31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Spatio-spectral feature selection based on robust mutual information estimate for brain computer interfaces

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

***Abstract*— This paper addresses the issue of selecting optimal spatio-spectral features, which is key to high performance motor imagery (MI) classification that is in turn one of the central topics in EEG-based brain computer interfaces. In particular, this work proposes a novel method which first formulates the selection of features as maximizing mutual information between class labels and features. It then uses a robust estimate of mutual information, within a filter-bank and common spatial pattern feature extraction framework, to select an effective feature set. We have assessed the proposed method on both BCI Competition IV Set I and a separate data set collected in our lab from 7 healthy subjects. The results indicate the method is effective in selecting optimal spatial-spectral features for classification.**

I. INTRODUCTION

Selection of optimal spatio-spectral features constitutes the kernel issue in motor imagery classification for EEG-based brain computer interfaces (BCIs) [1], [2], [3], which allow a user to interact with the environment through motor activity in the brain alone, without using muscular output channels.

The discriminative spatio-spectral characteristics of motor imagery EEG vary, considerably, from one person to another. Thus, motor imagery classification relies on selecting an effective set of EEG features in the vast spatio-spectral feature space, while practically only a limited number of examples (say, less than 100 trials) for an individual are available.

The prevalent technique for extracting discriminant spatial features is the *common spatial pattern* (CSP) method [4], [5]. From recorded motor imagery data, it constructs spatial filters which basically maximize the variance for one class while at the same time minimize it for the other one. The CSP usually works on bandpass-filtered EEG, i.e. a specific rhythm, and captures a strong or attenuated rhythmic activity, linking to ERD/ERS effects of motor imagery [2].

Simultaneous optimization of spatial filters and spectral filters is needed for selecting effective features in the joint spatio-spectral space. In [6], an algorithm was proposed to optimize simple frequency filters (one tap delay) together with spatial filters. Later on, an extension, termed CSSSP, enabled simultaneous optimization of an arbitrary FIR filter within the CSP analysis [7]. More recently, Wu et al. proposed an algorithm termed ISSPL which directly optimizes spectral filters (FIR in spectral domain) to achieve maximal classification accuracy [8].

In [9], an alternative method, termed FBCSP (filter-bank common spatial pattern), introduced a combination of filter banks and CSP. It first processed EEG waveforms with an array of band-pass filters (filter banks), and used CSP to construct spatial filters for each frequency band. Importantly, the method used a maximal mutual information criterion to select a feature set, which is in turn classified using a naïve-Bayesian Parzen window algorithm. That method has been validated in the BCI Competition IV, where it served as the basis for the winning algorithms in all three EEG categories (see http://ida.first.fhg.de/projects/bci/competition_iv/results/index.html).

This paper presents an extension to FBCSP. It formulates feature extraction&selection as maximizing mutual information between class labels and features. Importantly, it introduces a robust estimate of mutual information to the FBCSP framework, as well as an efficient feature selection algorithm. To validate the method, we have tested it on both BCI Competition IV Set I (human data sets only) and a separate data set from 7 healthy subjects collected in our lab. The results indicate the method is effective in selecting optimal spatial-spectral feature sets. And, it also consistently outperforms the FBCSP method in terms of classification accuracy on the test data sets.

It's worthwhile to mention that a separate article [10] also proposed a mutual information approach (termed ITFE) for extracting features for motor imagery classification. That differs from this work in two aspects: first, it does not address the issue of selecting spectral filters; second, its approximation models were validated in multi-class settings only, while for two class paradigms, the authors showed that ITFE would approximate CSP. In contrast, this paper not only addresses spectral filters together with spatial filters, but also demonstrates that the proposed method yields superior performance to both CSP and FBCSP on the test data sets (i.e. unseen data from the training phase).

II. MAXIMUM MUTUAL INFORMATION FORMULATION FOR FEATURE SELECTION

Let's denote a trial of motor imagery in multi-channel EEG by $\mathbf{x}(t) = \{x_1(t), x_2(t), \ldots, x_L(t)\}$ where L is the number of channels. According to [9], it is first processed by an array of band-pass filters (i.e. filter banks) to extract specific rhythm activities. For each filter bank (say, the m-th bank), a CSP is constructed that consists of N_w spatial filters $\mathbf{w}_{mn}$,

This work was supported by the Science and Engineering Research Council of A*STAR (Agency for Science, Technology and Research). H. Zhang, K. K. Ang, C. Guan, C. Wang are with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #21-01 Connexis, Singapore 138632. (email: {hhzhang, kkang, ctguan, ccwang}@i2r.a-star.edu.sg).

