Artifact Removal From Electroencephalograms Using a Hybrid BSS-SVM Algorithm

Leor Shoker, *Member, IEEE*, Saeid Sanei, *Senior Member, IEEE*, and Jonathon Chambers, *Senior Member, IEEE*

Abstract—Artifacts such as eye blinks and heart rhythm (ECG) cause the main interfering signals within electroencephalogram (EEG) measurements. Therefore, we propose a method for artifact removal based on exploitation of certain carefully chosen statistical features of independent components extracted from the EEGs, by fusing support vector machines (SVMs) and blind source separation (BSS). 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 assessment of the simulation results shows that the algorithm is successful in mitigating the interference within EEGs.

Index Terms—Artifact removal, blind source separation (BSS), electroencephalogram (EEG), electrooculogram, support vector machines (SVMs).

I. INTRODUCTION

IN electroencephalogram (EEG) measurements, ocular artifacts (OAs), such as eye blinks, remain a significant nuisance to the clinician. For example, in event-related potential (ERP) analysis, data segments containing OAs are rejected, resulting in substantial data loss, which may lead to failure of a clinical assessment.

Interfering eye blinks generate a signal within EEGs that is on the order of ten times larger in amplitude than cortical signals and can last between 200–400 ms. 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 0.4–1 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, even those in the occipital area.

In previous attempts to remove OAs, some researchers have tried to estimate the propagation factors, as discussed in [2], based on regression techniques in both the time and frequency domains. These methods are limited in their success, and furthermore, all of the methods mentioned therein required a measured reference electrooculogram (EOG) channel during the measurement of the EEG. Principal component analysis (PCA) and singular value decomposition (SVD) [3] have been used in the past by some researchers to remove OAs. These methods assume that the underlying sources within EEG data
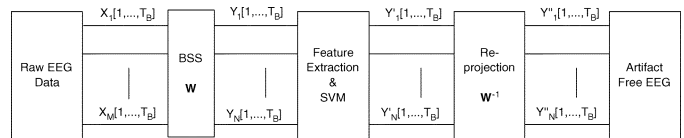
Fig. 1.　Block diagram of the BSS and SVM system.

are algebraically orthogonal, which is an assumption that 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 [4]. 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 eliminate the ICs that are believed to contribute to the artifact sources [5]. The authors in [5] used a graphical user interface (GUI) to display and highlight the ICs that are believed to correspond to such artifacts. The user would then decide if the highlighted component should be discarded, and the remaining ICs would be remixed.

An automated method for removing OAs from EEGs has been described in [6]. The authors used a blind source separation (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 that have been found after inversion are compared with the ICs of the previous separation, and those that 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 letter, 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 artifacts. We focus on the eye blinking artifact in this letter, but the method extends naturally to the removal of heart rhythm. A block diagram of the proposed system is shown in Fig. 1. In the first stage of the block diagram, the EEGs are acquired and stored; these are denoted in the diagram as $X_{1,\ldots,M}(1,\ldots,T_B)$, where M is the number of electrodes, and T_B is the data block length. Similarly, the intermediate signals within the figure are denoted as $Y_{1,\ldots,N}, Y'_{1,\ldots,N}, Y''_{1,\ldots,N}$, where N is the number of the estimated sources. In the second stage of Fig. 1, we separate the EEGs into statistically independent sources using BSS. Then, features are extracted from the independent sources and used to establish whether the source contains an eye blinking artifact.

Finally, the sources that are not identified as artifact are used to reconstruct the artifact-free EEGs through reprojection.

II. METHODS

A. BSS

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

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

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., $E\{\mathbf{A}\mathbf{s}(t)\mathbf{v}^T(t)\} = \mathbf{0}$, and temporally uncorrelated, i.e., $E\{\mathbf{v}(t)\mathbf{v}^T(t+\tau)\} = \mathbf{0} \,\forall \tau > 0$ and $M \geq N$. The output of the ICA system (i.e., the estimated sources) is given by

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

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

We used a gradient-based BSS algorithm that exploits the temporal structure of the underlying EEG sources as in [7] to process the EEGs. The goal of the proposed algorithm in [7] 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 employ multiple time lags instead of a single time lag, as in the AMUSE algorithm [8], to minimize the chance of the time-lagged covariance matrix having duplicate eigenvalues [9] 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 [9].

