 1406-1408, Dec. 2007.

[11] K. Thomas, C. Guan, L. Tong, and V. Prasad, "An Adaptive Filter Bank for Motor Imagery Based Brain Computer Interface," *Proc. 30th Ann. IEEE Int'l Conf. Eng. in Medicine and Biology Soc.,* pp. 1104-1107, Aug. 2008.

[12] H.-I. Suk and S.-W. Lee, "Subject and Class Specific Frequency Bands Selection for Multi-Class Motor Imagery Classification," *Int'l J. Imaging Systems and Technology,* vol. 21, no. 2, pp. 123-130, June 2011.

[13] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller, "Optimizing Spatial Filters for Robust EEG Single-Trial Analysis," *IEEE Signal Processing Magazine,* vol. 25, no. 1, pp. 44-56, Jan. 2008.

[14] B. Blankertz, M. Kawanabe, R. Tomioka, F. Hohlefeld, V. Nikulin, and K.-R. Müller, "Invariant Common Spatial Patterns: Alleviating Nonstationarities in Brain-Computer Interfacing," *Advances in Neural Information Processing Systems,* vol. 20, pp. 113-120, 2008.

[15] H. Wang and W. Zheng, "Local Temporal Common Spatial Patterns for Robust Single-Trial EEG Classification," *IEEE Trans. Neural Systems and Rehabilitation Eng.,* vol. 16, no. 2, pp. 131-139, Apr. 2008.

[16] F. Lotte and C. Guan, "Regularizing Common Spatial Patterns to Improve BCI Designs: Unified Theory and New Algorithms," *IEEE Trans. Biomedical Eng.,* vol. 58, no. 2, pp. 355-362, Feb. 2011.

[17] M. Krauledat, M. Tangermann, B. Blankertz, and K.-R. Müller, "Towards Zero Training for Brain-Computer Interfacing," *PLoS One,* vol. 3, no. 8, p. e2967, Aug. 2008.

[18] S. Fazli, C. Grozea, M. Danóczy, F. Popescu, B. Blankertz, and K.-R. Müller, "Subject Independent EEG-Based BCI Decoding," *Advances in Neural Information Processing Systems,* vol. 22, pp. 513-521, 2009.

[19] K. Ang, Z. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," *Proc. Int'l Joint Conf. Neural Networks,* pp. 2391-2398, June 2008.

[20] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller, "Spatio-Spectral Filters for Improving the Classification of Single Trial EEG," *IEEE Trans. Biomedical Eng.,* vol. 52, no. 9, pp. 1541-1548, Sept. 2005.

[21] G. Dornhege, B. Blankertz, M. Grauledat, F. Losch, G. Curio, and K.-R. Müller, "Combined Optimization of Spatial and Temporal Filters for Improving Brain-Computer Interfacing," *IEEE Trans. Biomedical Eng.,* vol. 53, no. 11, pp. 2274-2281, Nov. 2006.

[22] R. Tomioka and K.-R. Müller, "A Regularized Discriminative Framework for EEG Analysis with Application to Brain-Computer Interface," *NeuroImage,* vol. 49, no. 1, pp. 415-432, 2010.

[23] W. Wu, X. Gau, B. Hong, and S. Gau, "Classifying Sing-Trial EEG During Motor Imagery by Iterative Spatio-Spectral Patterns Learning (ISSPL)," *IEEE Trans. Biomedical Eng.,* vol. 55, no. 6, pp. 1733-1743, June 2008.

[24] K. Thomas, C. Guan, C. Lau, A. Vinod, and K. Ang, "A New Discriminative Common Spatial Pattern Method for Motor Imagery Brain-Computer Interfaces," *IEEE Trans. Biomedical Eng.,* vol. 56, no. 11, pp. 2730-2733, Nov. 2009.

[25] H. Zhang, Z. Chin, K. Ang, C. Guan, and C. Wang, "Optimum Spatio-Spectral Filtering Network for Brain-Computer Interface," *IEEE Trans. Neural Network,* vol. 22, no. 1, pp. 52-63, Jan. 2011.

[26] L. Haiping, E. How-Lung, G. Cuntai, K. Plataniotis, and A. Venetsanopoulos, "Regularized Common Spatial Pattern with Aggregation for EEG Classification in Small-Sample Setting," *IEEE Trans. Biomedical Eng.,* vol. 57, no. 12, pp. 2936-2946, Dec. 2010.

[27] U. Grenander, Y. Chow, and K. Keenan, *A Pattern Theoretical Study of Biological Shapes.* Springer-Verlag, New York, 1991.

[28] C. Andrieu, "An Introduction to MCMC for Machine Learning," *Machine Learning,* vol. 50, no. 1, pp. 5-43, Jan. 2003.

[29] A. Doucet, N. De Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice.* Springer, 2001.

[30] M. Isard and A. Blake, "CONDENSATION-Conditional Density Propagation for Visual Tracking," *Int'l J. Computer Vision,* vol. 29, no. 1, pp. 5-28, 1998.

[31] T. Lan, E. Erdogmus, A. Adami, M. Pavel, and S. Mathan, "Salient EEG Channel Selection in Brain Computer Interfaces by Mutual Information Maximization," *Proc. 27th Ann. IEEE Int'l Conf. Eng. in Medicine and Biology Soc.,* pp. 7064-7067, Sept. 2006.

[32] F. Oveisi and A. Erfanian, "A Minimax Mutual Information Scheme for Supervised Feature Extraction and Its Application to EEG-Based Brain-Computer Interfacing," *EURASIP J. Advances in Signal Processing,* vol. 2008, pp. 1-8, Jan. 2008.

