

Automated Artifact Detection in BrainStream

An Evaluation of An Online Eye and Muscle Artifact Detection Method

Danny Oude Bos

Human Media Interaction, EEMCS Faculty, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
d.oudebos@student.utwente.nl

ABSTRACT

Electroencephalography (EEG) is often used to acquire brain signals for input for brain-computer interaction (BCI) systems. Unfortunately, EEG is very susceptible to artifacts. In 2007, an automated artifact detection method was implemented at the Music Mind Machine (MMM) group of the Radboud University in Nijmegen for use in their online system.

This article describes a formal evaluation of this artifact detection method based on the expertise of two professionals who work with EEG. The result is a 1.00 AUC for eye artifacts and 0.99 AUC for muscle artifacts, with accuracies of 97% and 93% respectively.

Whether the statistical parameters used by the algorithm were personal or based on a grouped average did not significantly influence the performance results. Neither did using a special artifact session instead of a normal experiment session for determining these parameters. Using bipolar EOG may provide a slight advantage for detecting eye artifacts.

In a comparison with the offline artefact detection method from the Fieldtrip toolkit, the online version obtained both higher AUCs and higher accuracies.

Keywords

Electroencephalography (EEG), brain-computer interfacing (BCI), machine learning, artefact detection, physiological artefacts, eye artefacts (EA), muscle artefacts (MA)

1. INTRODUCTION

Brain activity can be analyzed by a computer so it can be used to control robots or software applications. This technology is called brain-computer interfacing or BCI [13]. For paralyzed patients, BCI can provide new means of interacting with the outside world [3, 8, 15], but BCI is also proven useful to treat patients via neurofeedback and to evaluate neurological diseases [9]. Besides, this technology may be interesting for healthy people where it can be used for example as a novel way of interacting with a game [10, 11].

At the moment the most-used method to obtain input for a BCI is EEG. Electrodes are mounted on the head to record the voltage differences that arise because of brain activity. EEG has a number of advantages over other methods: it is non-invasive (no implants are necessary), fast (reaction time), has a high temporal resolution (sample frequency), and is relatively cheap. It does not require big machines in a laboratory setting, and it is even possible to create wireless EEG head-sets. Unfortunately EEG has one important

drawback: its susceptibility to noise.

Noise coming from sources other than neuronal activity from the brain, create disturbances called artifacts in the recorded electrical activity. The recordings are obscured by these artifacts, which can influence the results of signal analysis and classification. According to Fatourechi *et al* "physiological artifacts, especially those generated by eye or body movements, remain a significant problem in the design of BCI systems" [4].

One way of dealing with artifacts is trying to avoid them. During experiments, the test subject is instructed to refrain from blinking, eye movement, and to stay relaxed in order to avoid muscle tension. Test subject instruction however is not without its drawbacks. The reflex tendency (e.g. in the case of eye blinking) and the following inhibition can be detected from the EEG. Secondly it presents an additional mental task for the subject which can influence the test results. Another issue is that these instructions can or will not always be adhered to, for example in the case of children or with very intensive tasks. Besides, some artifacts are caused by sources that cannot be controlled, like heart beats [4]. Therefore this practice cannot eliminate all the influence of artifacts.

What is left is some method of dealing with the artifacts that cannot be avoided. A common practice is visual screening: the researcher analyzes all the records visually for artifacts. This manual process is very labor intensive, and because of its subjectivity (the decision when the data is considered clean enough) it cannot yield consistent results [4]. This issue makes quantitative research impossible. Besides, in online systems, where time is of the essence, this method is just not practical.

1.1 Motivation

With the advance of BCI, the need for fast, automated artifact detection or removal has only grown. It can be used to remove contaminated brain activity from the data that is fed to the analysis and classification algorithms, or to inform the subject of the occurrence of certain artifacts which they can then try to avoid.

To address the problem of artifacts, an online detection function has been designed and implemented for BrainStream, the online system used by the MMM group of the Radboud University in Nijmegen [12]. The current algorithm focuses on eye artifacts (EA) and muscle artifacts (MA) specifically.

Although the system has been tested both *offline* (the processing happens after the experiment) and *online* (the processing takes place during the experiment), these tests were quite informal. A formal evaluation of the artifact detec-

tion method is necessary, so scientifically valid conclusions can be drawn about its performance and on the appropriate parameter selection methods.

