Loughborough University

This item was submitted to Loughborough's Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

A Hybrid Algorithm for Removal of Eye Blinking Artifacts from Electroencephalograms

Leor Shoker *Member, IEEE*, Saeid Sanei *Senior Member, IEEE*, and Jonathon Chambers *Senior Member, IEEE*
Centre of Digital Signal Processing, Cardiff School of Engineering,
Cardiff University Queen's Building, PO Box 925, Cardiff, CF24 3AA, U.K.
Tel: +44(0)29 2087 5946 Fax: +44(0)29 2087 4716   e-mail: {shokerl,saneis,chambersj}@cf.ac.uk

Abstract— A robust method for removal of artifacts such as eye blinks and electrocardiogram (ECG) from the electroencephalograms (EEGs) has been developed in this paper. The proposed hybrid method fuses support vector machines (SVMs) based classification and blind source separation (BSS) based on independent component analysis (ICA). The carefully chosen features for the classifier mainly represent the data higher order statistics. We use the second order blind identification (SOBI) algorithm to separate the EEG into statistically independent sources and SVMs to identify the artifact components and thereby to remove such signals. The remaining independent components are remixed to reproduce the artifact free EEGs. Objective and subjective results from the simulation studies show that the algorithm outperforms previously proposed algorithms.

Index Terms— Eye blinking artifact, blind source separation, support vector machines, electroencephalogram and electrooculogram.

I. INTRODUCTION

In EEG analysis ocular artifacts (OAs) pose a significant problem to the clinician. OAs such as eye blinks generate a signal that is on the order of ten times larger in amplitude than cortical signals and can last between 200 to 400ms. The eyeball can be considered as a dipole rotating in a socket. This means that as the eye rotates, the cornea remains at a potential 0.4 to 1 mV positive with respect to the retina. Rotations of the eyeball in saccadic eye movements cause large external field variations that can contaminate EEG readings [1]. Due to the magnitude of the blinking artifacts and the high resistance of the scalp, OAs can contaminate the majority of electrodes.

Principal component analysis (PCA) and singular value decomposition (SVD) [2] have been used in the past by some researchers to remove OAs. These methods assume that the underlying sources within EEG data are algebraically orthogonal, an assumption which from a physiological perspective is believed to be untrue. Therefore, PCA type methods will not completely remove the OAs. Application of adaptive filtering has also been investigated [3]. This, however, has limited success since it ignores the mutual information between the electrodes. There are several techniques that use independent component analysis (ICA) to separate the EEG into its constituent independent components (ICs) and then manually cancel the ICs that are believed to contribute to the OAs [4].

An automated method for removing OAs from EEGs has been described in [5]. The authors used a BSS algorithm based on second order statistics, to separate the EEG and measured EOG into statistically independent sources. The separation is then performed a second time on the raw EEGs but with a selection of EOG channels inverted. The ICs which have been found after inversion are compared with the ICs of the previous separation and those which invert are removed. In addition, the ICs that are above a threshold of correlation with the measured reference are removed, as are the ICs with high power in the low frequencies. The main drawback of this method is that it is restricted to having the reference EOG channels, which may not be available if one would like to process previously recorded data.

In this paper we therefore demonstrate the use of BSS to extract the separated sources and use an SVM to identify and thereby eliminate the sources contributing to the eye blinking artifacts.

II. METHODS

A. Blind Source Separation

BSS generally relies on the fundamental assumption that the source signals, $\mathbf{s}(t) = [s_1(t), s_2(t), \ldots, s_N(t)]^T$ are statistically independent and zero mean, where t is the discrete time index and N is the number of sources. The mixtures can be modeled by

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{v}(t) \quad (1)$$

where $\mathbf{A}$ is the $M \times N$ full column rank mixing matrix, M is the number of observed electrodes, $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_M(t)]^T$ contains the linear mixtures observed at the electrodes and $\mathbf{v}(t) = [v_1(t), v_2(t), \ldots, v_M(t)]^T$ is the additive zero mean sensor noise. We assume that the sensor noise is spatially uncorrelated with the sensor data i.e. $E\{\mathbf{x}(t)\mathbf{v}^T(t)\} = \mathbf{0}$. The output of the ICA system (i.e. the estimated sources) is given by

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \quad (2)$$

where $\mathbf{y}(t) = [y_1(t), y_2(t), \ldots, y_N(t)]^T$ is the vector of the estimated sources, $\mathbf{W}$ is the $N \times M$ separation matrix.

