Combining Feature Extraction and Classification for fNIRS BCIs by Regularized Least Squares Optimization

Dominic Heger, Christian Herff, and Tanja Schultz

Abstract— **In this paper, we show that multiple operations of the typical pattern recognition chain of an fNIRS-based BCI, including feature extraction and classification, can be unified by solving a convex optimization problem. We formulate a regularized least squares problem that learns a single affine transformation of raw HbO_2 and HbR signals. We show that this transformation can achieve competitive results in an fNIRS BCI classification task, as it significantly improves recognition of different levels of workload over previously published results on a publicly available n-back data set. Furthermore, we visualize the learned models and analyze their spatio-temporal characteristics.**

I. INTRODUCTION

Functional Near-Infrared Spectroscopy (fNIRS) is an non-invasive brain imaging technology that has been shown to be suitable for many applications in brain-computer interfacing and user state detection. However, studies focusing on single-trial analysis of fNIRS signals are still rare and common standards for signal processing, feature extraction and classification are still not established.

The core challenge in brain-computer interfacing is to extract informative features from the raw brain signal time series and to learn models to predict mental states or intentions of the user that can be used for communication and control. Therefore, BCIs typically follow a pattern recognition approach, in which multiple operations are sequentially applied to the raw signal data to derive recognition estimates. In such a processing chain, data are usually pre-processed to remove noise and artifacts, subsequently, the signals are temporally and spatially filtered, informative features are extracted, selected and combined. Finally, these features are transmitted to a machine learning component to predict unseen data based on previously learned models. For fNIRS-based BCIs, there is no single standard feature extraction method. Commonly, studies convert the raw optical measurements into relative estimates for oxygenated (HbO_2) and de-oxygenated (HbR) hemoglobin concentration using the modified Beer Lambert Law (MBLL) [1]. Features are calculated from these chromophore concentrations to represent informative properties of the hemodynamics. Typically, simple statistical properties of the time-domain signal amplitudes, such as mean, variance, slope, kurtosis, skewness or laterality features have been calculated as features [2], [3], [4], [5]. Other studies have also used frequency domain signals and wavelet decomposition [6], [5]. After feature extraction, machine learning methods, such as linear or quadratic discriminant analysis, support vector machines, or neural networks are used to model and predict unseen data [7], [6], [8], [9], [5].

Previous work has shown that multiple operations of a BCI pattern recognition chain can be unified into a single transformation of the raw signal data, which can be formulated as an optimization problem (e.g. [10], [11], [12]).

This approach is has the benefit that it performs a global optimization, instead of multiple intermediate steps that may have different explicit and implicit assumptions. Furthermore, it is usually difficult to find optimal parameters for multiple individual operation steps, in particular, as they may depend on each other. The global optimization approach requires no expert knowledge to identify relevant activity, whereby it allows to analyze and verify neurophysiological plausibility (see section III-A).

In the general case it may not be possible to solve such optimization problems efficiently, however for some problems the solution can be analytically determined, for example CSPs can be calculate by solving a generalized eigenvalue problem [13]. Furthermore, the family of convex optimization problems can be efficiently solved and there exist multiple generic numerical solvers, such as CVX [14] and more specialized solving algorithms, such as DAL [15], L1 General [16], ADMM [17] and others.

Convex optimization is especially relevant for BCI problems, since linear processing methods, such as frequency filters, spatial filters, signal transformations, and classifiers, are widely used in BCI research and approaches using linear methods only are among those that have achieved the most competitive performance in many BCI tasks (see e.g. [18]). In the context of EEG-based BCIs, Tomioka and Müller [12] have introduced a regularized discriminative framework using convex optimization that combines operations such as feature extraction, feature selection, feature combination, and classification in a single optimization process. They used logistic regression as a predictor function and evaluated different methods for regularization. They evaluated their approach for a P300 speller and self-paced finger tapping.

In this paper, we present a first approach showing that a single transformation of raw HbO_2 and HbR signals can be learned by solving a convex optimization problem that unifies multiple operations of a traditional pattern recognition processing chain. We show that our approach can achieve competitive results in an fNIRS BCI classification task as it significantly improves classification above previously published results on our publicly available n-back data set [4]. Additionally, we visualize the learned models and analyze spatio-temporal characteristics that are

All authors are with the Cognitive Systems Lab, Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Adenauerring 4, 76131 Karlsruhe, Germany. dominic.heger@kit.edu

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

optimally discriminative under least squares assumptions. Our approach is especially interesting for BCI problems where assumptions should be as small as possible, or little is known about the discriminative effects in the data. To the best of our knowledge convex optimization for solving feature extraction, feature selection and classification has not been applied in fNIRS-based BCIs before.

