Implication of non-stationarity in single-trial detection performance of event-related potentials

H. Cecotti[1] and A.J. Ries[2]

Abstract— **The electroencephalographic (EEG) signal is known to fluctuate over time due to ongoing brain activity related to various tasks that a subject can do or think of. For this reason, it is typically expected in Brain-Computer Interface (BCI) that the extracted brain responses will be non-stationary. The non-stationarity of the EEG signal can have an impact on the performance of the system during long sessions. In addition, BCI systems should aim at reducing the calibration procedure or include the calibration stage during the test phase in an invisible manner. In this paper, we propose to evaluate through different cross-validation approaches to what extent the non-stationarity of the EEG signal has an impact on single-trial detection, and if this effect can be taken into consideration for optimizing the design of BCI based on event-related potentials detection with applications for the triage of images during rapid serial visual presentation (RSVPs) tasks. We use the data obtained from sixteen healthy subjects performing an RSVP task where participants had to count a particular class of images to evaluate single-trial detection performance. The results support the conclusion that the cross-validation technique, i.e. the order of the examples in the training database, has an impact on the performance, and that existing labeled trials that are set regularly during the test phase can provide a novel way to avoid a calibration procedure in particular BCI settings.**

I. Introduction

Non-invasive Brain-Computer Interface (BCI) systems embed various applications that include the detection and analysis of brain signals. A current trend in BCI is the development of a holistic approach that puts the user in the center of the system in order to satisfy his needs. The increase of hybrid BCIs that combine several brain responses sequentially or in parallel [1], and the development of virtual keyboards that take into account the specificities of the required visual stimuli to evoked brain responses show how different elements should be combined to enhance the users experience [2]–[4]. This research direction stays faithful to the pioneer work in BCI that aimed at helping disabled people who have BCI systems as their only means of communication with the world [5], [6].

Regular BCI applications such as virtual spellers require a voluntary control and a relatively fast online feedback that does not allow a direct transductive approach. In fact, for BCI based on motor imagery detection, the feedback of the detection is typically translated directly into a command. For the P300 speller, the binary classification of event-related potentials is not directly translated as a command, but the binary decisions of several ERPs are combined, and sorted, to obtain a decision in the application. The decision of the P300 speller includes usually the single-trial detection scores from different repetitions of different visual stimuli displayed on a computer screen (e.g. a flashing row, a flashing column) [7]. Despite the use of the detection of up to 120 scores in a P300 speller (6 rows, 6 columns, and 10 repetitions), a decision based on a pure transductive approach is not used because the number of trials remain too low to cluster targets (20 trials) from non-target trials (100 trials). However, the decision that occurs with the P300 speller may be considered as transductive as it doesn't correspond exactly to a pure binary classification of ERPs, but to the output of a sorted list of binary classification scores, i.e. the number of target and non-target is known a priori. BCI based on ERPs detection during RSVP tasks that are used for the triage of images do not belong to the group of applications that can be considered as human-computer interface [8], [9]. In fact, RSVP tasks for the triage of images take advantage of the human visual process system. The subject is only asked to pay attention to a particular type of images. The detection of an image may not be directly translated into a command. Therefore, there is an absence of voluntary control, and the user is used as an advanced visual processing unit.

Contrary to gaze-independent virtual keyboards based on RSVP tasks where the visual stimuli are known, and the detection of a brain evoked response after the presentation of a visual stimulus can be assigned to a command, the goal of BCI RSVP tasks for the triage of images is to determine the class of the visual stimuli [10]. Because the goal of the application is to sort images, and a direct feedback may not be necessary, it is possible to process a large number of images within a single session. Moreover, contrary to other BCI applications where the calibration session has to be set before the test session, BCIs for the triage of images do not require to fix the calibration session before the test session. Brain responses vary over time due to fluctuation of the attention and other ongoing brain activity. The spatial distribution and the characteristics of the recorded brain evoked responses can change over time. When a short calibration is set, it is assumed that the recorded signal does not capture all the variability of the signal compared to different recorded signals. For this reason, we propose to investigate the choice of the evaluation procedure. Particularly, the order of the trials that are used in a session for training the classifier can have an impact on single-trial performance. The remainder of the paper is organized as follows. First, we present the experimental protocol. Second, we describe

[1] School of Computing and Intelligent Systems, Ulster University, Magee Campus, Derry, Northern Ireland, UK. h.cecotti@ulster.ac.uk
[2] Human Research and Engineering Directorate, US Army Research Laboratory, Aberdeen Proving Ground, MD 21005, USA.

the signal processing and classification methods with the evaluation techniques. Finally, the results are presented and discussed in the last two sections.

