STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Billiet L., Hunyadi B., Matic V., Van Huffel S., Verleysen M, De Vos M. (2015), **Single Trial Classification in Mobile BCI: A Multiway Kernel Approach** 8th International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS2015), 5-11.
Archived version	Final publisher's version / pdf
Published version	NA
Journal homepage	http://www.biosignals.biostec.org/?y=2015
Author contact	lieven.billiet@esat.kuleuven.be *+ 32 (0)16 327685*
IR	NA

(article begins on next page)

Single Trial Classification for Mobile BCI: a Multiway Kernel Approach

Lieven Billiet[1,2], Borbála Hunyadi[1,2], Vladimir Matic[1,2], Sabine Van Huffel[1,2], Michel Verleysen[3] and Maarten De Vos[4,5]

[1]*STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, Department of Electrical Engineering, KU Leuven, Leuven, Belgium*

[2]*iMinds Medical Information Technologies, Belgium*

[3]*Institute for Information and Communication Technologies, Electronics and Applied Mathematics, Universit catholique de Louvain, Louvain-la-Neuve, Belgium*

[4]*Methods in Neurocognitive Psychology Lab, Department of Psychology, Cluster of excellence Hearing4all, Carl von Ossietzky University, Oldenburg, Germany*

[5]*Research Center Neurosensory Science, Carl von Ossietzky University, Oldenburg, Germany*

lieven.billiet@esat.kuleuven.be

Keywords: single-trial ERP BCI, mobile BCI, tensors, subspaces.

Abstract: Subspace methods have been applied in various application fields to obtain robust results. Using multilinear algebra, they can also be applied on structured tensorial data. This work combines this principle with the power of non-linear kernels to investigate its merits in single trial classification for a mobile BCI ERP classification task. The accuracy difference with regard to more conventional vector kernels is evaluated for sitting and walking condition, increasing training data set and averaging over multiple trials. The study concludes that in general, the tensorial approach does not yield any advantage, though it might for specific subjects.

1 INTRODUCTION

BCI is a multidisciplinary field aiming at using brain activity to drive applications. It has various possible uses, particularly as assistive technology (e.g. for people suffering from neural or muscle degradation, locked-in syndrome). It led to the development of brain spellers, robot, wheelchair or prosthesis control and even coma detection (Jackson and Mappus, 2010). One way to steer such systems is a synchronised reactive approach: the user is presented with stimuli and his selective attention for a particular stimulus can be discovered as a positive Event Related Potential (ERP) approx. 300ms after stimulus onset (P300).

Brain activity can be measured in various ways: using cortical electrodes, with fMRI etc. However, many systems aim at non-invasiveness and *mobility*. Therefore, wireless EEG is a suitable method. Mobile EEG systems are already available commercially, several of which offer wireless communication. Some have been shown to perform as good as more classical systems. The system of Debener et al., used in this work, has been subject to several tests for ERP detection (De Vos et al., 2013). This study aims to confirm the system's practical usefulness.

The brain response on a single stimulus is called a *single trial*. Most systems average over several trials to obtain a higher signal-to-noise ratio. Yet, when single trials can be classified accurately, the system can be controlled faster, an advantage for real-time use. For this purpose, mostly relatively simple features are extracted from EEG data and concatenated in a vector, after which a classifier can be trained. However, these features often disregard structural information of the data. Representing EEG data as tensors allows a multilinear comparison based on signal subspaces rather than classical feature values. For example, Onishi et al. use a tensorial expansion and dimensionality reduction to extract an informative feature vector (Onishi et al., 2012). Other studies focus on regularization using the nuclear norm of a tensor (Hunyadi et al., 2013). Tensor discriminant analysis has been used to obtain informative subspaces from wavelets (Li and Zhang, 2010) and with Gabor features on motor imagery tasks and EEG seizure detection (Nasehi and Pourghassem, 2011).

Tensorial approaches can be combined with SVM

classification by kernels for structured information (Zhao et al., 2013). Signoretto et al. (Signoretto, 2011) developed a tensorial kernel. They apply it among others for classification of MEG signals in a BCI task.