1.2 Artifact Detection Algorithm

The method used for online artifact detection we used was derived from the offline z-value-based artifact detection function of the Fieldtrip Matlab toolkit. It is founded on the idea that artifacts are anomalies with more prominent amplitudes than general brain activity. Common experience with EEG shows this to be a generally valid assumption for eye and muscle artifacts. This is also demonstrated in Figure 1.

The processing steps are:

1. *Channel selection* limits the data to the channels closest to the artifact sources.
2. *Bandpass filtering* limits the data to the frequencies in which the artifacts are most dominant.
3. *Hilbert analytic amplitude* yields the envelope of the signal for each channel.
4. *Normalization* by calculating the z-scores for each channel.
5. *Summation* obtains one summed z-value for each moment in time by adding the z-scores of all selected channels, normalizing the sum by dividing it by the root of the number of channels summed.
6. *Threshold comparison* is done to determine if an artifact is detected.

Eye artifacts (EAs) are most dominant in special EOG electrodes. The electrodes used are positioned above and below the right eye and near the temples. EAs are seen in the lower frequency band, which is reflected in the standard Fieldtrip setting to bandpass the data to 1–15Hz for detection of this type of artifact [7].

Muscle artifacts (MAs) can be detected from especially positioned EMG electrodes. As these are not available in our datasets, electrode positions at the neck and temporal lobes were used. For MA detection, the standard Fieldtrip setting is to bandpass at 110–140Hz as these artifacts are high frequency. To speed up for online processing, the EEG data is downsampled by BrainStream to a 256Hz sample frequency. As the upper frequency cannot be higher than the Nyquist frequency as it cannot be accurately sampled, for the evaluation the bandpass is set to 110–128Hz.

To give each channel the same amount of influence, the amplitude distribution is normalized by calculating their z-scores. A z-score indicates the distance in standard deviations from the mean to the currently analyzed data sample.

$$z_{ch,t} = \frac{(x_{ch,t} - \mu_{ch})}{\sigma_{ch}}$$

where

$$\mu_{ch} = \frac{1}{N} \sum_{t=1}^N x_{ch,t}$$

with N the total number of time samples and

$$\sigma_{ch} = \sqrt{\frac{1}{N} \sum_{t=1}^N (x_{ch,t} - \mu_{ch})^2}$$

To obtain one value per time sample, the z-values are summed over each of the analyzed channels, and divided by the root of the number of channels taken into account for normalization.

$$zsum_t = \frac{\sum_{ch}^C z_{ch,t}}{\sqrt{C}}$$

with C the number of channels.

If the z-sum exceeds the predefined threshold then the data is said to be contaminated with an artifact at sample time t .

1.3 Related Work

Because artifact detection and removal are so important for EEG analysis, many different approaches have been developed and tested, each with its own advantages and disadvantages. This section describes some of this research.

Halder *et al* used blind source separation and independent component analysis for feature selection and support vector machines for classification to detect and remove EAs and MAs. They report a classification rate of above 90%, using 20-fold cross validation on the data set with labels obtained from visual screening by one expert [6].

Van de Velde *et al* evaluated a number of muscle artifact detection methods. One neurologist screened the data twice to judge intra-expert performance as well as the performance of the algorithms for muscle artifact detection. As the data was not split up in epochs, the performance was evaluated in the amount of time in the results matched the expert labeling in sensitivity (TP/P , see section on Performance Measure for interpretation) and specificity (TN/N). The absolute power of the high beta band resulted in the best discrimination with a sensitivity of 80% and specificity of 90% [14].

Similar to our method but not exactly the same, Moretti *et al* use statistical characteristics of the EEG to detect EA and MA artifacts. Two skilled electroencephalographers were asked to classify selected epochs for evaluation. The data sets were constructed from the epochs that had been assigned the same labels by both experts. 98% of the epochs with MAs were detected (TP/P), and 95% of EAs [9].

To detect blink artifacts, Bogacz *et al* compared three different neural classification algorithms. Labels for the EEG data were derived from the video recordings of the subject's faces. Where this information was not available, one domain expert screened the recordings. The neural network using back propagation as training method performed the best, resulting the smallest error percentage ($(FP + FN)/(P + N)$) of 1.40% [2].

1.4 Research Questions

The main objective of this research is to answer the following question:

- Q1. What is the performance of an online adaptation of z-based artifact detection for EA and MA?

As a performance comparison provides more meaning about how well an algorithm performs, the performance is compared to the results of the original offline method from Fieldtrip.