II. Experimental protocol

A. Subjects

The data corresponds to 16 healthy volunteer participants (33.5 years, 13 males, 15 right handed). They provided written informed consent, reported normal or corrected-to-normal vision and reported no history of neurological problems. The voluntary, fully informed consent of the persons used in this research was obtained as required by federal and Army regulations. The investigator has adhered to Army policies for the protection of human subjects [11], [12].

B. Visual stimuli and procedure

Participants were seated 75 cm from a Dell P2210 monitor, and they viewed a series of simulated images from a desert metropolitan environment in a rapid serial visual presentation (RSVP) paradigm (Fig. 1(a)). Images (960x600 pixels, 96 dpi, subtending 36.3 x 22.5) were presented using E-prime software on a Dell Precision T7400 PC. Images were presented for 500 ms (2 hz) with no inter-stimulus interval. Images contained either a scene without any people (non-target) or a scene with a person holding a gun (target). A total number of 110 target images and 1346 non-target images were presented to each participant. Scenes in which a target appeared were also presented without the person in the non-target condition. All stimuli appeared within 6.5 degrees of center of the monitor. The goal of the task was to classify target images from non-target images. Behavioral analysis was conducted on a session in which subjects responded to target stimuli by pressing a key while also counting the number of target images. Single-trial detection was conducted on a second session in which the subjects had only to count the number of target images.

C. Signal acquisition

Electrophysiological recordings were digitally sampled at 1024 Hz from 64 scalp electrodes arranged in a 10-10 montage using a BioSemi Active Two system (Amsterdam, Netherlands). Impedances were kept below 25 $k\Omega$. External leads were placed on the outer canthus of both eyes and above and below the right orbital fossa to record electrooculogram.

III. Methods

A. Cross-validation

A large number of cross-validation (CV) techniques are available, and the type of CV can have a significant impact on the classification results [13]. CV is a model validation technique for assessing how the results of a classifier will generalize to a new independent data set. It is principally used for classification and prediction, and then the goal is to estimate how accurately a classifier will perform in practice. In a supervised classifier, a model is usually given a data set of labeled data on which training is run, and a data set of unlabeled data against which the model is tested. Exhaustive CV are typically not used because it is computationally expensive to learn and test a classifier on all possible ways to divide the original data set into a training and a validation set. Hence, non-exhaustive CV approaches are used because they do not compute all ways of splitting the original data set. Leave-one-out cross-validation involves using a single data point as the validation set, and the remaining data as the training set. This type of CV may not be appropriate for the evaluation of BCI systems because the signals that are before and after the signal that is tested, are used during training, allowing the classifier to better capture the variability of the signal over time. Because the signal is assumed to be non-stationary with fluctuations due to the subject's attention on the task, and the current subject's fatigue level, leave-one-out CV may not represent a realistic estimation of the classifier performance. In k-fold cross-validation, the original data set is partitioned into k equal size subsets. With the k subsets, two evaluations can be performed. First, a single subset is retained as the validation data for testing the model, and the remaining $k1$ subsets are used as training data. Second, a single subset is used for training the classifier, and the remaining $k-1$ subsets are used for the test. The k-fold CV leads to the evaluation of k classifiers. However, only the case where the training data precedes the test data corresponds to the online evaluation of a BCI.

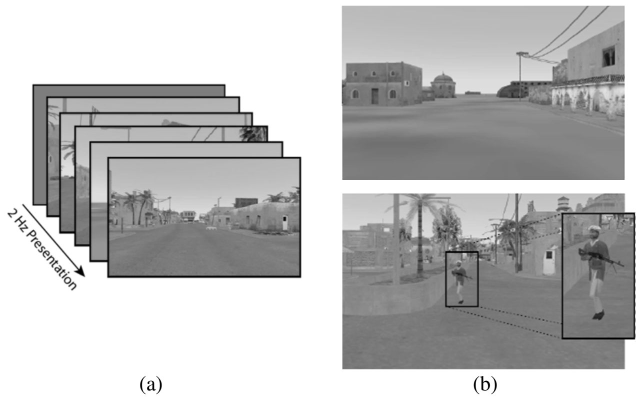

Fig. 1. **(a)** Rapid Serial Visual Presentation task. **(b)** Representative examples of stimuli on target (**bottom**) and non-target trials (**top**). The inset showing a target is for illustration purposes only, it did not appear in actual stimuli.

