Journal of Neuroscience Methods 268 (2016) 31–42

Contents lists available at ScienceDirect

Journal of Neuroscience Methods

journal homepage: www.elsevier.com/locate/jneumeth

An improved artifacts removal method for high dimensional EEG

Jidong Hou*, Kyle Morgan, Don M. Tucker, Amy Konyn, Catherine Poulsen, Yasuhiro Tanaka, Erik W. Anderson, Phan Luu

Electrical Geodesics Inc., Eugene, OR 97401, USA

HIGHLIGHTS

- An automated EEG artifact removal method is developed.
- A novel independent component analysis strategy is proposed and applied.
- Dimension reduction is achieved with permutation and resampling.
- Improved eye blinks removal performance is demonstrated.

ARTICLE INFO

Article history:
Received 18 January 2016
Received in revised form 7 April 2016
Accepted 4 May 2016
Available online 5 May 2016

Keywords:
EEG
ICA
Artifact removal
Template matching
Permutation
Resampling

ABSTRACT

Background: Multiple noncephalic electrical sources superpose with brain signals in the recorded EEG. Blind source separation (BSS) methods such as independent component analysis (ICA) have been shown to separate noncephalic artifacts as unique components. However, robust and objective identification of artifact components remains a challenge in practice. In addition, with high dimensional data, ICA requires a large number of observations for stable solutions. Moreover, using signals from long recordings to provide the large observation set might violate the stationarity assumption of ICA due to signal changes over time.

New method: Instead of decomposing all channels simultaneously, subsets of channels are randomly selected and decomposed with ICA. With reduced dimensionality of the subsets, much less amount of data is required to derive stable components. To characterize each independent component, an artifact relevance index (ARI) is calculated by template matching each component with a model of the artifact. Automatic artifact identification is then implemented based on the statistical distribution of ARI of the numerous components generated.

Results: The proposed permutation resampling for identification matching (PRIM) method effectively removed eye blink artifacts from both simulated and real EEG.

Comparison with existing method: The average topomap correlation coefficient between the cleaned EEG and the ground truth is 0.89 ± 0.01 for PRIM, compared with 0.64 ± 0.05 for conventional ICA based method. The average relative root-mean-square error is 0.40 ± 0.01 for PRIM, compared with 0.66 ± 0.10 for conventional method.

Conclusions: The proposed method overcame limitations of conventional ICA based method and succeeded in removing eye blink artifacts automatically.

1. Introduction

Some of the most important advances in neuroscience research and clinical diagnosis have been made with electroencephalography (EEG). However, the utility of EEG is limited due to artifacts caused by eye movements, eye blinks, muscle activity, heart signals, line noise, or noise from joint recordings conducted with other imaging modalities (e.g., MRI). These artifacts superpose with the EEG data, impacting statistical and physical analysis of the brain's contribution to the EEG signal.

Numerous strategies can be used to remove artifacts from the EEG. A straightforward approach is template subtraction. Based on the assumption of that the occurrence of the artifact is identical in each instance and the EEG signal is random relative to the artifact, a pure artifact template can be created through averaging

* Corresponding author at: Electrical Geodesics Inc., 500 East 4th Ave., Suite 200, Eugene, OR 97401, USA. Fax: +1 541 687 7963.
E-mail address: jhou@egi.com (J. Hou).

http://dx.doi.org/10.1016/j.jneumeth.2016.05.003
0165-0270/

multiple occurrences of the artifact. The template is then used to subtract each instance of the artifact from the raw signal without removing the signal of interest. However, the obvious limitation is that if the artifact is highly variable, the template does not match each instance, resulting in imperfectly cleaned data, even with several modifications and improvements to this template subtraction approach (Allen et al., 1998; Sijbers et al., 1999; Niazy et al., 2005).

Regression-based approaches have been applied to correct ocular artifacts (Gratton et al., 1983; Wallstrom et al., 2004) using the electrooculogram (EOG) recorded adjacent to the eyes as the reference signal. The problem is that the EOG is not free of EEG signals, such as from the frontal lobes. Similarly, Kalman filters and multichannel recursive least squares (M-RLS) algorithm were developed to remove motion and ballistocardiogram (BCG) artifacts, using accelerometers, wire loops or electrodes on a conducting layer to generate movement-tracking reference signals (Bonmassar et al., 2002; Masterton et al., 2007; Chowdhury et al., 2014). Although these techniques have proved useful, the relationship between the reference signal and the induced artifact may not always be linear, causing difficulty for a linear filtering approach. In addition, it is difficult to acquire a reference signal for some artifacts, such as electromyographic (EMG) activity that occurs at the scalp.