We used a gradient based BSS algorithm that exploits the temporal structure of EEGs as in [6] to process the EEGs. The goal of the proposed algorithm in [6] is to jointly diagonalize a set of output covariance matrices $\mathbf{R}_Y(k) \triangleq E\{\mathbf{y}(t)\mathbf{y}^T(t+k)\}$ $k \in 1, 2, \ldots, K$, where K is the maximum time lag. We use multiple time lags instead of a single time lag, as in the AMUSE algorithm [7], to minimize the chance of the time lagged covariance matrix having duplicate eigenvalues [8] and hence the separation to fail. It is sufficient that only one of the time lagged covariance matrices has unique eigenvalues for successful separation, hence the algorithm is more robust at the slight cost of increased computational complexity [8].

The output covariance matrix $\mathbf{R}_Y(k)$ is given by,

$$\mathbf{R}_Y(k) = \mathbf{W}[\mathbf{R}_X(k) - \mathbf{R}_V(k)]\mathbf{W}^T \quad (3)$$

where $\mathbf{R}_X(k)$ is the time lagged covariance matrix of the signal mixtures and similarly $\mathbf{R}_V(k)$ is the covariance matrix of the sensor noise. Since we assume that the noise is spatially white and temporally uncorrelated, $\mathbf{R}_V(k)$ will be a diagonal matrix for $k = 0$ and $\mathbf{R}_V(k) = \mathbf{0}$ for $k \neq 0$ [9].

$$\mathbf{R}_X(k) - \mathbf{R}_V(k) = \mathbf{A}\mathbf{R}_S(k)\mathbf{A}^T - \mathbf{R}_V(k) \quad (4)$$

where $\mathbf{R}_S(k)$ is a diagonal covariance matrix of the independent source signals. A suitable practical cost function is therefore defined based upon minimizing the off-diagonal elements for multiple lagged covariance matrices, as

$$\mathbf{W}_{opt} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} J_M(\mathbf{W}, k) \quad (5)$$

and

$$J_M(\mathbf{W}, k) = ||\mathbf{R}_Y(k) - diag(\mathbf{R}_Y(k))||_F^2 \qquad (6)$$

where $diag(\cdot)$ is an operator which zeros the off-diagonal elements of a matrix. To implement this approach we use a gradient decent algorithm to find $\mathbf{W}_{opt}$ iteratively, which is given by

$$\mathbf{W}(\kappa + 1) = \mathbf{W}(\kappa) + \Delta\mathbf{W}(\kappa) \qquad (7)$$

where κ is the iteration number and

$$\Delta\mathbf{W}(\kappa) = -\mu \sum_{k=1}^{K} \left.\frac{\partial J_M(\mathbf{W}, k)}{\partial\mathbf{W}}\right|_{\mathbf{W}=\mathbf{W}(\kappa)} \qquad (8)$$

where μ is the learning rate and $\frac{\partial J_M(\mathbf{W},k)}{\partial\mathbf{W}}$ is the gradient of the cost function in (5) evaluated at $\mathbf{W} = \mathbf{W}(\kappa)$. The non negative parameter μ is typically $\ll 1$ and $\mathbf{W}(0) = \mathbf{I}$. In reality the algorithm can only approximately jointly diagonalize the matrices as the linear model in (3) may not accurately describe the generation of the sources and due to the estimation errors in the sample covariance matrices.

B. Feature Extraction

The four most effective features we have found which efficiently discriminate the artifact signal from the normal EEG are as follows:

1) Feature 1: A large ratio between the peak amplitude and the variance of a signal suggests that there is an unusual value in the data. This is a typical identifier for the eye blink because it causes a large deflection on the EEG trace. The equation describing this feature is given by

$$f_1 = \frac{\max(|\mathbf{u}_n|)}{\sigma_u^2} \quad n = 1, \ldots, N \qquad (9)$$

where $\mathbf{u}_n$ is one of the N ICs, $\max(\cdot)$ is a scalar valued function that returns the maximum element in a vector, σ_u is the standard deviation of $\mathbf{u}_n$ and $|\cdot|$ is the absolute value applied element-wise in (9). The normal EEG activity is tightly distributed about its mean value, therefore a low ratio is expected for it in contrast to ICs containing eye blink sources where a high value is expected.