[33] N. Kwak and C.-H. Choi, "Input Feature Selection by Mutual Information Based on Parzen Window," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 12, pp. 1667-1671, Dec. 2002.

[34] H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 8, pp. 1226-1238, Aug. 2005.

[35] J. Leiva-Murillo and A. Artés-Rodríguez, "Maximization of Mutual Information for Supervised Linear Feature Extraction," *IEEE Trans. Neural Networks,* vol. 18, no. 5, pp. 1433-1441, Sept. 2007.

[36] E. Parzen, "On Estimation of a Probability Density Function and Mode," *Annals of Math. Statistics,* vol. 33, pp. 1065-1076, Sept. 1962.

[37] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery,* vol. 2, no. 2, pp. 121-167, June 1998.

[38] https://wiki.ml.tu-berlin.de/wiki/Main/SS09_AnalysisOfNeuronalData, 2012.

[39] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Boosting Bit Rates in Non-Invasive EEG Single-Trial Classifications by Feature Combination and Multi-Class Paradigms," *IEEE Trans. Biomedical Eng.,* vol. 51, no. 6, pp. 993-1002, June 2004.

[40] C. Brunner, R. Leeb, G. Müller-Putz, A. Schlógl, and G. Pfurtscheller, "BCI Competition 2008—Graz Data Set a," Laboratory of Brain-Computer Interfaces, Inst. for Knowledge Discovery, Graz Univ. of Technology, http://www.bbci.de/competition/iv/#dataset2a, 2008.

[41] D. Garrett, D. Peterson, C. Anderson, and M. Thaut, "Comparison of Linear, Nonlinear, and Feature Selection Methods for EEG Signal Classification," *IEEE Trans. Neural Systems and Rehabilitation Eng.,* vol. 11, no. 2, pp. 141-144, June 2003.

[42] G. Garcia, T. Ebrahimi, and J.-M. Vesin, "Support Vector EEG Classification in the Fourier and Time-Frequency Correlation Domains," *Proc. First IEEE Int'l EMBS Conf. Neural Eng.,* pp. 591-594, Mar. 2003.

[43] K.-R. Müller, M. Krauledat, G. Dornhege, G. Curio, and B. Blankertz, "Machine Learning Techniques for Brain-Computer Interfaces," *Biomedical Technology,* vol. 49, pp. 11-22, Dec. 2004.

[44] C. Babiloni, F. Garducci, F. Cincotti, P. Rossini, C. Neuper, G. Pfurtscheller, and F. Babiloni, "Human Movement-Related Potentials versus Desynchronization of EEG Alpha Rhythm: A High-Resolution EEG Study," *NeuroImage,* vol. 10, no. 6, pp. 658-665, Dec. 1999.

[45] G. Dornhege, B. Blankertz, G. Curio, and K.-R. Müller, "Combining Features for BCI," *Advances in Neural Information Processing Systems,* vol. 15, pp. 1115-1122, 2003.

Heung-Il Suk received the BS and MS degrees in computer engineering from Pukyong National University, Busan, Korea, in 2004 and 2007, respectively, and the PhD degree in computer science and engineering from the Korea University, Seoul, in 2012. He is currently a postdoctoral research associate in the Department of Radiology and the Biomedical Research Imaging Center (BRIC) at the University of North Carolina, Chapel Hill. From 2004 to 2005, he was a visiting researcher in the Computer and Vision Research Center at the University of Texas, Austin. His current research interests include machine learning, computer vision, brain-computer interfaces, and neuroimaging analysis. He received the Outstanding Paper Award at the Korea Computer Congress in 2007. He also received the Silver Award at the 18th Samsung Human-Tech Thesis Prize in 2012. He is a student member of the IEEE.

Seong-Whan Lee received the BS degree in computer science and statistics from Seoul National University, Korea, in 1984, and the MS and PhD degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Seoul, Korea, in 1986 and 1989, respectively. Currently, he is working as the Hyundai-Kia Motor Chair Professor at Korea University, Seoul, where he is the head of the Department of Brain and Cognitive Engineering and the director of the Institute for Brain and Cognitive Engineering. He is the principal investigator of the World Class University (WCU) project on “Brain and Cognitive Engineering” research, which is funded by the Ministry of Education, Science and Technology of Korea. His current research interests include pattern recognition, computer vision, and brain informatics. He has more than 250 publications in international journals and conference proceedings, and has authored 10 books. He was the winner of the Annual Best Student Paper Award of the Korea Information Science Society in 1986. He received the First Outstanding Young Researcher Award at the Second International Conference on Document Analysis and Recognition in 1993, and the First Distinguished Research Award from Chungbuk National University in 1994. He received the Outstanding Research Award from the Korea Information Science Society in 1996. He received the Lotfi Zadeh Best Paper Award at the IEEE International Conference on Machine Learning and Cybernetics in 2011. He also received the Scientist of the Month Award from the Ministry of Education, Science and Technology of Korea in 2012. He was the founding co-editor-in-chief of the*International Journal of Document Analysis and Recognition*. He has been an associate editor of several international journals, including *Pattern Recognition*, *ACM Transactions on Applied Perception*, *IEEE Transactions on Affective Computing, Image and Vision Computing*, *International Journal of Pattern Recognition and Artificial Intelligence*, and *International Journal of Image and Graphics*. He was a general or program chair of many international conferences and workshops and was also on the program committees of numerous conferences and workshops. He is a fellow of the IEEE, IAPR, and Korean Academy of Science and Technology; he has served several professional societies as a chairman or governing board member.