$n = 1, \ldots, N_w$, each linearly combines the filtered multi-channel EEG into a new waveform.

Consider the waveform generated by the m-th band-pass filter (h_m) and the associated n-th spatial filter ($\mathbf{w}_{mn}$). We define its feature (a_{mn}) as the logarithmic mean power.

$$a_{mn} = \phi(\mathbf{x}, h_m, \mathbf{w}_{mn}) = \log\left[\frac{1}{T}\int_{t_0}^{t_0+T}\left[\mathbf{w}_{mn}^T(\mathbf{x}\otimes h_m)\right]^2 dt\right] \quad (1)$$

Then the whole feature set $\{a_{mn}\}$ is comprised of $M \times N$ features from the M band-pass filters and the $M \times N$ spatial filters. The problem becomes how to select an effective subset of $\{a_{mn}\}$ to represent the discriminative information between motor imagery classes.

To address this problem, we consider that the features from a feature subset η constitute a random vector A_η. The corresponding class label (discrete value from 1 to N_c) is a random uni-variate denoted by C. A particular feature vector and the class label are denoted respectively by $\mathbf{a}_\eta$ and c. From information theory, the mutual information [11] between A_η and C is

$$I(A_\eta, C) = H(A_\eta) - H(A_\eta|C) \quad (2)$$

where $H(A_\eta)$ denotes the entropy of the random feature vector, and $H(A_\eta|C)$ is the conditional entropy

$$\begin{aligned} H(A_\eta|C) &= -\sum_{c=1}^{N_c}\int_{\mathbf{a}} p(\mathbf{a}_\eta, c)\log(p(\mathbf{a}_\eta|c))d\mathbf{a}_\eta \\ &= -\sum_{c=1}^{N_c} H(A_\eta|c)P(c) \end{aligned} \quad (3)$$

The mutual information is a quantity that measures the mutual dependence of the two variables. And maximum mutual information (MMI) criterion has been established as the basis for discriminative learning procedures in various machine learning techniques such as hidden Markov models machines. Following the same principle, we define the objective of selecting the optimal set of features as below.

$$\max_{\eta} I(A_\eta|C) \quad (4)$$

There are two problems towards achieving the objective: 1. how to robustly estimate the mutual information given a limited set of samples, 2. how to efficiently select the optimal feature subset from a large number of possibilities (theoretically $\sum_{k=1}^{MN}\binom{MN}{k}$). In the next section we will propose a solution to address the two problems.

III. Feature selection algorithm with robust estimate of mutual information

The objective above involves joint probability density functions (PDFs) that need to be estimated from a given training data set. Note: to simplify the following elaboration, we hereafter omit the subset symbol η in the expressions unless otherwise specified. Using kernel density estimator, we have the PDF of a random feature vector

$$p(\mathbf{a}) = \frac{1}{N}\sum_{i=1}^{N}\varphi(\mathbf{a} - \mathbf{a}_i) \quad (5)$$

where φ is a smoothing kernel, and $\mathbf{a}_i$ is a given sample of the random vector. A Gaussian kernel is used.

$$\varphi(\mathbf{t}) = (2\pi)^{-\frac{n}{2}}|\psi|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{t}^T\psi^{-1}\mathbf{t}\right) \quad (6)$$

where ψ is the covariance matrix.

We adopt a method proposed in [12] to approximate the entropy $H(A)$ with a given set of samples. From the definition of entropy, it can be viewed as an expectation

$$\begin{aligned} H(A) &= -\int_{\mathbf{a}} p(\mathbf{a})\log(p(\mathbf{a}))d\mathbf{a} \\ &= -E[\log(p(\mathbf{a}))] \\ &\cong -\frac{1}{N}\sum_{i=1}^{N}\log(p(\mathbf{a}_i)) \end{aligned} \quad (7)$$

Combining the above equations, the entropy of the random vector A can be approximated by

$$H(A) = -\frac{1}{N}\sum_{i=1}^{N}\log\left\{\frac{1}{N}\sum_{j=1}^{N}\varphi[\mathbf{a}_i - \mathbf{a}_j]\right\} \quad (8)$$

Similarly, the within-class entropy $H(A|c)$ can also be easily estimated by using samples from the class c only.