In this work, several vectorial and tensorial representations for single-trial BCI data are compared, with a focus on the influence of training set size and the interference of motor activity when walking. Finally, though the main focus is on single trials, averaging over multiple trials is studied as well, since the effectiveness of a BCI system is often a trade-off between fast classification (less averaging) and high accuracy (more averaging). This trade-off is expected to be different for different methods.

2 DATA AND METHODS

2.1 Data

The dataset for this study was acquired for the study of De Vos et al. (De Vos et al., 2013). 20 subjects were asked to participate in a three stimulus auditory oddball paradigm (Halder et al., 2010). Subjects were presented with a train of standard tones (900Hz) interspersed with rare target and distractor (deviant) tones (600Hz and 1200Hz, randomly assigned) of 62ms duration with a mean interstimulus interval of 1000ms. A focus on the target tone leads to a more pronounced P300 compared to the deviant, which allows to detect binary choices. The recordings were performed in two conditions, walking and sitting, both outdoor, to estimate the impact of motor activity. Subjects were equipped with the wireless EEG system developed by Debener et al. (Debener et al., 2012). It combines a 14 channel sintered Ag-AgCl electrode cap based on the international 10-20 system with a light-weight amplifier. The system is presented in Figure 1. It has a sample frequency of 128Hz.

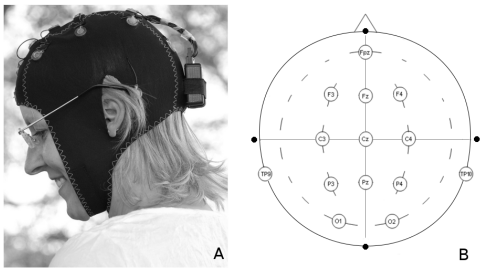


Figure 1: Acquisition system (De Vos et al., 2013) (A) with electrode positions (B)

2.2 Data Processing

The data processing steps are shown in Figure 2. First, eye blinks were removed semi-automatically with ICA. Since the P300 response is time-locked to a stimulus, epochs could subsequently be extracted from the resulting signal. Each recorded trial (epoch) starts at 200ms before stimulus onset and ends at 800ms after it. The leading 200ms was used for baseline correction. Furthermore, high frequency noise was removed with a 20Hz low-pass filter. Next, the P300 response was be made more prominent by rereferencing using the average of the Tp9 and Tp10 channel. One last additional data-driven FIR filtering step was performed by truncation of the SVD of a channel's hankel matrix to its three most prominent components (Hansen and Jensen, 1998). The resulting data set consists of 94 target and 94 deviant 12-channel single trials for each subject.

2.3 Classification Methods

The data is classified with Least Squares SVM (LS-SVM), an SVM variant solving the training optimization as a system of linear equations (Suykens and Vandewalle, 1999). Three types of kernel functions were used: a linear kernel, an RBF kernel and a tensorial kernel (Signoretto, 2011). The tensorial kernel is a factor kernel with a factor for each dimension ('mode') of the tensorial inputs, given as:

$$K(\mathcal{X},\mathcal{Y}) = \prod_i e^{\frac{d(\mathcal{X}^{(i)},\mathcal{Y}^{(i)})^2}{2\sigma^2}} \qquad (1)$$

$d(\mathcal{X}^{(i)},\mathcal{Y}^{(i)})$ is a distance function between the mode-i tensors unfoldings (De Lathauwer et al., 2000). The kernel is equivalent to an RBF kernel with the euclidean distance replaced by a sum of distances $d()^2$, one for each mode. The distance function is a subspace measure which can be calculated using the SVD as:

$$d(\mathcal{X}^{(i)},\mathcal{Y}^{(i)}) = \|\mathbf{V}_{\mathcal{X}^{(i)}} \cdot \mathbf{V}_{\mathcal{X}^{(i)}}^T - \mathbf{V}_{\mathcal{Y}^{(i)}} \cdot \mathbf{V}_{\mathcal{Y}^{(i)}}^T\| \qquad (2)$$