II. MATERIAL AND METHODS

A. Experiment Description and Data Corpus

We used the publicly available *n*-back data corpus that has recently been recorded at the Cognitive Systems Lab [4]. In this data set subjects performed the *n*-back task, in which a sequence of rapidly appearing stimuli is visually presented. Subjects respond by button press if the current stimulus is identical to the one that has been shown *n* stimuli before. This task requires substantial attention and memory processing and induces higher workload for increasing *n*.

The data corpus consist of fNIRS signals from 10 subjects (4 females, 8 right-handed). All subjects performed 10 trial blocks of three *n*-back tasks $n \in \{1,2,3\}$ in pseudo-randomized order. During each trial block 22 stimuli with 3 ± 1 targets were visually presented for 2 seconds each. Relaxation periods were recorded after each trial block to ensure that hemoglobin levels returned to baseline. The data corpus, additionally, contains RELAX periods an a break in the middle of the experiment. Only the 30 *n*-back trial blocks were used for the analyzes in this paper. fNIRS signals were recorded using an Oxymon Mk III (Artinis Medical Systems, Netherlands). Four transmitter and four receiver optodes were attached to the forehead with a source-detector distance of 3.5cm (8 measurement locations). The available data set contains the concentration changes of HbO_2 and HbR for each channel sampled at 25Hz. A detailed description of the recording setup, experiment design, subjective evaluations and behavioral results can be found in [4].

B. Signal Processing

Signal pre-processing was performed similar to [4]: A moving average filter was applied to attenuate slow signal trends and low frequency effects (e.g. Mayer Waves), which subtracted the mean of the 120 seconds before and after every sample from each HbO_2 and HbR sample. Interferences from heartbeat and higher frequency signal components were attenuated using an elliptic IIR low-pass filter with cutoff frequency 0.5Hz (filter order 6). Trial blocks were extracted from the filtered signals using different window lengths.

C. Optimization Problem

The problem of predicting the target variable from a small chunk of fNIRS signal data can be formulated as the following regularized least squares optimization problem (i.e. regression problem) [19], which estimates an affine transformation of the data that minimizes the squared distance to the corresponding target variable:

$$\underset{w,b}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - w^\top \cdot vec(X_i) + b)^2 + \lambda \|w\|, \qquad (1)$$

where n is the number of trials, $y_i \in \{-1,1\}$ is the target variable indicating the ground truth class label of trial i, $vec(X_i)$ is the vectorized data of trial i, i.e. a vector containing all HbO_2 and HbR samples of all channels and $b \in \mathbb{R}$ is a bias term.

The model vector w can be high dimensional, for example, for 16 data channels and a trial length of 40 seconds sampled by 25 Hz there are 16000 coefficients in w. Nonetheless, models can be robustly estimated from small amounts of data as the second term in (1) controls for model complexity by regularization. The regularization weight λ that penalizes the optimization by the norm $\|w\|$ can be chosen to avoid overfitting due to small sample size. Multiple penalty terms for regularization, in particular sparsity inducing norms, are used in machine learning literature and may be applied within this framework. For the evaluations in this paper we applied ℓ_1-norm regularization (aka. lasso), which is well known for learning sparse models (i.e. many coefficients in w have values near zero). Predictions $\hat{y}$ from unseen chunks of fNRIS data X can be made by $\hat{y} = sgn(w^\top \cdot vec(X) + b)$.

The optimization problem (1) has only small constraints and simultaneously optimizes temporal and spatial weights of the least squares predictor function in a purely data driven way, which allows to consider the whole spatiotemporal structure of the data. For the evaluations in this paper we used CVX [14] to solve the optimization problem.

D. Baseline Data Processing

For the evaluation of our approach we consider the results presented in [4] as the baseline performance. In [4] feature extraction was performed by calculating the slopes of the HbO_2 and HbR signals of each trial, which results in a 16 dimensional feature vector. Subsequently, the 8 most relevant features were selected using Mutual Information based feature selection. A Linear Discriminant Analysis (LDA) was employed for classification.