Now we consider how to use this estimate to efficiently select a feature subset for classification.

As in conventional CSP, it's reasonable to assume that most discriminant features are associated with a relevant rhythm (i.e. a user-specific frequency band.) Thus, we consider only those feature subsets whose elements are from a single frequency band. Besides, since the top and bottom few components from a CSP projection contain the most discriminant information, the other features are discarded for feature selection. To further reduce computational cost in this study, we select the feature subsets with two elements only. In other words, the selected feature subset is a pair of CSP features from a single frequency band.

Hence our algorithm for determining the optimal feature subset can be expressed in the following pseudo-code.

---

1) Process raw EEG by filter banks and respective CSPs and obtain features $\{a_{mn}\}$, where only the top 2 and the bottom 2 CSP components are used for each filter bank;
2) For each filter bank k, $k = 1, \ldots, K$;
   a) For each pair (η) of CSP features: $\eta \in \mathcal{N}$ where $\mathcal{N}$ is the set of all possible combination of choosing 2 features from the 4 CSP features; thus the size of $\mathcal{N}$ is $\binom{4}{2} = 6$;
      i) Compute the entropy $H(A_\eta)$;
      ii) Compute the within-class entropy $H(A_\eta|c)$ for each class: $c = 1, \ldots, N_c$;
      iii) Compute the mutual information $I(A_\eta|C)$ according to Eq. 2 and Eq. 3;
   b) Select the optimal subset for the k-th filter bank: $\eta^{(k)} = \text{argmax}_{\eta\in\mathcal{N}} I(A_\eta|C)$;

3) Select the optimal subset for all filter banks: $\eta^{\text{opt}} = \text{argmax}_{\eta^{(k)}} I(A_{\eta^{(k)}}|C)$

IV. Experiments

A. Description of BCI Competition Dataset I

Data set 1 from BCI Competition IV were used for assessing the proposed method. A full description can be found at http://ida.first.fhg.de/projects/bci/ competition_iv/desc_1.html. Here is a brief introduction. The data were recorded from 4 human subjects and 3 artificial ones. During the data collection, motor imagery was performed without feedback. Each subject chose two classes of motor imagery from left hand, right hand, and foot (side chosen by the subject; optionally also both feet.) In each motor imagery trial, visual cues were displayed for a period of 4s during which the subject was instructed to perform the cued motor imagery task. These periods were interleaved with 2s of blank screen and 2s with a fixation cross shown in the center of the screen. The EEG contains 59 channels.

In this study, we use the calibration data sets from the 4 real human subjects, where each subject contributed 200 trials that are evenly distributed over the two chosen classes. And to facilitate the computation, our experiment uses the 100Hz version of the data.

B. Description of Our Own Dataset

A separate data set was collected in our lab, involving 7 healthy, male and young subjects. Each subject contributed 160 motor imagery trials that were evenly distributed over two classes: left-hand motor imagery and right right motor imagery. At the beginning of each trial, a fixation symbol appeared on the screen that prompted the subject to prepare; after 2 seconds, a random visual cue appeared to indicate either left-hand or right-hand; 4 seconds later, a stop signal presented to marked the end of the trial. These trials were interleaved with 6 seconds of break. EEG was acquired with a Neuroscan amplifier working at a sampling rate of 250Hz. And 25 EEG channels surrounding the motor cortex areas were selected, without channel-selection tuning for individual subjects.

C. Evaluation Setting

In this study, we used the same setting for both datasets.

- **Time interval** (i.e. the time segment of a motor imagery trial for analysis). It starts at 1 second and ends at 4 seconds after the cue. The starting point was set as such to remove the effect of spontaneous responses (evoked potentials) to the cue.
- **Frequency ranges** for filter banks. The center frequency of the banks linearly span the range from 5.5Hz to 29.5 Hz, with a fixed bandwidth of 1.5Hz. The filters are all Chebyshev Type II.
- **Channels**. No channel selection was performed. For the BCI Competition dataset, all the 59 channels were used. For our own dataset, all the 25 channels were used.