$\mathbf{V}$ are the right singular vectors of the tensor unfolding. The distance is called a projection Frobenius norm (or chordal distance). The procedure is summarized in Figure 3

2.4 Data Representations

2.4.1 Vector Representations

A fast way to obtain robust and informative features is by averaging over the informative part (0-800ms)

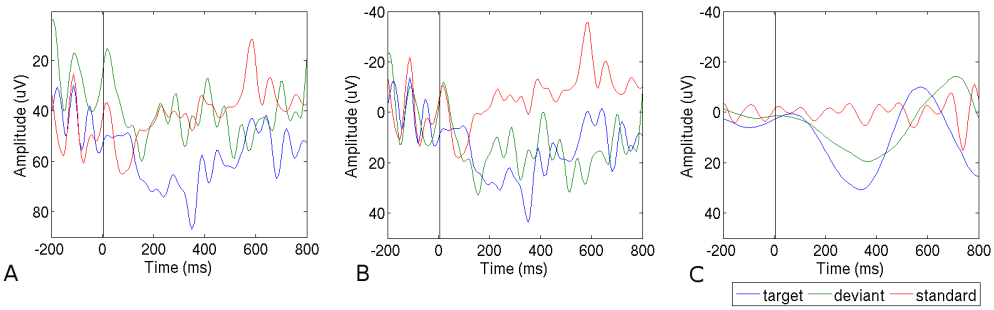


Figure 2: Data acquisition steps for a single trial: after ICA (A), filtered and baseline corrected (B), re-referenced and SVD truncated (C)

of the time signals. For this study, each channel has been divided in 23 bins of 31ms. The binned data can be presented as a channel-time matrix, as in the left part of Figure 4. For the linear and RBF kernels, the data is vectorized. The dimensionality of the problem is reduced to 60 subject-dependent features using reliefF (Liu and Motoda, 2008). In Figure 4, the importance of selected features is given as a gray value. Note the most important lie around t = 300ms (bin 9-11). These binning values are used with both the linear (BINLIN) and RBF (BINRBF) kernel.

As another type of features, the right side of Figure 4 displays examples of Gabor atoms (Gabor, 1946), defined as gaussian windowed cosines (for real signals), with a certain frequency f, time shift t_0 and window parameter σ. A vectorial Gabor atom approach starts from a dictionary of generated atoms. Here, 5696 atoms have been used with $t_0 \in \{1+2\cdot i | i = 1..63\}$ [samples], $f \in \{\frac{\pi}{8} i | i = 0..26\}$ [Hz] and $\sigma \in \{2\cdot i | i = 1:6\}$. The dictionary is used to approximate signals as linear combinations of atoms by means of orthogonal matching pursuit (Mallat and Zhang, 1993). The combination coefficients form the feature vector. A channel can be approximated almost perfectly with 4 or 5 atoms. Feature vectors are therefore sparse and, additionally, non-consistent across trials. One way to cope with that is to select representative atoms based on the grand average (the average across all single trials). For each channel, atoms are selected for both target and non-target grand averages. These sets can be unified, leading to one representative atom set for each channel. Eventually, single trial channels are fitted to these sets in a least-squares sense and concatenated to form a single trial vector. A last optimization is a further feature selection with reliefF, yielding a final vector of 30 features. The use of these features with RBF will be referred to as GABRBF.

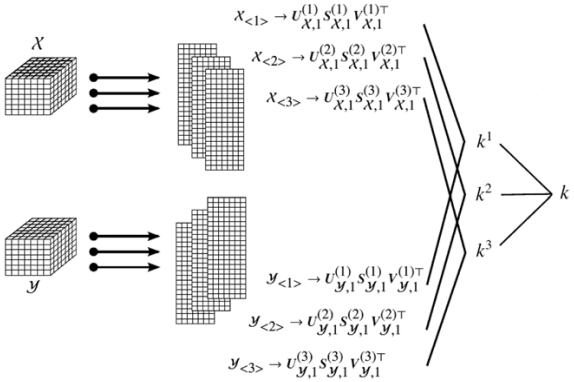


Figure 3: Graphical representation of the tensorial kernel (Signoretto, 2011)