For k-fold CV evaluation, the data can be split into k subsets of same size in different ways: first the k blocks are contiguous, second the k blocks are randomly chosen, and third the k blocks are split into two blocks, the first sub-block is $1/k$ of the size of the block, and the second sub-block is $(k-1)/k$ the size of the block. In each case, the distribution of the classes is kept identical to the complete database.

By comparing and evaluating several combinations of data subsets, we want to estimate to what extent the non-stationarity of the EEG signal has an impact in the classification performance. Indeed, it is expected that test data that is surrounded by training data will provide a better performance than test data that is isolated from the training data. In addition, because RSVP tasks can be used to sort images, predefined images with known labels can be placed in the list of images that are presented to the user. These images can then be used as part of the training database, replacing a calibration session. Hence, the evaluation of different CV procedures aims also at optimizing the place of predefined images during the presentation of images.

B. Signal processing and classification

To enhance and reduce the number of discriminating components, the EEG signal was first bandpass filtered using a 4^{th} order Butterworth filters [1-42.66 Hz], and then downsampled by a factor of 8. After preprocessing, the signal was epoched from stimulus onset to 640 ms after stimulus onset for subsequent analysis. The next step consisted of enhancing the relevant signal using the xDAWN spatial filtering approach [14], [15]. In this method, spatial filters are obtained through the Rayleigh quotient by maximizing the signal-to-signal plus noise ratio (SSNR), where the signal corresponds to the information contained in the ERPs corresponding to the presentation of a target [16], [17]. The first four spatial filters generated by xDAWN were used as inputs for the classification ($N_f = 4$). Bayesian linear discriminant analysis (BLDA) [18], [19] was used for the binary classification of the brain evoked responses corresponding to the presentation of target versus non-target images. Artificial trials based on shifted in time examples were added for training the classifier [20], [21].

Performance was evaluated across different subsets for training and the test. In conditions A_1, A_2, A_3, B_1, B_2, and B_3, the database is cut into 10 contiguous blocks of the same size, i.e. the data in block i were recorded before the data in block $i + 1$. In condition A_1, the first block is used for the test, the nine remaining blocks are used for training. In A_2, the last block is used for the test, the nine remaining blocks are used for training. In A_3, the fifth block is used for the test (the block in the middle of the experiment). In B_1, the first block is used for training, the nine remaining blocks are used for the test. In B_2, the last block is used for the test. In B_3, the fifth block is used for the test (the block in the middle of the experiment). In condition R, the database is cut into 10 blocks of the same size of the examples randomly selected over time while keeping the class distribution stable. In condition S, the examples are for training, and selected regularly over time in the test. The database is cut into 10 contiguous blocks, and in each block the first tenth is used for training, and the rest is used for the test. Hence, the training data is sampled regularly across the whole database. The different procedures are presented in Fig. 2. In the subsequent sections, performance is assessed by using the area under the receiver-operator characteristic (ROC) curve (AUC) [22].

IV. Results

Single-trial detection performance is depicted in Figure 3. The AUC for the conditions A_1, A_2, and A_3 is 0.973 ± 0.027, 0.924 ± 0.069, and 0.954 ± 0.050, respectively. Post-hoc analysis with Wilcoxon signed-rank test, with a Bonferroni correction, revealed a significant difference was observed between A_1 and A_2 (p=0.018), and between A_2 and A_3 (p=0.025). These results show a decrease of performance when the block corresponding to the test occurs at the end of the experiment. The AUC for the conditions B_1, B_2, and B_3 is 0.870 ± 0.068, 0.874 ± 0.070, and 0.872 ± 0.064, respectively. In these conditions, there was no significant difference. The performance of condition R is 0.837 ± 0.075. For condition S, the AUC is 0.899 ± 0.059, which is significantly higher than the other approaches using the same number of examples for training and the test (p<10e-2). This result proves that the position of the trials used for training has a significant effect on single-trial detection performance, translating variations of the features over time.