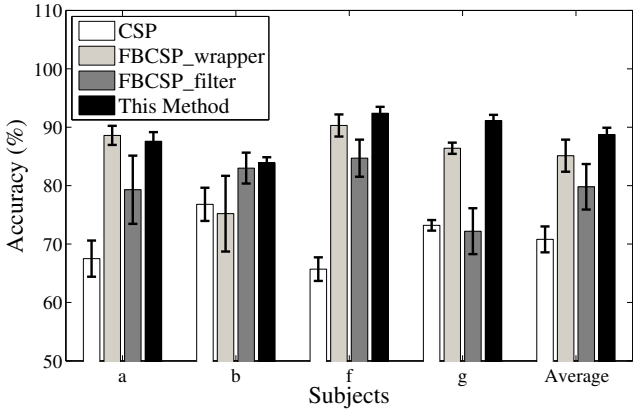

Fig. 1. Accuracy on BCI Competition IV Set I (with Naïve Bayesian Parzen window classifier)

- **Classifiers**. We tested two widely-used classifiers: Naïve-Bayesian Parzen window classifier, support vector machines (SVMs). The former was recommended as in prior studies on FBCSP. The SVMs library was provided by the bioinformatics toolbox in Matlab.
- **Methods under comparison**. In this study, we also evaluated other three methods on the same datasets and setting. The methods include
  - CSP. Conventional CSP with a broad passband from 6Hz to 28Hz (with a Chebyshev Type II filter).
  - Filter-bank CSP using a wrapper approach. See [9]. Its short name is **FBCSP_wrapper** in the graphs.
  - Filter-bank CSP using a filter approach. See [9]. Its short name is **FBCSP_filter** in the graphs.
- **Number of Features**. Only 2 features are selected for each of the methods under comparison.
- **Evaluation of classification accuracy**. A randomized 5x5fold cross validation was applied to compare the methods in terms of classification accuracy.

D. Results and Discussions

Figure 1 and 2 plot the statistics of classification accuracy for each method on individual subject as well as on average for each dataset. In particular, each bar represents an averaged (over 5 by 5 folds) classification accuracy rate, while the small bar on top of it denotes the standard deviation of the accuracy rate over the folds.

The results indicate that, with the Naïve-Bayesian Parzen window classifier, all feature extraction methods except for the conventional CSP could yield an accuracy rate close to or above 80% on the competition dataset, or above 75% on our dataset. Compared with FBCSP methods, the proposed method yielded a significant 3.6% boost and 1.5% boost in accuracy on the two datasets. Importantly, the more prominent improvement lies in the variance: the STD of the accuracy rate is approximately halved by the proposed method.

Figure 3 and 4 illustrate the comparative classification results with the SVM classifier. The results are nearly identical to that with the Naïve-Bayesian Parzen window classifier.

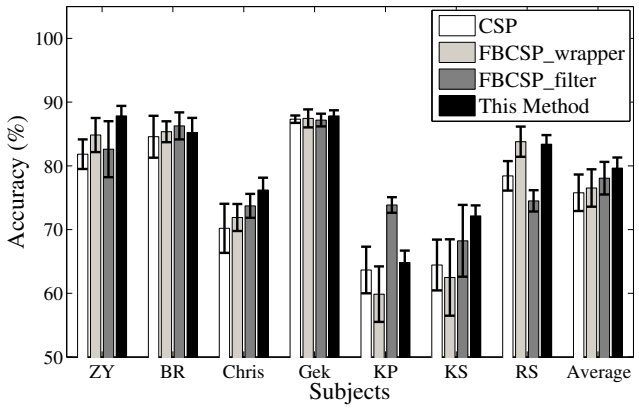

Fig. 2. Accuracy on Our 7 subjects (with Naïve Bayesian Parzen window classifier)