2.4.2 Tensor Representations

Three kinds of three-way tensor representations are tested for use with the tensorial kernel. Example slices of these tensors for the grand average differences of target and deviant can be seen in Figure 5. The Hankel tensor consists of channel hankel matrices, stacked together to form a third dimension. As Figure 5A shows, a hankel matrix is an anti-diagonal matrix whose entries can be defined by a signal laid out over the first column and last row. Its use is motivated by subspace identification for linear time invariant systems, closely linked to (damped) harmonic

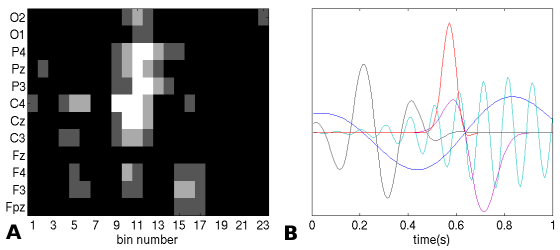


Figure 4: The bin-channel map with features 1-10 in white, 11-25 in gray and 26-60 in dark grey (A) example of Gabor atoms (B)

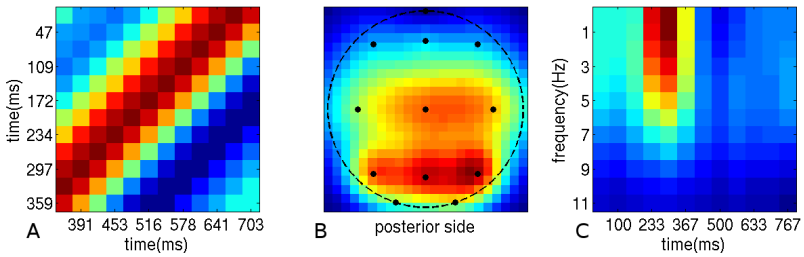


Figure 5: Hankel Pz slice (A) Topographic slice around 300ms (B) Pz Time-frequency map (C)

retrieval (Kung et al., 1983). Since the 23 bins are used to construct square matrices, the eventual tensors have dimensions $12 \times 12 \times 12$, two time and one channel dimension. Using the hankel tensor with the tensorial kernel is denoted by HANK.

A second data representation is topographic (TOPO). For each of the 23 time bins, a square topographic map of the brain is generated. Such a map is a 2D gaussian mixture model based on the electrode positions and their voltages for the given time bin. The maps are stacked to form a $23 \times 23 \times 23$ tensor, two spatial and one time dimension.

Finally, a time-frequency tensor is constructed. It is closely related to the Gabor transform (Wexler and Raz, 1990), which will here be considered as a windowed STFT. For each channel, a 12×12 matrix is constructed, corresponding to an equal division of time (0-800ms) and frequency (0-11Hz). The width of the gaussian window is optimized for each subject. Similar data representations have been used in BCI, though mostly for motion imagery. Using this representation for classification is indicated with GABT.

2.5 Experiments

Three kinds of experiments will be discussed, highlighting the influence of condition (sitting vs walking), the training set size and averaging.

3 RESULTS

3.1 Influence of Condition

As mentioned before, data was recorded in sitting and walking condition to estimate the influence of motor activity. A general comparison of the methods across all subjects and their accuracy changes due to the recording condition is given in Figure 6. Based on these average results, the influence of condition is clear: there is a significant decrease in accuracy from sitting to walking ($p < 0.01$, paired t-test). For individual subjects, 18 out of 20 have all methods non-increasing (either significantly decreasing or no significance).

Conclusions about a comparison between the methods can be drawn as well. In sitting condition, the following order can be established ($p < 0.05$): BINLIN $>$ BINRBF $>$ TOPO $>$ HANK, GABT $>$ GABRBF. The difference between HANK and GABT is not significant. BINLIN and GABRBF are least affected by the condition change (-4%), whereas GABRBF, HANK and TOPO lose around 6% and GABT even 8%. The ordering of the methods remains almost the same, only GABRBF and GABT switch place ($p < 0.05$). BINLIN strengthens its position.