During the earlier informal evaluation accuracies of about 90% have been attained. Because of the nature of that previous evaluation, the actual performance is expected to be

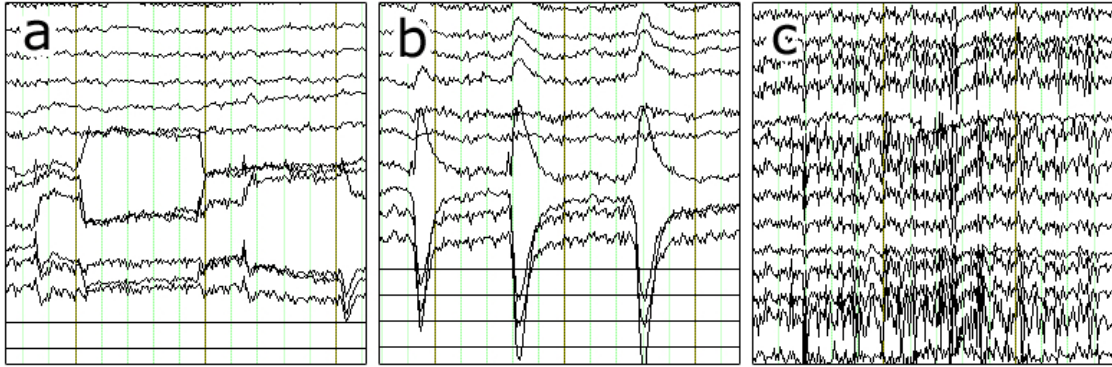


Figure 1: **Physiological artifacts in EEG.** a) eye movement, b) eye blink, c) muscle tension. Normal EEG, which has lower amplitudes, can be seen in the upper not-affected channels in (a).

a little lower. As both algorithms are based on the same method, the results are expected to be similar. However, there are some subtle differences which may work in the advantage of either implementation. See Section 2.5 for details.

Then there are some options for obtaining and selecting the algorithm parameters that we want to compare:

- Q2. When selecting the statistical parameters (μ , σ), what is the influence on the performance of choosing average, grouped parameters for all subjects versus individual parameters for each?

Individual parameters are expected to yield better results than grouped parameters as they are more personalized. However, individual parameters are also more vulnerable for anomalies during the training session.

- Q3. What is the influence on the performance of using the artifact session as training data compared to using an ordinary experiment session to obtain the statistical parameters (μ , σ)?

In the offline Fieldtrip implementation, the analyzed experiment itself is used to derive the parameters, which would be comparable to using an ordinary experiment. However, initial experiments show that the artifact session also shows a lot of promise. This artifact session deliberately contains many artifacts – see the description of the data sets in the next section for more information.

2. METHODOLOGY

This section details the methodology used to answer these research questions. It first describes the data sets used for this evaluation, and the processing steps that occur during an experiment. Then the chosen performance measure is explained, the different scenarios are detailed, to end with a description of the Fieldtrip comparison.

2.1 Data Sets

The data sets available are one artifact experiment session and one ‘other’ experiment session, from three different subjects. The ‘other’ experiment session is split up in two sets of 100 two-second epochs (3:20 min) where the first set is the training set and the second is the test set.

Data Acquisition

All the data sets were recorded at the MMM group in Nijmegen using the BrainStream platform and a 256-electrode set from BioSemi. During the artifact sessions, the subjects are asked to consciously create artifacts in the EEG. On cue, the subject blinks, moves the eyes, moves the head, breaths deeply, chews, coughs, smiles, and frowns. The whole session takes about five minutes.

The other experiments were recorded to test frequency tagging techniques. The subject gets to hear two modulated sounds and focuses on one of them.

Data Evaluation

The epochs from the training and test sets were labeled by two domain experts using the visual screening method, as this is still considered the gold standard in the case of artifact detection [16, 9].

Table 1 shows the number of artifacts detected by each expert for each data set. It shows the percentage of sightings, in 200 epochs per set, except for in the case of the consensus, where the percentage is based on just those epochs on which both experts agreed. Especially in set B1T this drastically reduced the number of epochs.