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

$$\mathbf{R}_Y(k) = \mathbf{W}\left[\mathbf{R}_X(k)\right]\mathbf{W}^T \qquad (3)$$

where $\mathbf{R}_X(k)$ is the time-lagged covariance matrix of the signal mixtures and is diagonal when $\mathbf{W}$ is a nonzero unmixing solution. A suitable practical cost function to find $\mathbf{W}$ is, therefore, defined based upon minimizing the off-diagonal elements for multiple-lagged covariance matrices, as

$$\mathbf{W}_{\text{opt}} = \arg\min_{\mathbf{W}} \sum_{k=1}^{K} J_M(\mathbf{W}, k) \qquad (4)$$

and

$$J_M(\mathbf{W}, k) = \|\mathbf{R}_Y(k) - \operatorname{diag}(\mathbf{R}_Y(k))\|_F^2 \qquad (5)$$

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

$$\mathbf{W}(\kappa + 1) = \mathbf{W}(\kappa) + \Delta\mathbf{W}(\kappa) \qquad (6)$$

where κ is the iteration number, and

$$\Delta\mathbf{W}(\kappa) = -\mu \sum_{k=1}^{K} \left.\frac{\partial J_M(\mathbf{W}, k)}{\partial \mathbf{W}}\right|_{\mathbf{W} = \mathbf{W}(\kappa)} \qquad (7)$$

where μ is the learning rate, and $(\partial J_M(\mathbf{W}, k))/\partial \mathbf{W}$ is the gradient of the cost function in (4) 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 that efficiently discriminate the artifact signal from the normal EEG are as follows.

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

$$f_1 = \frac{\max(|\mathbf{u}_n|)}{\sigma_u^2}, \qquad n = 1, \ldots, N \qquad (8)$$

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 (8). 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.

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

$$f_2 = \left|\frac{E\{u_n^3(i)\}}{\sigma_u^3}\right|, \qquad n = 1, \ldots, N \qquad (9)$$

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. The 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.

Feature 3: For the third feature, we find the correlation between the IC and an independent 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). The reference data set, i.e., the EEG from the aforementioned electrodes, is distinct from the training and test data sets. 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

$$f_3 = \frac{1}{6}\sum_{i=1}^{6}\max_{\tau}\left(\left|E\{x_i^0(t)u_n(t+\tau)\}\right|\right), \qquad n = 1, \ldots, N \qquad (10)$$

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 $\tau \approx \sqrt{T_B}$.

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. The OA reference IC is taken from a data set that is also distinct from the training and test data sets. Here, we assume that the PDF of the IC containing the artifact will be identical to that of the reference signal containing the artifact. To measure the statistical distance between the two PDFs, we used the Kullback–Leibler (KL) distance, given by

$$f_4 = KL\left(P(\mathbf{u}_n)\|P(\mathbf{x}_{\text{ref}})\right)$$
$$= \int P(\mathbf{u}_n) \ln \frac{P(\mathbf{u}_n)}{P(\mathbf{x}_{\text{ref}})} d\mathbf{u}_n, \qquad n = 1, \dots, N \quad (11)$$

where $P(\mathbf{u}_n)$ and $P(\mathbf{x}_{\text{ref}})$ 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 optimization:

$$\min_{\mathbf{z}, b, \gamma_{i=1,\dots,l}} \left(\frac{1}{2}\|\mathbf{z}\|^2 + C \sum_{i=1}^{l} \gamma_i\right)$$
$$\text{s.t. } q_i(\mathbf{z} \cdot \mathbf{g}_i - b) + \gamma_i \geq 0, \qquad i = 1, \dots, l \quad (12)$$

where l is the number of training vectors, and $q_i \in \{\pm 1\}$ are the output targets, $\|\mathbf{z}\|^2 = \mathbf{z}^T \mathbf{z}$ 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 ith positive slack parameter, and $\mathbf{g}_i$ is a vector containing the features $\mathbf{g}_i = [f_1(i)\ f_2(i)\ f_3(i)\ f_4(i)]^T$. 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 set to a value that yields the lowest cross-validation (CV) test error. Support vectors (SVs) are the points from the data set that fall closest to the separating hyperplane. Any vector $\mathbf{g}_i$ that corresponds to a nonzero α_i is an 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 as a decision surface of the form

$$f(\mathbf{g}) = \text{sgn}\left(\sum_{i=1}^{L_s} q_i \alpha_i K\left(\mathbf{g}_i^s, \mathbf{g}\right) + b\right) \quad (13)$$

where $\text{sgn}(\cdot) \in \{\pm 1\}$, $\mathbf{g}_i^s$ 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), and L_s is the number of support vectors. 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 $K(\mathbf{g}_i, \mathbf{g}) = \exp(-|\mathbf{g} -$