3.2 Influence of Training Set Size

Figure 7 sketches the influence of the training set size for two subjects. Although there is a big difference in the actual values for the two subjects, all methods increase in performance with a growing set, which is only logical: a better model can be derived. It should be remarked that the linear method proves superior in all cases, followed by BINRBF. For the second subject however, TOPO is seen to be better for increasing size of the data set.

3.3 Influence of Averaging

The average over all subjects (left part of Figure 8) shows that BINLIN and BINRBF have an almost parallel increase of accuracy when averaging. They increase more than all other methods. TOPO and HANK also follow this, though less clearly. All meth-

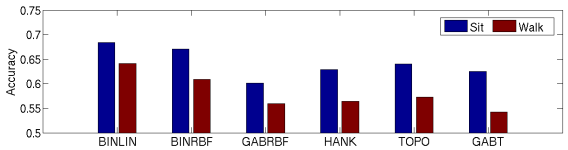


Figure 6: Accuracies for all methods for sitting and walking condition

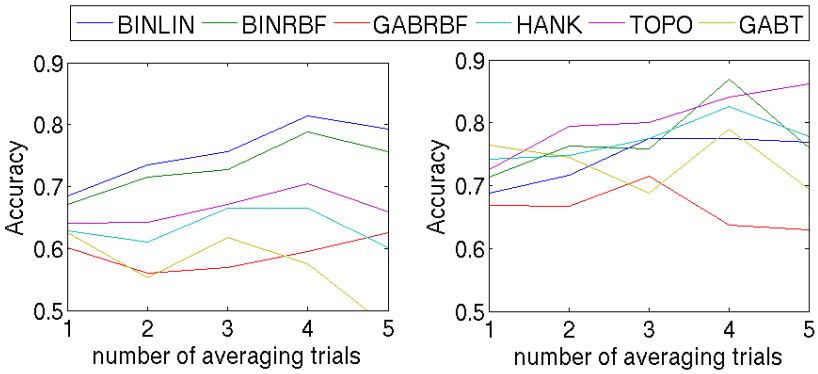


Figure 8: Influence of averaging over 1-5 trials: across subjects (left), for one specific subject (right)

ods except for GABRBF show an eventual decrease when using more trials for averaging, particularly clear for GABT. This will probably be due to the decrease in training set size, a direct result of the averaging process. Yet, this is not true for GABRBF. Since it is built using a model derived from the grand average, trials can be more accurately modelled when averaging is applied as it yields a lower variance. The right part of Figure 8 shows an interesting insight. TOPO is the only tensorial method which frequently can keep pace with or outperforms BINLIN and/or BINRBF. In the given example, HANK performs better than BINLIN as well. This is an example where it is already the case for single trials, averaging is therefore not more beneficial to vectorial methods. On the other hand, in most cases where TOPO becomes competitive, it performs worse on single trials. Therefore, for some subjects, the benefit of averaging is higher for TOPO than for the vectorial methods, which cannot be derived from the left part of the Figure.