Expert 1	EA	MA	Other
A1T	47.0%	0.0%	2.0%
B1T	85.5%	97.0%	13.0%
C1T	50.5%	24.5%	01.5%
<i>Average</i>	<i>61.0%</i>	<i>40.5%</i>	<i>5.5%</i>
Expert 2	EA	MA	Other
A1T	45.5%	0.0%	0.5%
B1T	30.5%	9.0%	5.0%
C1T	45.5%	26.5%	0.5%
<i>Average</i>	<i>40.5%</i>	<i>11.8%</i>	<i>2.0%</i>
Consensus	EA	MA	Other
A1T	46.2%	0.0%	0.0%
B1T	68.2%	75.0%	3.9%
C1T	47.9%	24.5%	0.5%
<i>Average</i>	<i>54.1%</i>	<i>33.2%</i>	<i>1.5%</i>

Table 1: Expert Scoring Results

Dataset	EA	MA	Average
A1T	98.5%	100%	99.3%
B1T	44.0%	12.0%	28.0%
C1T	94.0%	96.0%	95.0%
Average	78.8%	69.3%	74.1%

Table 2: Inter-Expert Consensus

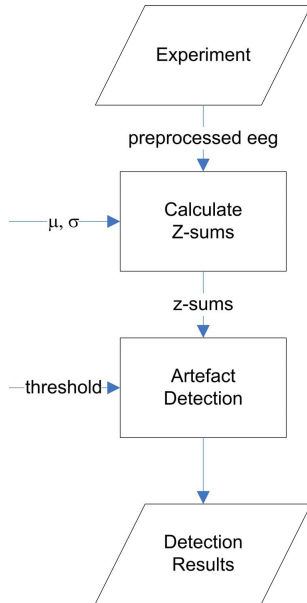


Figure 2: Artifact Detection Process.

The percentage of consensus between the two experts can be seen in Table 2. Overall, the domain experts agreed on 74.1% of the 600 screened epochs. However, the second data set was particularly problematic with a very low agreement rate of 28%. The average consensus for just the other sets is as high as 97.1%.

Based on these results, it was decided to leave data set B1T out of the evaluation. With a consensus rate of only 28% this data is not suitable material to evaluate a classifier with.

2.2 Experiments

The steps performed in the process of evaluating the artifact detection software are globally shown in Figure 2, with the inputs and outputs for the algorithm.

The EEG data from the experiment session is pre-processed by downsampling the data from 256 to 64 channels and to a 256Hz sample frequency. After this channel selection and bandpass filtering are performed for each type of artifact that will be detected.

The means and standard deviations for each of the selected channels are calculated based on the artifact or training data sets. Alternatively, one can use the means of the statistics over the different test subjects as more generalized values. These statistical parameters are necessary to be able to compute the z-sums for the training and test experiments.

Based on the the expert screening results and the z-sums calculated from the training set, an optimal threshold can be

determined at which the accuracy is highest. This threshold is then used to determine the actual artifact detection results of the test set. Again a general threshold can also be used, based on the average of the optimal thresholds over the test subjects.

2.3 Performance Measure

Now the classification results are known, the performance can be derived. It is common to express the results of a classifier compared to the labels of the ground truth in a *confusion matrix* giving four classes: *true positives* (TP) or ‘hits’, *true negatives* (TN) or ‘correct rejections’, *false positives* (FP) as ‘false alarms’, and *false negatives* (FN) also known as ‘misses’ [1].

A derived performance measure is *accuracy*, which expresses the correct classification rate $((TP + TN)/(P + N))$ where $P = TP + FN$, the number of actual positives, and $N = FP + TN$, the number of actual negatives). Because of the easy interpretation of accuracy, this performance measure is mentioned in the results.

The downside of accuracy is that the resulting value depends on the ratio of positives and negatives in the data set. An alternative measure to work around this problem is the *Area Under Curve* (AUC) of a *Receiver Operating Characteristic* (ROC) curve. To obtain this curve the *True Positive rate* (TPr, TP/P) is plotted against the *False Positive rate* (FPr, FP/N). For classifiers like this, a range of [FPr, TPr] points can be obtained by varying the classification threshold, resulting in a curve indicating the performance of the model. Ideally, the curve moves from [0,0] straight to point [0,1] where there are no FPs and all TPs to end in [1,1], resulting in an AUC of 1. So to interpret the AUC: a value closer to 1 is better [5].

To summarize, the performance measures used are accuracy, and AUC. Besides, the ROC curves are shown as they provide more information than just the area under curve value.

2.4 Scenarios

As mentioned previously, there are some options in what exact data to use for the artifact detection. These variations on the algorithm are detailed next.

Individual or Grouped Statistical Parameters

In the case of individual statistical parameters (μ and σ), both the data set on which the statistics are based and the test set are recorded from the same test subject. To provide better generalization and to make the system more robust to rare anomalies in the artifact or training data, one may decide to use statistical parameters averaged over the artifact or training sessions of *all* the test subjects.