2) Feature 2: This feature corresponds to a third order statistic of the data. The normalized skewness for each IC is given by

$$f_2 = \left|\frac{E\{u_n^3(i)\}}{\sigma_u^3}\right| \quad n = 1, \ldots, N \qquad (10)$$

for zero mean data. An EEG containing eye blinks typically has a positive or negative skewness since the eye blinking artifact increases locally the asymmetry of the signal segment. Hence we take the absolute value of the skewness. Significance of this feature in the overall classification is high since the eye blink signal has larger skewness than that of normal EEGs, which are approximately symmetrically distributed.

3) Feature 3: For the third feature we find the correlation between the IC and a data set containing eye blinking artifact from six electrodes including the frontal electrodes close to the eyes (F_{P1}, F_{P2}, F_3, F_4) and the electrodes on the occipital lobe (O_1, O_2). This will make the classification more robust by introducing a measure of the spatial location of the eye blinking artifact. We average the maximum value of cross-correlation between each of the electrode locations and the IC

$$f_3 = \frac{1}{6}\sum_{i=1}^{6}\max_\tau\left(\left|E\{x_i^0(t)u_n(t+\tau)\}\right|\right) \quad n = 1, \ldots, N \qquad (11)$$

where $u_n(t)$ is the nth independent component and $x_i^0(t)$ are eye blinking reference signals, where i indexes each of the aforementioned electrode locations. The value of this feature will be larger for ICs containing eye blinking artifact, since they will have a larger

correlation for a particular value of τ in contrast to ICs containing normal EEG activity, the maximum $\tau \approx \sqrt{T_B}$.

4) Feature 4: The fourth feature is the statistical distance between the probability density function (PDF) of an IC and the PDF of a reference IC known to contain OA. Here we assume that the PDF of the IC containing the artifact will be identical to that of the reference signal containing artifact. To measure the statistical distance between the two PDFs we used the Kullback-Leibler (KL) distance, given by

$$\begin{aligned} f_4 &= KL(P(\mathbf{u}_n)||P(\mathbf{x}_{ref})) \\ &= \int P(\mathbf{u}_n)\ln\frac{P(\mathbf{u}_n)}{P(\mathbf{x}_{ref})}\, d\mathbf{u}_n \quad n = 1, \ldots, N \end{aligned} \qquad (12)$$

where $P(\mathbf{u}_n)$ and $P(\mathbf{x}_{ref})$ are the PDFs of one of the N ICs and a previously measured artifact, respectively. When the IC contains OAs the KL distance between its PDF and the PDF of the reference IC will be approximately zero, whereas the distance to the PDF of a normal EEG signal will be larger. Since the KL distance is related to the mutual information it reflects effectively the information shared between the IC and the reference.

C. Classification

We use an SVM as our classification method, due to its generalization performance, and its established empirical performance [10]. The goal of an SVM is to find an optimal separating hyperplane (OSH) for a given feature set. The OSH is found by solving the following constrained optimisation,

$$\min_{\mathbf{z},b,\gamma_{i=1,\ldots,l}}\left(\tfrac{1}{2}||\mathbf{z}||^2 + C\sum_{i=1}^{l}\gamma_i\right)$$
$$\text{s.t.} \quad q_i(\mathbf{z}\cdot\mathbf{g}_i - b) + \gamma_i \geq 0 \quad i = 1, \ldots, l \qquad (13)$$