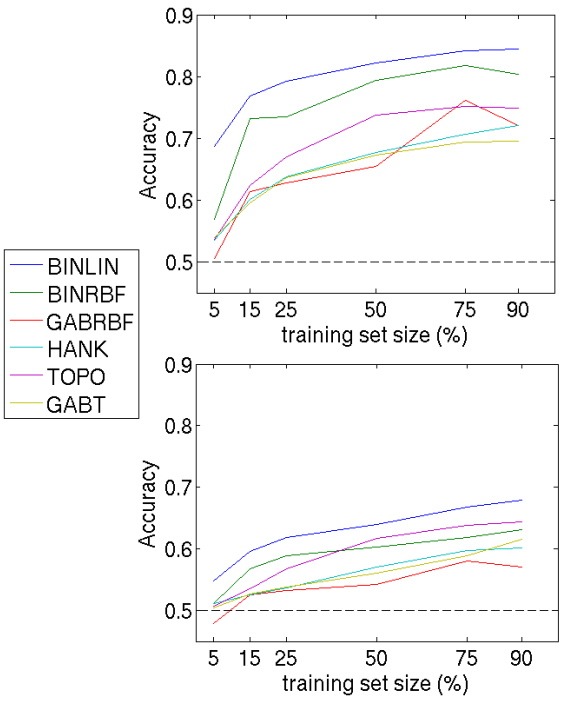


Figure 7: Influence of training set size for two subjects

4 DISCUSSION

The idea of using subspace methods was to obtain a more robust measure of class discrimination. It appears that this is not the case. Particularly, the expectations that such methods would better withstand noise (represented here by additional brain activity when walking) and capture essential discriminative information with smaller data sets do not hold. Several possible reasons and influences should be mentioned. First of all, the vector methods use an additional feature selection, which allow them to focus on essential local features. The subspace methods, as they are applied here, do not have such an optimization. An additional bin-channel selection could improve the discrimination. This theory is confirmed by the performance of GABRBF, which is significantly lower than BINRBF. One the one hand, the features are based on loose approximations of the trials based on the grand average, but more importantly, the features are global and cannot exploit feature selection as the binning methods.

A second reason is the possibility of generalization, essential for a good classification. The linear method mostly outperforms the others due to its necessary approximation of the decision boundary. The other methods are all RBF-based (even the tensorial), thus having the universal approximation property. Despite the regularization of (LS-)SVM, overfitting seems to occur. This becomes clear when com-

Table 1: Number of times the method mentioned in the row outperforms the method in the column ($p < 0.05$) for the single trial case in sitting condition

	BL	BR	GR	HNK	TOP	GT
BINLIN	0	5	16	13	13	13
BINRBF	2	0	14	9	8	11
GABRBF	0	0	0	1	1	2
HANK	2	1	7	0	4	6
TOPO	3	2	12	8	0	10
GABT	1	1	8	4	5	0

paring the high training accuracy with the significantly lower test accuracy. The gap shrinks with a larger training set, as was indeed observed in the second experiment. Yet, the total data set remains small for accurate learning, which cannot be overcome.

Thirdly, the tensorial kernel presupposes an underlying structure of the data. Each class is supposed to correspond to a congruence set; that is, to be generated from a multilinear basis which characterizes the class. Most tensor representations do not completely fulfil this condition e.g. due to noise. The extent to which they do can e.g. be estimated from kernel-target alignment (Cristianini et al., 2002), calculated for the tensorial kernel.

It is hard to draw general conclusions with regard to the future of subspace methods from this study. The subspace structure implied by the tensorial kernel is just one way of measuring the similarity between trials. Furthermore, one could consider other data representations.

Finally, it should be noted that there are large differences among the subjects. In the best performing subject, the results as presented above hold. Yet, other subjects have comparable results for vector and tensor methods, or even better results with tensors, particularly for TOPO. As already mentioned, this is particularly true when averaging, but Figure 7 (bottom) gives an example for single trials as well. Table 1 is an additional illustration. It gives the number of subjects for which the method on the row significantly outperforms the method in the column ($p < 0.05$). The table supports the conclusion of BINLIN and BINRBF being the superior methods, while TOPO dominates the tensorial methods. Even apart from these rather negative results, tensor methods are computationally more expensive as well.

Further study could involve additional data sets, other dimensions or other types of tensors and kernels. Furthermore, instead of creating data representations to directly measure similarities, tensors could also be used for regularization or to extract structural features for use in a vector classifier.

5 CONCLUSIONS

Subspace discrimination of tensorial data has been combined with kernels for BCI ERP classification. Three tensorial data representations were introduced: one based on Hankel matrices, one on topographic maps and one on time-frequency matrices. By means of analyses on the influence of condition, training set size and averaging, they were evaluated against more conventional local (time-channel bins) and global (gabor atom matching) feature vector methods with linear or RBF kernels. By and large, the tensorial approach does not yield any advantage, although it is at least competitive for some subjects, particularly the topographic method.