Artifact or Experiment Training Session

Apart from the choice of using individual or grouped statistical parameters, there are also two alternatives for the source of these parameters. The first option is to use the special artifact session which is sure to contain all types of artifacts. The second is to use the training set which is part of the same experiment that is analyzed for artifacts. As a result, the parameters of the training set are more likely to match the test set.

2.5 Comparison with Fieldtrip

Another part of the evaluation of the online artifact detection method is comparison with the original offline Fieldtrip implementation. The preprocessed epochs will be fed to the algorithm, just as is done in the online version. To derive the ROC curves and AUC scores, the original code was slightly adjusted to output the z-sums as well.

Some main differences between the Fieldtrip method and the online version are:

- Fieldtrip uses the whole dataset to determine the statistical parameters. In this case that means both the training and test set. It was decided not to alter this, as this is how the method is intended to work.
- Fieldtrip adds filter padding to the epoch which is removed again after the bandpass filter is applied. The online adaptation also removes filter padding to avoid possible artifacts at the borders caused by the filter, but this padding has not been added in advance. The reason for this is the nature of online processing: you cannot add padding at the future end of an epoch, as that part of the EEG has not occurred yet.
- Fieldtrip also appends trial padding around the epoch in order to detect artifacts at the borders which may still have some influence in the processed epoch itself.

Because of the preprocessing, the frequency band Fieldtrip would normally use for MA detection is not valid, as the upper threshold of 140Hz is higher than the Nyquist frequency of the processed EEG. Therefore the upper threshold is lowered to 128Hz, and now matches the frequency band used online.

3. RESULTS

The results obtained with the just-described methodology are discussed in this section, with respect to the aforementioned research questions.

3.1 Online Performance

The average ROC plot for EA in Figure 3a is derived from the test set using the algorithm with the following settings: bipolar EOG, and individual statistical parameters based on an ordinary experiment session. It is averaged over the results for the experts plus consensus, and the two different data sets.

The average ROC curve for MA displayed in Figure 3c excludes some varieties in which no MAs were marked at all. In those cases (dataset A1T) there is no TPr defined as there are no actual positives. The gray crosses mark the horizontal and vertical standard deviations from the mean at those points.

Average AUC scores with these settings are, with two fractional digits of precision, 1.00 for EA and 0.99 for MA (ignoring the non-existent AUCs). Average [threshold, optimal accuracy] combinations obtained over the two data sets and the two experts plus consensus ground truths are: [0.65, 0.97] for EA and [3.19, 0.93] for MA. Although the AUC score seems perfect, this is a rounded value. Therefore, the not-perfect accuracy is not unexpected.

3.2 Fieldtrip Performance

Figures 3b and 3d show the average ROC curves for EA and MA detection by the offline Fieldtrip artifact detection method. Like with the ROCs for the online situation, the crosses indicate the standard deviations in FPr and TPr from the mean at those points.

Average AUCs with these settings are 0.68 for EA and 0.62 for MA. The average [threshold, optimal accuracy] pairs are [0.70, 0.80] for EA and [21.97, 0.82] for MA.

3.3 Scenarios

For the different options in the use of data, the results for the alternative pairs have been compared with t-tests. If not separately stated, the mentioned results hold true for both the training and test sets.

Individual or Grouped Statistical Parameters

The difference between using individual or grouped parameters was not significant for the AUC scores, nor the threshold, nor the corresponding accuracy. There was a trend towards a higher threshold for individual statistics for MA however ($t(46) = 1.94, p < 0.10$) with a 2.08 average for individual and 1.20 for grouped parameter values.

Artifact or Experiment Training Session

When comparing the results obtained with statistical parameters based on the artifact session versus the training session, no significant differences were seen for the AUC scores or accuracies.

For the thresholds however, the artifact session resulted in significantly lower values (EA: $t(46) = 7.58, p < 0.001$; MA: $t(46) = 7.36, p < 0.001$). The means for EA for the artifact session were about 0.17 and about 1.50 for the training session. For MA this was 0.47 versus 2.80.

4. DISCUSSION

Most of these results deserve a number of side notes or an attempt at an explanation.