TABLE I

PERFORMANCE OF THE CLASSIFIER BASED ON THE AVERAGE NUMBER OF CORRECTLY CLASSIFIED POINTS. THREE KERNELS ARE COMPARED IN THE CLASSIFICATION

Kernel	Average classification rate (%) (s.d.)		
	Overall	Normal	Eye Blinks
Gaussian RBF	98.50 (1.00)	98.26 (1.17)	99.03 (1.35)
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)

$\mathbf{g}_i|^2/(2\rho))$, where the adjustable parameter ρ governs the variance of the function, is widely used due to having quasi-Gaussian distribution for data sets with a large number of samples.

III. EXPERIMENTS

A. Data Set for Analysis

The data were provided by King's College Hospital, London, U.K., and are available from our website [11]. The data represent a wide 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 200 Hz and bandpass filtered with cutoff frequencies of 0.3 Hz and 70 Hz. We obtained the independent components by applying BSS to blocks of data, 10 s 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 [12] was tested using a variety of kernels. For each kernel, the average error values were estimated with four-fold cross-validation, i.e., using 75% of the data as training examples and 25% for testing with no overlapping. The cross-validation was performed ten times, each time the data were randomly rearranged, in order to yield a better estimate of the error. To find the value of parameter C, we calculated the average CV test error for a range of values for C. The optimum value of C was found to be 64 in the case of the linear and cubic polynomial. For the RBF kernel, we adjusted C and ρ and found the optimum values for the RBF kernel as $C = 72$ and $\rho = 7$. The CV error results are shown in Table I.

The average number of support vectors found when using the RBF kernel was 7% of the training data set size. In the case of the cubic polynomial and linear kernels, the number of support vectors found were 18% and 3.3%, respectively, of the training data set size.

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 of the training data. It is calculated by applying the classification function in (13) without the $\text{sgn}(\cdot)$ function. The result from the training data using the linear kernel is shown in Fig. 2(b). The ICs containing eye blinks are clustered around and above the $+1$ value and the ICs containing normal
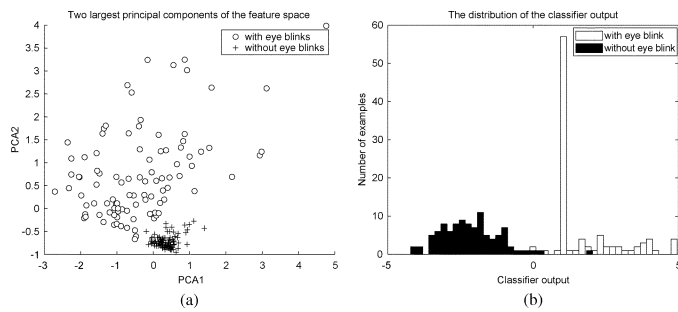
Fig. 2. (a) Plot of the two largest principal components of the feature space. There are 200 feature vectors: 100 from normal EEG ($+$) and 100 from EEG containing eye blinks ($\circ$). (b) Histogram plot showing the output of the classifier pre sgn($\cdot$) using the linear kernel.
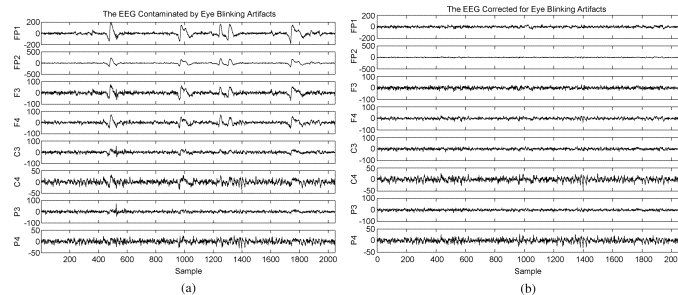

Fig. 3. Selection of eight electrodes from a 16-electrode EEG recording. The OAs are clear in (a) between samples 400–600, 900–1400, and 1700–1900. They are more prominent on the frontal electrodes (FP1, FP2, etc.). (b) Same segment of EEGs after being corrected for eye-blinking artifacts by using the proposed BSS-SVM algorithm.