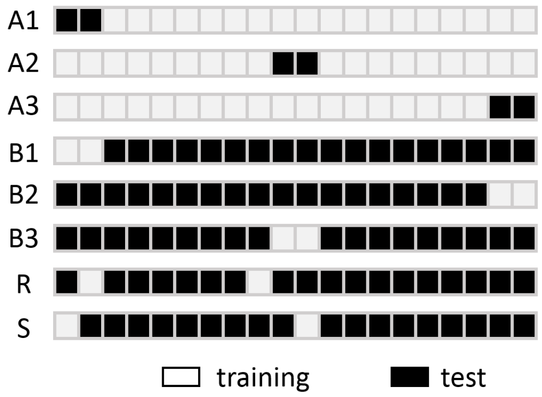

Fig. 2. Graphical representation of the performance evaluation procedures (examples with 10 blocks).

The information transfer rate (ITR) in bits per minute (bpm) is defined by $ITR = \frac{60}{T} \cdot \psi$ where ψ, the information transfer rate, in bits per symbol, is defined by:

$$\psi = \vartheta_0 - \vartheta_1 \tag{1}$$

$$\vartheta_0 = -\sum_{j=1}^{N_{out}} p(w_j) \cdot log_2(p(w_j)) \tag{2}$$

$$\vartheta_1 = -\sum_{i=1}^{N_{out}} \sum_{j=1}^{N_{out}} p(w_i) \cdot p(w_j|w_i) \cdot log_2(p(w_j|w_i)) \tag{3}$$

N_{out} is the number of classes, and T is the time in seconds of recorded EEG signal that is required to take the decision among the N_{out} outputs. In this case, T includes the time required to record all the trials (test and training) as they are merged in the same database. Due to the low target probability, we considered the Nykopp definition of the ITR, and $N_{out} = 2$. $p(w_j|w_i)$ being the element (i, j) in the confusion matrix of the classification obtained with a threshold set to maximize the f-score in the training database. The average ITR was estimated to 33.19 ± 0.40 bits/minute with condition (S).

V. Discussion and conclusion

In this paper, we have shown that the type of evaluation procedure has a significant impact on the classification performance. Results using cross validation evaluation report

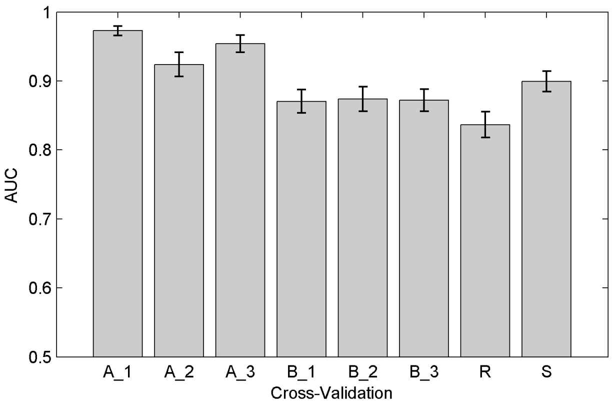

Fig. 3. Average AUC across subjects for each performance evaluation procedure. The error-bars represent the standard error.

only the mean and standard error, we have shown that there exists a significant difference of performance across the different blocks in a cross-validation, suggesting that performance does degrade over time, hence confirming a change in the data. First, this effect should be carefully taken into account when reporting results when the features of a signal are highly dependent of the signal processed in the past. Second, the difference of performance observed over time can be exploited online by providing predefined examples over time that will be used to train the classifier.

In addition to the low accuracy and problems related to the portability of the EEG recording devices [23], BCI systems should also aim at improving the usability aspect by removing the cumbersome calibration sessions. This research work is relevant because it aims at removing the calibration step that is required in BCI based on the detection of event-related potentials by integrating directly the required training data during the test stage. We have shown that by considering only 10% of the data, with the addition of artificial examples for training classifiers, it is possible to obtain an information transfer rate of 33 bits/minute. When the calibration session and the test session are merged, the user does not feel a gap between the calibration of the system and the test. Furthermore, by spreading predefined labeled training trials across the whole test session, it allows to remove issues related to the non-stationarity of the brain evoked responses as the predefined trials are present along the experiment, and they are not put at the beginning of the session such as in traditional calibration procedures. Further works will be carried out to investigate the robustness of the approach in sessions that last several hours.

Acknowledgment

The data were collected under U.S. Army Prime Contract No. W911NF-09-D-0001. This project was also supported in part by the Office of the Secretary of Defense ARPI program MIPR DWAM31168. H.C. would like to thank the NI Functional Brain Mapping Facility project (1303/101154803) funded by InvestNI and Ulster University.

References

[1] G. Pfurtscheller, B. Z. Allison, G. Bauernfeind, C. Brunner, T. S. Escalante, R. Scherer, T. O. Zander, G. Mueller-Putz, C. Neuper, and N. Birbaumer, "The hybrid BCI," *Frontiers in Neuroscience*, vol. 4, no. 42, pp. 1–11, 2010.