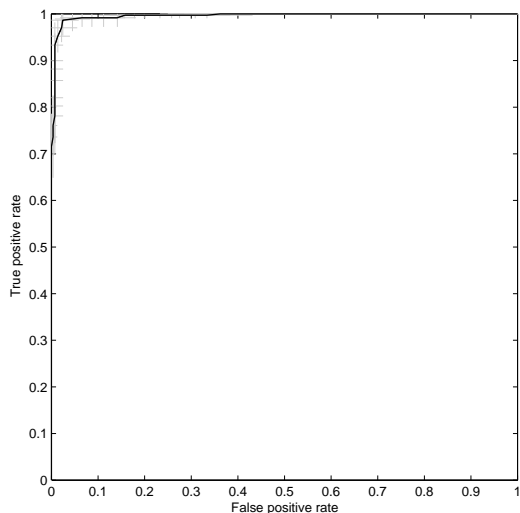
4.1 Online Performance

The AUC score for EA was higher than MA. It should be noted that the scores for MA are based solely on set C1T, as no MAs were marked in A1T. When there are no instances for one class, a ROC curve cannot be determined.

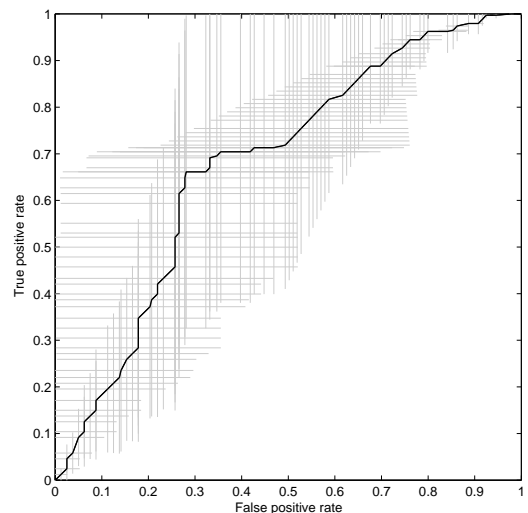
Furthermore, both experts remarked on having problems marking the MAs, as it is more difficult to draw a line between contamination and clean than it is for EAs. This is somewhat reflected in the higher standard deviations in Figure 3c showing the online MA detection performance.

The difference in threshold levels for EA and MA, can partially be explained by the use of bipolar electro-oculogram (EOG). Four electrodes were used to record the eye movements: two to the sides of the eyes for horizontal movement, and two above and below one eye to detect vertical eye movement plus eye blinks. For the EA detection it was chosen to use bipolar measurements: the two signals for horizontal movement were subtracted to obtain one bipolar value, and the same for the two signals for vertical movement. This procedure reduces the waves caused by brain activity, but as a result of the positions of the electrodes it amplifies the wave forms caused by eye movement and blinking.

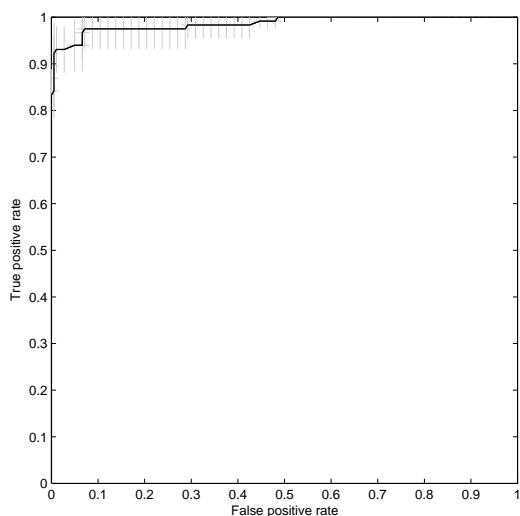
Another element that may have influenced the values for



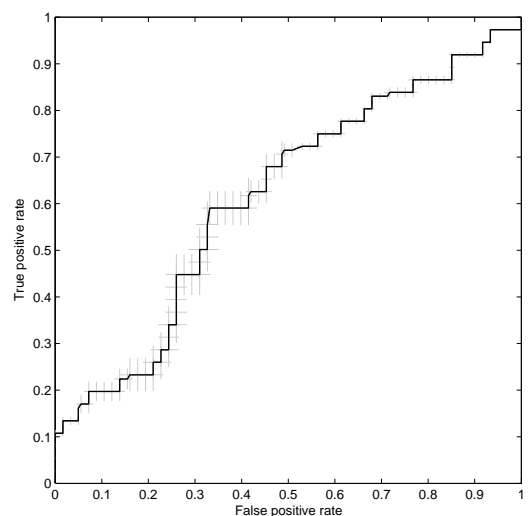
(a) ROC Online EA Detection



(b) ROC Fieldtrip EA Detection



(c) ROC Online MA Detection



(d) ROC Fieldtrip MA Detection

Figure 3: **Average ROC Curves.** A comparison between the performance of the online and offline artefact detection methods for eye and muscle artifacts.

the optimal thresholds is the fact that much more EAs have been scored by the experts than MAs. With an occurrence of only about 12%, it is profitable for the accuracy to select a high threshold. Even if no artefacts would have been detected at all, the accuracy would still be as high as 88%.