III. EVALUATION AND RESULTS

To evaluate the approach proposed in this paper, we discriminated different *n*-back conditions corresponding three different levels of workload against each other. Comparable to our previous article [4], we evaluated three binary classification tasks (**1-2**, **1-3**, **2-3**) and one three class task (**1-2-3**) for different window lengths.

10-fold cross-validations have been performed to calculate estimates of classification accuracy. Regularization weights $\lambda \in \{i^2/100 | i = 0,1,\ldots,10\}$ were chosen using a nested 5-fold cross-validation on the training data of each fold.

Figures 1 and 2 show classification accuracies averaged over all subjects for different window lengths, for the 2 class and 3 class tasks, respectively. Dashed lines show the baseline from [4] for comparison. Note that in this analysis, the number of instances is dependent on the window length, i.e. with increasing window size fewer instances are available (see [4] for more details).

The recognition results of the proposed approach are in general superior to the baseline results, with the exception

of results for task **1-2** and the result for 25 seconds window length of **1-3**. Pairwise one-sided t-tests indicated that the improvements in accuracy are significant for the results of the tasks **1-3**, **2-3** and **1-2-3** ($p < 0.05$, $p < 0.0005$, $p < 0.01$, respectively). The performance of the task **1-2** was not significantly lower than the baseline results (pairwise one-sided t-test, $p > 0.08$).

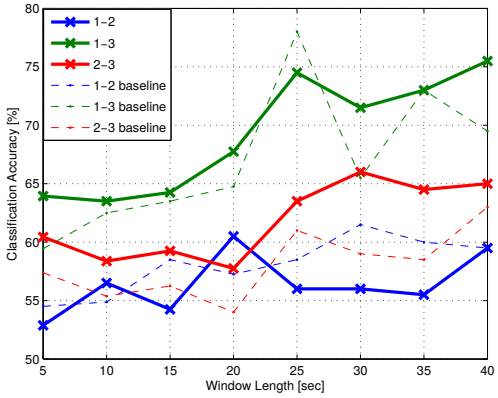

Fig. 1. Classification accuracies for the tasks **1-2**, **1-3**, and **2-3** using different window lengths. Dashed lines indicate baseline results, solid lines results of the proposed approach.

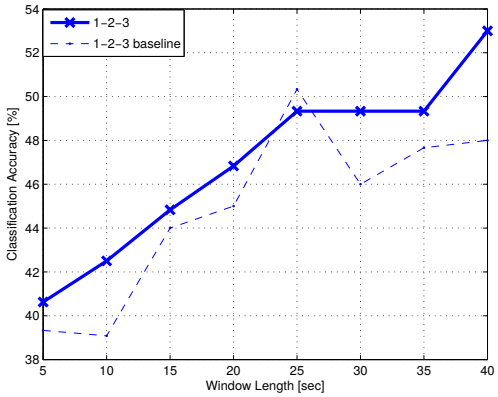

Fig. 2. Classification accuracies for the task **1-2-3** using different window lengths. Dashed lines indicate baseline results, solid lines results of the proposed approach.

During the optimization of (1), the number of relevant coefficients in the model w is automatically determined by regularization. In most cases a non-zero ℓ_1-regularization parameter λ was chosen and the number of active weight coefficients (i.e. absolute value larger than 10^{-5}) in the weight vector w was between 11 and 20 (mean 16.5, standard deviation 2.4). For the models where $\lambda = 0$ has been chosen, more than 99% of the coefficients in w were active weights. For long windows, classification results of the proposed approach appear to be more stable than for the baseline, which can be explained by the regularization that controls for the complexity of the classifier and avoids overfitting when only few instances are available.

Instead of the ℓ_1-norm regularization penalty that was used for the results presented above, we also calculated results for the ℓ_2-norm (i.e. equivalent to Frobenius norm; producing overall small coefficients in w) regularization penalty. Results were not significantly better than the baseline and performance was significantly lower for nearly all window sizes and tasks than ℓ_1-norm results of the proposed approach.

A. Model Analysis

Using optimization problem (1), discriminative models (w,b) are learned purely data driven from chunks of the raw HbO_2 and HbR time series. Investigating the spatiotemporal structure of the learned models may give insights on how predictions are made by the model and indicate how differences between different levels of workload are represented in fNIRS data. A direct interpretation of the discriminative (backward) model weight vectors as learned in (1) can be misleading, since, e.g. task irrelevant activity can have significant weights in w in order to filter out such activity. Haufe et al. [20] recently proposed a method to convert a linear backward model w into a forward model a (generative model), which allows interpretation of the weight vector:

$$a = \Sigma_X \cdot w \cdot \Sigma^{-1}_{w^\top X}, \tag{2}$$

where Σ_X is the covariance matrix of the zero mean signal data X and $\Sigma_{w^\top X}$ is the covariance matrix of the predicted data.