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 data sets.

For our data set, there is only 0.5% difference in the overall classification rate between the linear kernel and the RBF kernel. The cubic polynomial had the lowest overall classification rate. The largest difference in classification performance was between the RBF and cubic polynomial kernel when classifying normal EEGs—there was a difference of 7.1%. The reason for the close overall classification rates is mainly due to the separability of the feature space. Since the linear kernel requires fewer SVs in calculating the OSH and due to its computational simplicity, the linear kernel will be used to classify eye blinks in the following experiments. In order to test the significance of the proposed features, we evaluated the principal components. We found the values of the components as 2.97, 0.68, 0.66, and 0.25; this testifies that the proposed features are significant to the detection of eye blinks in EEGs. A plot of the two largest principal components is shown in Fig. 2(a). From Fig. 2(a), it can be verified that the multidimensional feature space is nonseparable due to the overlapping region.

We applied the BSS-SVM algorithm to ten real EEG data sets, each being 7 min long. The performance of the algorithm can be seen by comparing the EEG data obtained at the electrodes [see Fig. 3(a)] and the same segment of data after being processed by the proposed algorithm [see Fig. 3(b)]. 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, whereas 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 does not contain any artifact, and measured the waveform similarity $\varepsilon_{dB} = 10 \log(1/M \sum_{i=1}^{M}(1 - E\{(x_{i,seg}[n] - \tilde{x}_{i,seg}[n])\}))$. 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 $\varepsilon_{dB} = .0.009$ dB (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

We have presented a robust method for removing OAs from EEGs by fusing BSS and SVM methods. The results show that the proposed algorithm identifies and removes the effect of eye-blinking artifacts. The experiments herein demonstrate that for the test data set, the signals 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. Based on our experimental data, the BSS-SVM algorithm consistently removes the effect of eye-blinking artifacts from the EEGs, whereas, in comparison, the removal by eye generally fails.

REFERENCES

[1] D. A. Overton and C. Shagass, "Distribution of eye movement and potentials over the scalp," *Electroencephalography Clin. Neurophys.*, vol. 27, p. 546, 1969.
[2] G. Gratton, "Dealing with artifacts: the EOG contamination of the event-related brain potential," *Behav. Res. Method., Instrum., Comput.*, vol. 30, no. 1, pp. 40–41, 1998.
[3] 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.
[4] P. Celka, B. Boashash, and P. Colditz, "Preprocessing and time-frequency analysis of newborn EEG seizures," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 5, pp. 30–39, Sep.-Oct. 2001.
[5] T. P. Jung, C. Humphries, T. Lee, S. Makeig, M. McKeown, V. Iragui, and T. Sejnowski, "Extended ICA removes artifacts from electroencephalographic recordings," *Adv. Neural Inf. Process. Syst.*, vol. 10, pp. 894–900, 1998.
[6] C. A. Joyce, I. Gorodnitsky, and M. Kautas, "Automatic removal of eye movement and blink artifacts from EEG data using blind component separation," *Psychophysiol.*, vol. 41, pp. 313–325, 2004.
[7] M. Joho and H. Mathis, "Joint diagonalization of correlation matricies by using gradient methods with application to blind signal processing," in *Proc. 1st Annu. Conf. Sensor Array Multichannel Signal Process.*, 2002, pp. 273–277.
[8] L. Tong, V. Soon, R. Liu, and Y. Huang, "AMUSE: a new blind identification algorithm," in *Proc. ISCAS*, New Orleans, LA, 1990.
[9] A. Hyvarinen, E. Oja, and J. Karhunen, *Independent Component Analysis (Adaptive & Learning Systems for Signal Processing, Communications & Control)*. New York: Wiley, 2001.
[10] S. Gunn, "Support vector machines for classification and regression," Tech. Rep., Univ. of Southampton, Southampton, U.K., 1998.
[11] Centre of DSP, Cardiff University website.. [Online]. Available: http://www.engin.cf.ac.uk/research/groups/cdsp/index.html
[12] S. Gunn. Support Vector Machines. University of Southampton. [Online]. Available: http://www.isis.ecs.soton.ac.uk/resources/svminfo/