4.2 Comparison with Fieldtrip

The accuracies obtained by the offline Fieldtrip method are more than 10% lower than the accuracy values of the online version. The lower performance is also reflected by the ROC curves. The reason for this is not immediately clear, as both methods are based on the same algorithm. Nevertheless, there are some small differences that could have induced this difference.

The Fieldtrip artifact detection function uses all the 200 epochs (per data set) to base the statistical parameters on, whereas the online version only used the first 100 for training. As a result, the online implementation may provide

better generalization.

Secondly, the Fieldtrip method originally looks for muscle artifacts in the 110–140Hz band range. Because of the downsampling to 256Hz and the Nyquist frequency¹, Fieldtrip could only take into account 110–128. Therefore this performance may not be fully representative of its capabilities. Of course, the online implementation is subject to the same limitation.

But perhaps the most important difference is the fact that Fieldtrip adds filter padding and trial padding to the epochs, extending the time range it will look at for artifacts. Normally, human experts would probably do the same. In this case however, the epochs presented with the screening system were pieces of exactly those two seconds. Apart from moving to the previous and next epochs, there was no easy means provided to look at the EEG around the borders.

¹http://en.wikipedia.org/wiki/Nyquist_frequency

Another striking phenomenon is the high variation for the ROC curves for the EA detection with Fieldtrip shown in Figure 3b. None of the other ROC plots show this same feature. The reason for this high variation is as of yet unclear.

4.3 Scenarios

Based on the analysis using independent-samples t-tests, it can be concluded that it does not matter significantly to the resulting AUC scores whether individual parameters were used or grouped (average) parameters. Neither did it matter whether the statistical parameters were based on the artifact training set which deliberately contained many artifacts or on a normal experiment. Even the decision of using bipolar or unipolar EOG for EA detection did not influence the performance in a relevant way.

The thresholds were significantly lower when the statistical parameters were based on the artifact session instead of the training session. This is to be expected, as the artifact session will contain more artifacts and will therefore probably have a higher mean and higher standard deviations. As the training session values will be normalized with higher statistical parameters, the resulting z-sums will be lower, hence the lower threshold. This is also in line with the premise that EAs and MAs have higher amplitudes than brain activity.

Although not the main research of this article, a quick comparison between using bipolar EOG and unipolar EOG showed a slightly (not significantly) better performance for the bipolar variety. However, this difference was not significant.

5. CONCLUSIONS

This document described a formal evaluation of an online artifact detection method that was implemented recently for use in the Brainstream platform that is developed by the F.C. Donders Centre and the Music Mind Machine group of the Radboud University in Nijmegen.

As a ground truth the visual screenings of two domain experts have been obtained for three data sets of 200 epochs. Each of these data sets was then split into a training set and a test set of 100 epochs each. Except for one data set, the expert consensus was very high considering the subjectiveness of the task. There was a 99.3% agreement for set A1T and a consensus of 95.0% for C1T.

The performance of the online classifier (individual statistical parameters from training set, bipolar EOG) as indicated by the AUCs are (rounded) 1.00 for EA and 0.99 for MA detection. The average optimal accuracies are 97% for EAs and 93% for MAs, so it can be concluded that the classifier works quite well.

Compared to the performance obtained from the original Fieldtrip implementation, the online version seems to perform better. This is strange, as both methods are based on the same algorithm. At the moment it is not easy to discern what difference causes this incongruity.

General recommendations are to use a bipolar EOG instead of unipolar channels (although the difference was not shown significant). A normal training session to determine the statistical parameters is probably more practical, as this way no special artifact session is required.

Also, if for future work one would be interested in updating the statistical parameters of the classifier with new data from the currently running experiment, it is interesting to

know that statistics from a normal experiment will do.

For this purpose, the fact that there is no significant difference in performance when using grouped statistical parameters instead of individual ones is also quite valuable. It is even possible to skip a training session altogether if a set of grouped parameters is already available.