where $||\mathbf{z}||^2 = \mathbf{z}^T\mathbf{z}$ is the squared Euclidean norm and $(\cdot)$ is the dot product. The parameter $\mathbf{z}$ determines the orientation of the separating hyperplane, γ_i is the i-th positive slack parameter, $\mathbf{g}_i$ is a vector containing the features $\mathbf{g}_i = [f_1(i)\ f_2(i)\ f_3(i)\ f_4(i)]^T$. Here, l is the number of training vectors and $q_i \in \{\pm1\}$ are the output targets. The non negative parameter C is the (misclassification) penalty term, and can be considered as the regularization parameter and is selected by the user. A larger C is equivalent to assigning a higher penalty to the training errors. The parameter C is usually set to a high value to avoid any training error. SVs are the points from the dataset that fall closest to the separating hyperplane. Any vector $\mathbf{g}_i$ that corresponds to a non-zero α_i is a support vector (SV) of the optimal hyperplane. It is desirable to have the number of SVs small to have a more compact and parsimonious classifier. The OSH (generally nonlinear) is then computed by solving (II-C) using Karush-Kuhn-Tucker conditions [11] as a decision surface of the form

$$f(\mathbf{g}) = \text{sgn}\left(\sum_{i=1}^{L_s}q_i\alpha_i K(\mathbf{g}_i^s, \mathbf{g}) + b\right). \qquad (14)$$

In this formula $\text{sgn}(\cdot) \in \{\pm1\}$, $\mathbf{g}_i^s$ are SVs, $K(\mathbf{g}_i^s, \mathbf{g})$ is the nonlinear kernel function (if $K(\mathbf{g}_i^s, \mathbf{g}) = \mathbf{g}_i^s \cdot \mathbf{g}$ the SVM is linear). A Kernel for a nonlinear SVM projects the samples to a feature space of higher dimension via a nonlinear mapping function. Among nonlinear kernels the radial based function (RBF) defined as $K(\mathbf{g}_i, \mathbf{g}) = \exp(-|\mathbf{g} - \mathbf{g}_i|^2/(2\rho))$, where the adjustable parameter ρ governs the variance of the function, is widely used due to having quasi-Gaussian distribution for datasets of large samples.

III. EXPERIMENTS

A. Dataset for analysis

The data were provided by King's College Hospital, London U.K. and are available from our website [12]. The data represent a wide

Kernel	Average classification rate (%) (s.d.)		
	Overall	Normal	Eye Blinks
Cubic Polynomial	94.50 (1.92)	91.15 (2.31)	97.91 (2.04)
Linear	99.00 (1.15)	99.24 (1.11)	99.21 (0.97)
Gaussian RBF	98.50 (1.00)	98.26 (1.17)	99.03 (1.35)

range of patients and therefore gives a comprehensive set of data for the evaluation of our method. The scalp EEG was obtained using Silver/Silver-Chloride electrodes placed at locations defined by the 10-20 system. The data were acquired using a Beekeeper Telefactor EEG amplifier, sampled at 200Hz and bandpass filtered with cutoff frequencies of 0.3Hz and 70Hz. We obtained the independent components by applying BSS to blocks of data, 10 seconds in length. We assumed that the number of sources is the same as the number of electrodes (i.e. N=M). Then, we extracted the features from each of the ICs. The classifier was trained using the ICs from different patients.

B. Testing the Features

In our study we tested the features using 200 ICs; 100 ICs containing eye blinks and 100 free of artifact. The classifier was tested using a variety of kernels, for which the error results are shown in Table I. The value chosen for the parameter C was 64 and in the case of the RBF kernel the value of ρ was empirically found as 0.75. For each kernel the average error values were estimated with 4-fold cross validation (CV) i.e. using 75% of the data as training examples and 25% for testing with no overlapping. The cross-validation was performed 10 times, each time the data were randomly rearranged, in order to yield a better estimate of the error.

The average number of support vectors found when using the RBF kernel was 37% of the training examples. In the case of cubic polynomial and linear kernels the number of support vectors found were 18% and 3.3% respectively.

The training error was found by using the training data to test the SVM. The training error was found to be 0.5% (av) and the test error was 0.7% (av). This avoids any overfitting since the training error is close to the training error.

The classifier was further evaluated by plotting the distribution of the classifier output. It is calculated by applying the classification function in (14) without the sgn($\cdot$) function. The result from the test data using the linear kernel is shown in Fig. 1. The ICs containing eye blinks are clustered around and above the +1 value and the ICs containing normal EEG activity around and below the -1 value. There is minimal overlap in the classifier output, indicating that the proposed features are significant to the detection of eye blinking artifacts for the test datasets.