Figure 3 shows the forward models for the task **1-3** for a window length of 40 seconds. The models illustrated in figure 3 were calculated from the complete 1-back and 3-back data in the data set. Visual inspection showed that the models trained in the evaluations (section III) during the 10 folds of the cross-validation are very similar and did not show systematic differences to those in figure 3. We chose the task **1-3** as it produced the highest recognition accuracies and therefore has the most validity to show physiological effects. Similar conclusions can be drawn from other tasks, which are not shown here due to page limitations.

Each row of figure 3 shows a plot of the 8 HbO_2 (left column) and a plot of the 8 HbR (right column) channels over time. Comparing the plots along the columns indicates strong inter-subject variabilities, which can be attributed to individual differences in brain activity and measurement variations, such as small differences in measurement location due differences in brain anatomy and optode locations.

The forward models of this data set are subject to noise, but temporal and laterality effects can be recognized. For multiple subjects, the models of most channels show negative activations (blue) in the first half and positive activations (red) in the second half of the HbO_2 trials, or positive activations (red) in the first half and negative activations (blue) in the second half of the HbR trials. Such temporal effects are for example pronounced for subjects 2, 4, 6 in the HbO_2 channels and for subjects 5, 6, 7, 10 in HbR channels. The effect can be interpreted in the light of a typical shape of a hemodynamic response that increases for HbO_2 and decrease for HbR signals around 10 seconds after the stimulus onset, whereby the effect is most pronounced in the 3-back condition (c.f. Figure 4 in [4] for grand averages of the task conditions).

Differences between the activations of channels 1-4 (right hemisphere) and channels 5-8 (left hemisphere) indicate laterality effects. Figure 3 shows that laterality effects are

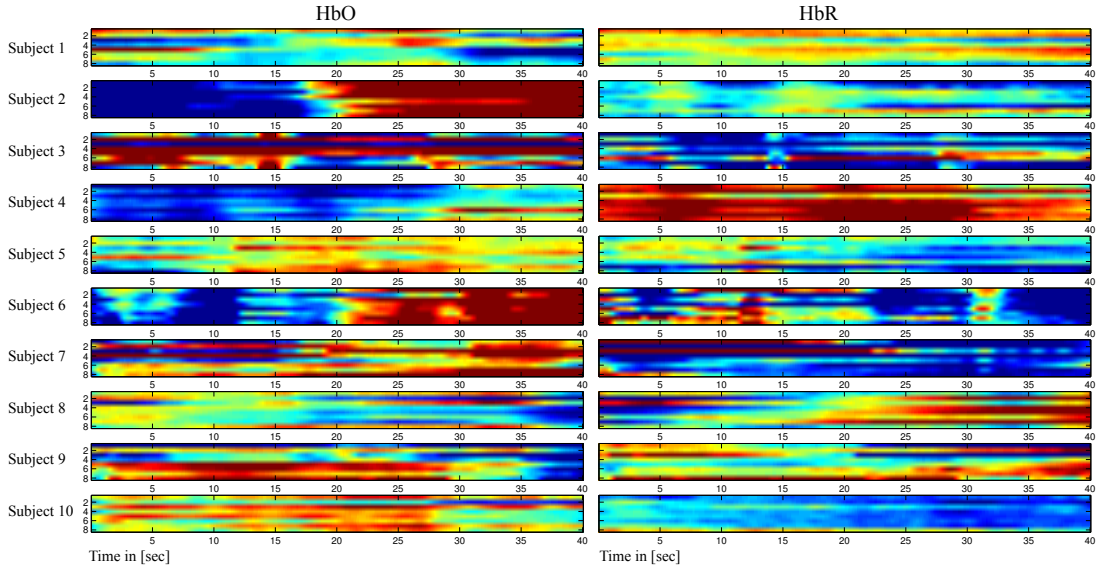

Fig. 3. Visualization of the forward models for the condition 1-back versus 3-back (**1-3**), 40 seconds window length.

more subject to temporal variability, but can be observed, for example, for subject 9 in HbO_2 channels and for subjects 7 in HbO_2 and HbR channels.