6. FUTURE WORK

In line with the already mentioned idea of an online updating classifier, and the use of grouped parameters to start with, it is important to research the influence of time and subjects on these parameters.

It could also be worthwhile to find out what causes the performance drop in the offline Fieldtrip method, and what causes the high standard deviations in the EA detection curve.

And then there is the issue of dataset BIT. For this evaluation, this dataset has not been used because of the little amount of consensus. Analysis including this set showed a greatly reduced performance of the classifier. It could be interesting to see what exactly was the cause of this disagreement between the experts, and what would be the proper way to deal with it.

For other research groups to continue to design and evaluate artifact detection or perhaps removal methods: the datasets are available on request. Contact the author if you are interested.

7. ACKNOWLEDGEMENTS

I'd like to thank Philip van den Broek and Mannes Poel for their advice and guidance during this project, and Peter Desain and Anton Nijholt for their support. The advice of Kim Verhoef, Christian Mühl, and Rebecca Schaefer concerning the Online Screening System has been invaluable. And last but not least I'd like to thank the domain experts who screened the data as without their help there would have been no evaluation.

8. REFERENCES

- [1] E. Alpaydin. *Introduction To Machine Learning*. MIT Press, 2004.
- [2] R. Bogacz, U. Markowska-Kaczmar, and A. Kozik. Blinking artefact recognition in EEG signal using artificial neural network. *Proc. of 4th Conference on Neural Networks and Their Applications*, 4:502–507, 1999.
- [3] L. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and clinical neurophysiology*, 70(6):510–523, 1988.
- [4] M. Fatourehchi, A. Bashashati, R. Ward, and G. Birch. EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, 118(3):480–494, 2007.
- [5] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.
- [6] S. Halder, M. Bensch, J. Mellinger, M. Bogdan, A. Kübler, N. Birbaumer, and W. Rosenstiel. Online artifact removal for brain-computer interfaces using support vector machines and blind source separation. *Intell. Neuroscience*, 2007(2):1–9, 2007.

- [7] V. Krishnaveni, S. Jayaraman, P. Kumar, K. Shivakumar, and K. Ramadoss. Comparison of Independent Component Analysis Algorithms for removal of ocular artifacts from Electroencephalogram. *Measurement Science Review Journal*, 5:67–79, 2005.
- [8] R. Leeb. Self-Paced (Asynchronous) BCI Control of a Wheelchair in Virtual Environments: A Case Study with a Tetraplegic. *Computational Intelligence and Neuroscience*, 2007:1–8, 2007.
- [9] D. Moretti, F. Babiloni, F. Carducci, F. Cincotti, E. Remondini, P. Rossini, S. Salinari, and C. Babiloni. Computerized processing of eeg–eog–emg artifacts for multicentric studies in eeg oscillations and event-related potentials. *International Journal on Psychophysiology*, 47(3):199–216, 2003.
- [10] A. Nijholt, D. Tan, B. Allison, J. del R. Millán, and B. Graimann. Brain-computer interfaces for hci and games. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3925–3928, New York, NY, USA, 2008. ACM.
- [11] A. Nijholt, D. Tan, G. Pfurtscheller, C. Brunner, J. Millán, B. Allison, B. Graimann, F. Popescu, B. Blankertz, and K. Müller. Brain-Computer Interfacing for Intelligent Systems. *IEEE Intell. Systems*, 23(3):72–79, 2008.
- [12] D. Oude Bos. Automated EEG Artefact Detection and Removal. Internship Report, 2007.
- [13] T. J. Sejnowski, G. Dornhege, J. del R. Millán, T. Hinterberger, D. J. McFarland, and K.-R. Müller. *Toward Brain-Computer Interfacing (Neural Information Processing)*. The MIT Press, 2007.
- [14] M. van de Velde, G. van Erp, and P. Cluitmans. Detection of muscle artefact in the normal human awake EEG. *Electroencephalography and Clinical Neurophysiology*, 107(2):149–158, 1998.
- [15] S. Wills and D. MacKay. DASHER: An Efficient Writing System for Brain-Computer Interfaces? *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 14(2):244, 2006.
- [16] J. Wu, E. Ifeachor, E. Allen, S. Wimalaratna, and N. Hudson. Intelligent artefact identification in electroencephalographysignal processing. *Science, Measurement and Technology, IEE Proceedings-*, 144(5):193–201, 1997.