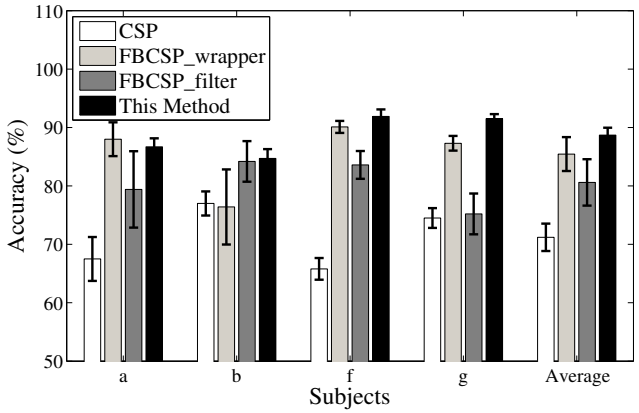

Fig. 3. Accuracy on BCI Competition IV Set I (with SVM classifier)

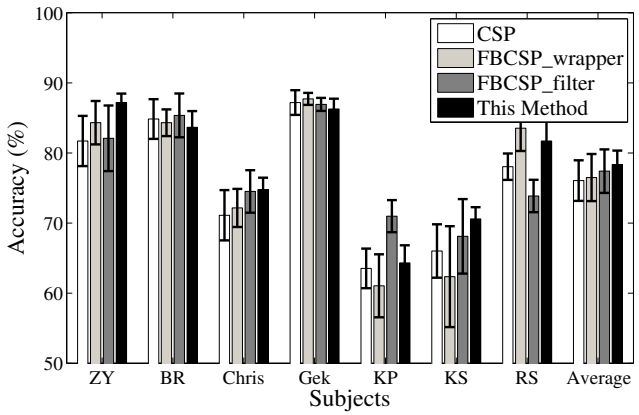

Fig. 4. Accuracy on Our 7 subjects (with SVM classifier)

The proposed method outperformed the other three by 3% on the competition dataset and by 1% on our dataset, all on the test sets in cross-validation. Again, the prominent improvement lies in the STD of the accuracy rate. This indicates that the new method clearly stands out in terms of robustness against changes in training-test sets.

Besides, it is evident that all the three feature extraction methods are not very sensitive to the choice between the two classifiers.

V. Conclusions

We have presented an efficient feature extraction method for motor imagery classification. The method has been validated with a total of 11 subjects from 2 independent datasets. It's noteworthy to emphasize that the method is a fully autonomous learning and classification mechanism. And this work demonstrates that, without painstaking manual-tunning of the parameters, the autonomous technique can select effective discriminative features for high performance motor imagery BCIs.

References

[1] J.R. Wolpaw, N. Birbaumer, D.J. MacFarland, G. Pfurtscheller, and T.M. Vaughan. Brain-computer interface for communication and control. *Clinical Neurophyiology*, 113:767–791, 2002.

[2] G. Pfurtscheller, C. Neuper, D. Flotzinger, and M. Pregenzer. Eeg-based discrimnation between imagination of right and left hand movement. *Clinical Neurophysiology*, 103:642–651, 1997.

[3] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Müller. Optimizing spatial filters for robust eeg single-trial analysis.

[4] J. Muller-Gerking, G. Pfurtscheller, and H. Flyvbjerg. Designing optimal spatial filtering of single trial EEG classification in a movement task. *Clinical Neurophysiology*, 110:787–798, 1999.

[5] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Transactions on Rehabilitation Engineering*, 8(4):441–446, 2000.

[6] S. Lemm, B. Blankertz, G. Curio, and K.-R. Müller. Spatio-spectral filters for robust classification of single trial EEG. *IEEE Transactions on Biomedical Engineering*, 52(9):1541, 2005.

[7] Guido Dornhege, Benjamin Blankertz, Matthias Krauledat, Florian Losch, Gabriel Curio, and Klaus-Robert Mller. Combined optimization of spatial and temporal filters for improving brain-computer interfacing. *IEEE Transactions on Biomedical Engineering*, 53(11):2274–2281, 2006.

[8] Wei Wu, Xiaorong Gao, Bo Hong, and Shangkai Gao. Classifying single-trial eeg during motor imagery by iterative spatio-spectral patterns learning (isspl). *IEEE Transactions on Biomedical Engineering*, 55(6):1733–1743, 2008.

[9] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. In *International Joint Conference on Neural Networks (IJCNN2008)*, pages 2391–2398, 2008.

[10] M. Grosse-Wentrup and M. Buss. Multiclass common spatial patterns and information theoretic feature extraction. 55(8):1991–2000, 2008.

[11] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. 2 edition, 1984.

[12] P. Viola and W.M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24:2, 1997.