ACKNOWLEDGEMENTS

- Research Council KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC);
  PhD/Postdoc grants
- Flemish Government:
  - FWO: projects: G.0427.10N (Integrated EEG-fMRI), G.0108.11 (Compressed Sensing) G.0869.12N (Tumor imaging) G.0A5513N (Deep brain stimulation); PhD/Postdoc grants
  - IWT: projects: TBM 080658-MRI (EEG-fMRI), TBM 110697-NeoGuard; PhD/Postdoc grants
  - iMinds Medical Information Technologies SBO 2014, ICON: NXT_Sleep Flanders Care: Demonstratieproject Tele-Rehab III (2012-2014)
- Belgian Federal Science Policy Office: IUAP P7/19/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017)
- Belgian Foreign Affairs-Development Cooperation: VLIR UOS programs
- EU:
  - EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Advanced Grant: BIOTENSORS (n 339804).This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information.
- other EU funding: RECAP 209G within INTERREG IVB NWE programme, EU MC ITN TRANSACT 2012 (n 316679), ERASMUS EQR:

Community service engineer (n 539642-LLP-1-2013)

REFERENCES

Cristianini, N., Kandola, J., Elisseeff, A., and Shawe-Taylor, J. (2002). On kernel-target alignment. In *Advances in Neural Information Processing Systems 14*, pages 367–373. MIT Press.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278.

De Vos, M., Gandras, K., and Debener, S. (2013). Towards a truly mobile auditory brain-computer interface: Exploring the P300 to take away. *International Journal of Psychophysiology*.

Debener, S., Minow, F., Emkes, R., Gandras, K., and De Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? *Psychophysiology*, 49(11):1617–1621.

Gabor, D. (1946). Theory of Communication. *Journal of the Institution of Electrical Engineers*, 93(26):429–457.

Halder, S., Rea, M., Andreoni, R., Nijboer, F., Hammer, E. M., Kleih, S. C., Birbaumer, N., and Kuebler, A. (2010). An auditory oddball brain-computer interface for binary choices. *Clinical Neurophysiology*, 121:516–523.

Hansen, P. C. and Jensen, S. H. (1998). FIR filter representations of reduced-rank noise reduction. *IEEE Transactions on Signal Processing*, 46(6):1737–1741.

Hunyadi, B., Signoretto, M., Debener, S., Huffel, S. V., and Vos, M. D. (2013). Classification of structured EEG Tensors using Nuclear Norm Regularization: Improving P300 Classification. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 98–101. IEEE.

Jackson, M. M. and Mappus, R. (2010). Applications for Brain-Computer Interfaces. In *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*, chapter 1. Springer.

Kung, S. Y., Arun, K. S., and Rao, D. V. B. (1983). State-space and singular-value decomposition-based approximation methods for the harmonic retrieval problem. *J. Opt. Soc. Am.*, 73(12):1799–1811.

Li, J. and Zhang, L. (2010). Regularized tensor discriminant analysis for single trial EEG classification in bci. *Pattern Recognition Letters*, 31:619–628.

Liu, H. and Motoda, H., editors (2008). *Computational Methods of Feature Selection*. Chapman & Hall.

Mallat, S. and Zhang, Z. (1993). Matching Pursuit with Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415.

Nasehi, S. and Pourghassem, H. (2011). Real-Time Seizure Detection based on EEG and ECG Fused Features using Gabor Functions. *International Conference on Intelligent Computation and Bio-Medical Instrumentation*, 0:204–207.

Onishi, A., Phan, A. H., Matsuoka, K., and Cichocki, A. (2012). Tensor classification for P300-based brain computer interface. In *ICASSP*, pages 581–584. IEEE.

Signoretto, M. (2011). *Kernels and Tensors for Structured Data Modelling*. PhD thesis, KULeuven.

Suykens, J. A. K. and Vandewalle, J. (1999). Least Squares Support Vector Machine Classifiers. *Neural Process. Lett.*, 9(3):293–300.

Wexler, J. and Raz, S. (1990). Discrete Gabor Expansions. *Signal Processing*, 21(3):207–220.

Zhao, Q., Zhou, G., Adali, T., Zhang, L., and Cichocki, A. (2013). Kernelization of Tensor-Based Models for Multiway Data Analysis. *IEEE Signal Processing Magazine*, 30(4):137–148.