Blind source separation (BSS) is an approach that doesn't have the limitations noted for the above methods. The basic idea of BSS is to decompose the raw data into components that represent cortical activity and artifacts. Once the artifact components are identified, the brain signal can be reconstructed by excluding the artifact components. BSS can be carried out in many ways, e.g., canonical correlation analysis (CCA), principle component analysis (PCA) and independent component analysis (ICA). ICA has proven effective in separating unique components, particularly with powerful rotation methods such as Infomax (Bell and Sejnowski, 1995). Each independent component consists of a waveform that describes a source activity plus a topography vector that describes how the waveform contributes to the recorded signal. The advantage of ICA over PCA is that it doesn't require the source topography to be orthogonal, which is not a reasonable assumption of the brain's physiological activity in general. Instead of minimizing covariance among sources, ICA aims to maximize their independence. With the assumption that artifacts are statistically independent from the ongoing EEG, spatial filters derived by the ICA algorithm have been used to remove a wide variety of EEG artifacts, such as eye blinks, eye movements and electrode artifacts (Jung et al., 2000a,b; Ille et al., 2002; Flexer et al., 2005; Iriarte et al., 2003).

One practical challenge of the ICA approach to artifact removal is that manual interaction is usually required to identify artifact components following decomposition, based on either spatial topographies or temporal characteristics or both. Different components are generated for each data set, making manual selection cumbersome in addition to being a subjective process. Recently, Bartels et al. (2010) applied a support vector machine (SVM) algorithm to classify EOG and EMG components with supervised training. In addition, Mantini et al. (2007) proposed automatic identification of artifact components based on the distribution of correlation coefficients of all independent components with the reference signal that represents the artifact of interest.

A more serious limitation of the ICA approach is that it requires a large number of observations (data points) to generate stable independent components. The number of data points needed is typically kC^2 (Nolan et al., 2010) where k is a multiplier generally set to 25, as recommended in (Onton and Makeig, 2006), and C is the number of components. Given a 256-channel recording, for example, the amount of data required to find 256 independent components would be $25 \times 256^2 = 1{,}638{,}400$ time points. Even at 1000 samples/s sampling rate, this would require almost 30 min of data, during which there may be significant variations in both the artifact and the brain signal. These changes imply that the spatial stationarity assumption of ICA (Jung et al., 2000c) may be violated. The irony is that ICA then only works on the inadequate measurement set of low channel count EEG.

To address this high dimensional observation hunger (data requirement demand) as well as minimize the risk of violating the stationarity assumption associated with long recordings, PCA is often performed as a data reduction step, truncating the dimensionality of the data prior to ICA (Nolan et al., 2010; Kiviniemi et al., 2003; Haufe et al., 2014). As a rule of thumb, the minimum sample size of PCA is five samples per variable (five time points per channel in the case of EEG) (Gorsuch, 1983), which is much less than the kC^2 samples required for ICA. This is because PCA can be derived linearly by singular value decomposition of the correlation matrix, while ICA is derived with higher-order statistics that require more samples to achieve stable solutions. Although small data sets can be processed by using a PCA + ICA approach, truncation of dimensions risks losing important information in the data, if the information is not well captured by the orthogonal basis set defined by PCA.

In the present research, we aimed to improve and optimize the performance of ICA-based artifact removal approaches for high dimensional EEG data. We employed the "divide and conquer" strategy from computer science to address the high dimensionality challenge. Specifically, the divide and conquer approach that we propose employs a permutation resampling for identification matching (PRIM) strategy that can be applied to ICA (or other BSS methods) for EEG artifact removal. Instead of decomposing all channels simultaneously, or relying on PCA for data reduction, we divide the measurement channels into subspaces. Subsets of channels are randomly selected (with replacement) and ICA is conducted on each subset. The result is stable decomposition with smaller measurement sets, each of which is a random subset of the head surface topography captured by channel sub-sampling. The assumption of the method is that the true component structure, of both brain sources and artifacts, is captured for each subset, such that the artifact components can then be subtracted accurately from each subset.

To facilitate ease-of-use and remove subjective judgments associated with manual identification of artifact components, we also developed an automated method to clearly separate artifact and non-artifact ICA components. For each of the numerous components generated by PRIM, we calculate an artifact relevance index (ARI); this metric reflects a given component's similarity with the artifact template. The resulting ARI histogram highlights the difference between artifact and non-artifact components in the form of two peaks separated by a low valley. By fitting the ARI distribution with a cubic polynomial, the lowest point of the fitted curve is used to determine automatically the threshold that separates artifact from non-artifact components. We evaluate the performance of the proposed approach with eye blink removal, using both simulated EEG as well as real EEG signals and compare the results with the PCA + ICA method.

2. Materials and methods

2.1. The PRIM workflow

EEG data is acquired for a duration that can span minutes to days, depending on the nature of the study or clinical exam. PRIM operates on long data files using relatively small sequential segments, thus minimizing the risk that the stationarity assumption of ICA is violated. The workflow of the PRIM approach is shown in Fig. 1. Each segment is processed through the entire workflow independent of the other segments. Once all segments are processed, the complete data file can be reconstructed.