For our dataset mainly due to reasonably separable classes there are no significant differences between the results using different kernels. However, the linear kernel requires far fewer SVs in calculating the OSH, hence the linear kernel will be used to classify eye blinks in the following experiments.

We applied the BSS-SVM algorithm to 10 real EEG datasets, each were 7 minutes long. The performance of the algorithm can be seen by comparing the EEG data obtained at the electrodes (see Fig. 2) and the same segment of data after being processed by the proposed
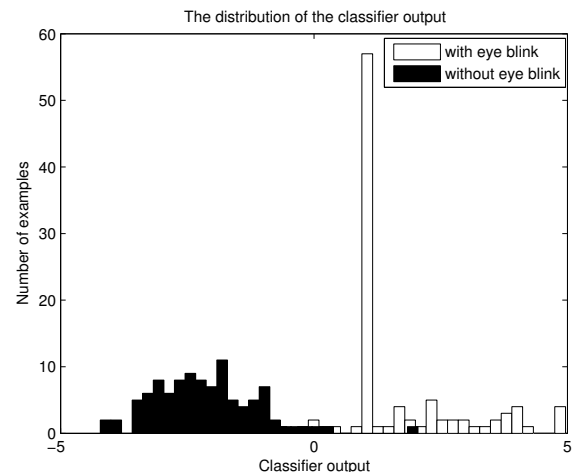

Fig. 1. A histogram plot showing the distance to the separating hyperplane using the linear kernel.

algorithm (see Fig. 3). The significance of our results was justified by clinicians at King's College Hospital. The proposed algorithm was compared to a manual artifact rejection, i.e. manually identifying and canceling the artifact by calculating the cross correlation between the BSS-SVM and the manually reconstructed EEG. The average value of cross correlation is 0.92 (s.d. 0.02). In a number of trials the effect of ECG has been automatically detected and removed, where as the complete removal has not been achieved with the method based on the manual selection. This had a detrimental effect on the cross-correlation measure since the BSS-SVM output will be less correlated with the manually reconstructed outputs, but has a positive effect on the output since there is less artifact present in the output.

As a second criterion for measuring the performance of the overall system we selected a segment of EEG, x_{seg}, and the reconstructed EEG, $\tilde{x}_{seg}$, which doesn't contain any artifact and measured the waveform similarity, $\varepsilon_{dB} = 10 \log \left(1/M \sum_{i=1}^{M} \left(1 - E\{ (x_{i,seg}[n] - \tilde{x}_{i,seg}[n]) \} \right) \right)$. When the value of ε_{dB} is zero, the original and reconstructed waveforms are identical. From ten sets of EEGs the average waveform similarity was $\varepsilon_{dB} = -0.009dB$ (standard deviation $10^{-4}dB$). These results suggest that the observations have been faithfully reconstructed both in terms of subjective visual inspection and objective performance metrics.

IV. CONCLUSION

Based on our experimental results, the BSS-SVM algorithm effectively removes the effect of eye blinking artifacts on the EEGs. The experiments herein demonstrate that for the test dataset the eye blinking artifacts are effectively classified by using the introduced features especially when the linear kernel is used for the SVM. The EEGs are separated using the time lagged SOBI algorithm and the identified artifacts are autonomously canceled, then the EEG is reconstructed from the remaining ICs.