IV. CONCLUSIONS

In this paper we have analyzed fNIRS signals recoded during 1, 2, and 3-back task conditions using least-squares optimal discriminative models learned by solving convex optimization problems. This approach has a straight forward problem formulation, and can efficiently be solved using numerical solvers. It is elegant as it unifies multiple operations in a typical BCI pattern recognition chain, including feature extraction, spatiotemporal filtering and classification. We have shown that the proposed approach significantly and consistently improved classification performance for the tasks **1-3**, **2-3**, and **1-2-3**. The learned models were analyzed using interpretable forward models and spatial and temporal effects were discussed. We believe that the techniques described in this paper are helpful for many problems in fNIRS-based brain computer interfacing and beyond.

REFERENCES

[1] A. Sassaroli and S. Fantini, "Comment on the modified beerlambert law for scattering media," *Physics in Medicine and Biology*, vol. 49, no. 14, 2004.

[2] K. Tai and T. Chau, "Single-trial classification of nirs signals during emotional induction tasks: towards a corporeal machine interface," *Journal of neuroengineering and rehabilitation*, vol. 6, p. 39, 2009.

[3] S. Moghimi, A. Kushki, S. Power, A. M. Guerguerian, and T. Chau, "Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy," *Journal of neural engineering*, vol. 9, no. 2, p. 026022, 2012.

[4] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back task quantified in the prefrontal cortex using fnirs," *Frontiers in human neuroscience*, vol. 7, 2013.

[5] D. Heger, C. Herff, F. Putze, R. Mutter, and T. Schultz, "Continuous affective states recognition using functional near infrared spectroscopy," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 113–125, 2014.

[6] T. Q. D. Khoa and M. Nakagawa, "Functional near infrared spectroscope for cognition brain tasks by wavelets analysis and neural networks," *Int J Biol Med Sci*, vol. 1, pp. 28–33, 2008.

[7] R. Sitaram, H. Zhang, C. Guan, M. Thulasidas, Y. Hoshi, A. Ishikawa, K. Shimizu, and N. Birbaumer, "Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface," *NeuroImage*, vol. 34, no. 4, pp. 1416–1427, 2007.

[8] S. D. Power, A. Kushki, and T. Chau, "Towards a system-paced near-infrared spectroscopy brain–computer interface: differentiating prefrontal activity due to mental arithmetic and mental singing from the no-control state," *Journal of neural engineering*, vol. 8, no. 6, p. 066004, 2011.

[9] K. S. Rahnuma, A. Wahab, N. Kamaruddin, and H. Majid, "Eeg analysis for understanding stress based on affective model basis function," in *Consumer Electronics (ISCE), 2011 IEEE 15th International Symposium on*. IEEE, 2011, pp. 592–597.

[10] J. Farquhar, J. Hill, and B. Schölkopf, in *NIPS 2006 Workshop on Current Trends Brain-Computer Interfacing, Whistler, Canada*, 2006.

[11] L. C. Parra, C. Christoforou, A. D. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. G. Philiastides, and P. Sajda, "Spatiotemporal linear decoding of brain state," *Signal Processing Magazine, IEEE*, vol. 25, no. 1, pp. 107–115, 2008.

[12] 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.

[13] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K.-R. Muller, "Optimizing spatial filters for robust eeg single-trial analysis," *Signal Processing Magazine, IEEE*, vol. 25, no. 1, pp. 41–56, 2008.

[14] M. Grant, S. Boyd, and Y. Ye, "Cvx: Matlab software for disciplined convex programming," 2008.

[15] R. Tomioka and M. Sugiyama, "Dual-augmented lagrangian method for efficient sparse reconstruction," *Signal Processing Letters, IEEE*, vol. 16, no. 12, pp. 1067–1070, 2009.

[16] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for l1 regularization: A comparative study and two new approaches," in *Machine Learning: ECML 2007*. Springer, 2007, pp. 286–297.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[18] K. Muller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 11, no. 2, pp. 165–169, 2003.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[20] S. Haufe, F. Meinecke, K. Görgen, S. Dähne, J.-D. Haynes, B. Blankertz, and F. Bießmann, "On the interpretation of weight vectors of linear models in multivariate neuroimaging," *NeuroImage*, vol. 87, pp. 96–110, 2014.