[2] L. Acqualagna, M. S. Treder, M. Schreuder, and B. Blankertz, "A novel brain-computer interface based on the rapid serial visual presentation paradigm," in *Proc. of the IEEE Annual International Conference of Eng. in Medicine and Biology Society (EMBC)*, 2010, pp. 2686–2689.

[3] H. Cecotti, "A self-paced and calibration-less SSVEP based brain-computer interface speller," *IEEE Trans. on Neural Systems and Rehab. Eng.*, vol. 18, pp. 127–133, 2010.

[4] J. Williamson, R. Murray-Smith, B. Blankertz, M. Krauledat, and K.-R. Müller, "Designing for uncertain, asymmetric control: Interaction design for braincomputer interfaces," *Int. J. Human-Computer Studies*, vol. 67, pp. 827–841, 2009.

[5] N. Birbaumer and L. G. Cohen, "Brain-computer interfaces: communication and restoration of movement in paralysis," *Journal of Physiology-London*, vol. 579, no. 3, pp. 621–636, 2007.

[6] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan, "Brain-computer interfaces for communication and control," *Clin Neurophysiol*, vol. 113, pp. 767–791, 2002.

[7] L. Farwell and E. Donchin, "Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials," *Electroencephalogr. Clin. Neurophysiol.*, vol. 70, pp. 510–523, 1988.

[8] A. Gerson, L. Parra, and P. Sajda, "Cortically-coupled computer vision for rapid image search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 174–179, 2006.

[9] E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S. Chang, and P. Sajda, "Closing the loop in cortically-coupled computer vision: a brain-computer interface for searching image databases," *J. Neural Eng.*, vol. 8, p. 036025, 2011.

[10] H. Cecotti, M. P. Eckstein, and B. Giesbrecht, "Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering," *IEEE Trans. Neural Networks and Learning Systems*, vol. 15, pp. 2030–42, Nov. 2014.

[11] "Use of volunteers as subjects of research," U.S. Department of the Army, Washington, DC: Government Printing Office, Tech. Rep. AR 70-25, 1990.

[12] "Code of federal regulations, protection of human subjects," U.S Department of Defense Office of the Secretary of Defense, Washington, DC: Government Printing Office, Tech. Rep. 32 CFR 219, 1999.

[13] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40–79, 2010.

[14] B. Rivet, H. Cecotti, E. Maby, and J. Mattout, "Impact of spatial filters during sensor selection in a visual P300 brain-computer interface," *Brain Topography*, vol. 12, no. 1, pp. 55–63, 2012.

[15] H. Cecotti, B. Rivet, M. Congedo, C. Jutten, O. Bertrand, E. Maby, and J. Mattout, "A robust sensor selection method for P300 brain-computer interfaces," *J. Neural Eng.*, vol. 8, p. 016001, 2011.

[16] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "xDAWN algorithm to enhance evoked potentials: application to brain-computer interface," *IEEE Trans Biomed. Eng.*, vol. 56, no. 8, pp. 2035–43, 2009.

[17] B. Rivet and A. Souloumiac, "Optimal linear spatial filters for event-related potentials based on a spatio-temporal model: Asymptotical performance analysis," *Signal Processing*, vol. 93, no. 2, pp. 387–398, 2013.

[18] D. J. C. MacKay, "Bayesian interpolation," *Neural Comput.*, vol. 4, no. 3, pp. 415–447, 1992.

[19] U. Hoffmann, J. Vesin, K. Diserens, and T. Ebrahimi, "An efficient P300-based brain-computer interface for disabled subjects," *Journal of Neuroscience Methods*, vol. 167, no. 1, pp. 115–125, 2008.

[20] H. Cecotti and B. Rivet, "Improving single-trial detection of event-related potentials through artificial deformed signals," in *Proc. of the 36nd Int. IEEE Conf. of the EMBC*, 2014, pp. 1–4.

[21] H. Cecotti, "Toward shift invariant detection of event-related potentials in non-invasive brain-computer interface," *Pattern Recognition Letters (in press)*, in press.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006.

[23] A. J. Ries, J. Touryan, J. Vettel, K. McDowell, and W. D. Hairston, "A comparison of electroencephalography signals acquired from conventional and mobile systems," *J. of Neuroscience and Neuroengineering*, vol. 3, no. 1, pp. 10–20, 2014.