REFERENCES

[1] D. A. Overton and C. Shagass, "Distribution of eye movement and eyeblink potentials over the scalp," *Electroencephalography and Clinical Neurophysiology*, vol. 27, p. 546, 1969.
[2] O. G. Lins, T. W. Picton, P. Berg, and M. Scherg, "Ocular artifacts in EEG and event-related potentials, i: Scalp topography," *Brain Topography*, vol. 6, pp. 51–63, 1993.
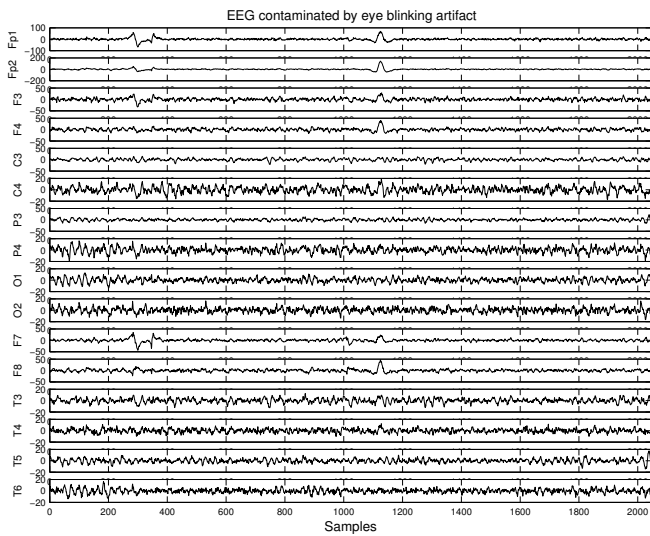
Fig. 2. A 16 electrode EEG recording. The OAs are clear between samples 200 to 400 and 1000 to 1200. They are more prominent on the frontal electrodes(FP1, FP2 etc.).
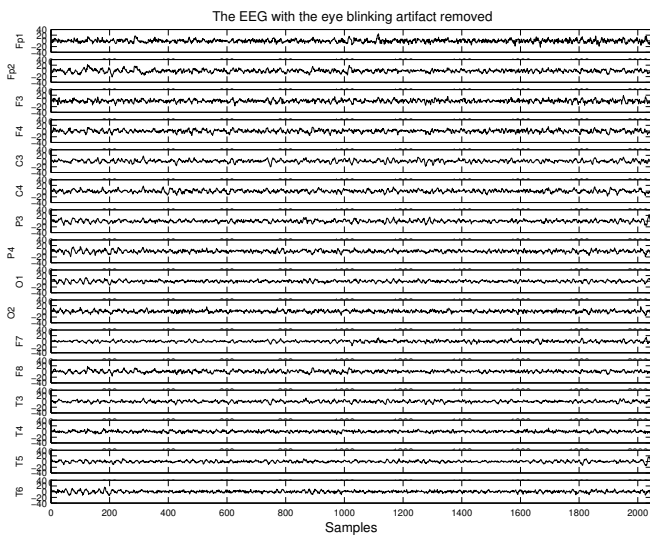

Fig. 3. The EEG after being processed by the proposed algorithm. The OAs have been removed from all of the electrodes.

[3] P. Celka, B. Boashash, and P. Colditz, "Preprocessing and time-frequency analysis of newborn EEG seizures," *IEEE Engineering Medicine and Biology Magazine*, vol. 20, pp. 30–39, 2001.

[4] T. P. Jung, C. Humphries, T. Lee, S. Makeig, M. McKeown, V. Iragui, and T. Sejnowski, "Extended ICA removes artifacts from electroencephalographic recordings," *Advances in Neural Information Processing Systems*, vol. 10, pp. 894–900, 1998.

[5] C. A. Joyce, I. Gorodnitsky, and M. Kautas, "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation," *Psychophysiology*, vol. 41, pp. 313–325, 2004.

[6] M. Joho and H. Mathis, "Joint diagonalization of correlation matricies by using gradient methods with application to blind signal processing," *in Proc. SAM2002*, pp. 273–277, 2002.

[7] L. Tong, V. Soon, R. Liu, and Y. Huang, "AMUSE: a new blind identification algorithm," in *Proc. ISCAS*, New Orleans, 1990.

[8] A. Hyvarinen, E. Oja, and J. Karhunen, *Independent Component Analysis.* Thomson Learning, Inc, 2001.

[9] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 320–327, May 2000.

[10] S. Gunn, "Support vector machines for clasification and regression," Univ. of Southampton, Tech. Rep., 1998.

[11] R. Fletcher, *Practical methods of optimisation, 2nd Ed.* John Wiley, 1987.

[12] Centre of DSP, Cardiff University website, [Online], Available: http://www.engin.cf.ac.uk/research/groups/cdsp/index.html.