NIRS-Based BCIs: Reliability and Challenges

Megan Strait and Matthias Scheutz

Tufts University, Medford MA 02155, USA

Abstract. Previously we contributed to the development of a brain-computer interface (Brainput) using functional near infrared spectroscopy (NIRS). This NIRS-based BCI was designed to improve performance on a human-robot team task by dynamically adapting a robot's autonomy based on the person's multitasking state. Two multitasking states (corresponding to low and high workload) were monitored in real-time using an SVM-based model of the person's hemodynamic activity in the prefrontal cortex. In the initial evaluation of Brainputs efficacy, the NIRS-based adaptivity was found to significantly improve performance on the human-robot team task (from a baseline success rate of 45% to a rate of 82%). However, failure to find any performance improvements in an extension of the original evaluation prompted a reinvestigation of the system via: (1) a reanalysis of Brainput's signal processing on a larger NIRS dataset and (2) a placebo-controlled replication using random (instead of NIRS-based) state classifications [1].

The reinvestigation revealed confounds responsible for the original performance improvements and underscored several challenges for NIRS-based BCIs in general. Specifically, it revealed the original performance improvements were due to a disparity in difficulty between experimental conditions of the original evaluation (i.e., the task being easier in the adaptive versus the baseline condition). Moreover, the reinvestigation showed Brainputs model of user multitasking (trained on the n-back task) generalized to neither the human-robot team task (the classifications showed systemic violations of basic hemodynamic principles) nor to other workload-inducing tasks (classifications of brain activity while users performed arithmetic were better than chance for only 1/4 of the subject population). Hence, in in an effort to identify ways forward, we first summarize the methods and results of this reinvestigation and then explore the challenges for achieving more reliable NIRS-BCIs.

1 Introduction

Brain-computer interfaces (BCIs) have been gaining traction in the field of human-computer interaction (HCI) for various application domains (e.g., [2,3]). Functional near infrared spectroscopy (NIRS, also referred to as fNIRS and fNIR), in particular, has been described as a suitable modality for BCIs given that it is relatively portable and reasonably robust to user movement [2]. Despite these useful characteristics, however, the reliability of these NIRS-based BCIs remains relatively unexplored [4,1]. Hence, here we summarize a reinvestigation

C. Stephanidis (Ed.): HCII 2014 Posters, Part I, CCIS 434, pp. 463–468, 2014.

of a NIRS-based passive BCI system, Brainput [5], of whose development the authors were part. Brainput was designed to react to fluctuations in cognitive workload by adapting a robot's level of autonomy, and in its initial evaluation, the passive NIRS-based adaptivity was found to significantly improve performance on a human-robot team task. However, when we attempted a follow-up extension, we did not find the improvements that we had observed originally. Hence, we set out to systematically evaluate the reliability of NIRS-based BCIs through a two-part reinvestigation of Brainput (first, via a reanalysis of Brainput's signal processing on a large NIRS dataset, followed by a placebo-controlled replication using random state classifications) which revealed confounds in the original study responsible for the initial performance improvements [1]. The goal of this experience report is to illustrate, in addition to the utility of the replication, some of the challenges and limitations involved as well.

2 Motivation

We previously participated in the development of the NIRS-BCI, Brainput, with the aim of classifying a user's multitasking state during a human-robot team task [5]. We hypothesized that brain-based adaptive autonomy would improve task performance. A two-probe NIRS instrument was used to image subjects' (N=11) prefrontal cortex while subjects explored a simulated environment with two robots to find a target location. The task ran until both robots found the goal location (success), or until five minutes had elapsed (failure). During the task, Brainput used classification of subjects' NIRS data to dynamically adapt the autonomy of one of the robots according to the subject's level of workload. The initial results showed that the brain-based adaptivity substantially improved performance (82% of subjects successfully completed the task) versus a baseline (no adaptivity) of 45%. Moreover, *mal-adaptive autonomy* (enabling of autonomy during low workload) caused performance to significantly worsen (18% success), indicating that the autonomy must be appropriately timed for it to be effective in human-robot teams. Inspired by these initial findings, we employed Brainput in an extension of the original protocol to test whether the performance improvements would persist with real robots (versus simulated robots in simulated environments as used originally). We replaced one of the two previously-simulated robots with a real robot in a real environment, but both were still controlled by the system architecture used in [5]. Here we expected to see the same performance improvements as we did originally; however, the results showed *no* performance improvement (for either simulated or real robot) from baseline performance. Given the nearly identical setups, these results suggested some degree of unreliability of the Brainput system.

3 Reinvestigation

Suspecting the Brainput classifier to have limited extensibility to larger populations (i.e., N=24 in the extension versus N=11 in the original investigation),

we conducted a reanalysis of Brainput's signal processing using a larger NIRS dataset, followed by a placebo-controlled replication of the original study [1] – the methods and results of which we summarize here.

Reanalysis of signal processing. Here we investigated two questions: (1) whether the original performance improvements persist over a larger sample and (2) whether the improvements generalize across variants of the same type of task. We obtained a larger NIRS datasest (N=40) consisting of low and high workload PFC samples induced by an arithmetic-based variant of the n-back task [4]. We preprocessed the data and trained the Brainput classifier on samples of each workload class (low, high), mirroring the procedure we used originally (see [5]), and then ran ten-fold cross-validation to predict model performance. Between subjects, the average classification accuracy was 54.5% ($SD = 14.3\%$). While the overall accuracy was statistically significant, only 10/40 subjects showed individual accuracies significantly above chance [1]. Moreover, Brainput's performance on this dataset was substantially lower than than what was found originally (68.4%). This discrepancy may have been due to the differences either in sample size (N=40 vs. N=3) and/or task (numeric vs. alphameric). Regardless, the lower overall performance was particularly worriesome in that the human-robot team task differed substantially from the alphameric task used to train the classifier in the original evaluation [5]. That is, if the classification schema does not extend well to a variant of the same type of task (numeric vs. alphameric n-back task), then it suggests that it might not generalize to more realistic applications (such as human-robot interaction tasks).

As Brainput's performance on the novel dataset differed so substantially from its preliminary evaluation, we revisited the original dataset (from [5]) to investigate Brainput's behavior during the human-robot team task. Using the logs of the realtime classifications, we constructed plots of the robot's autonomous behavior (autonomy-disabled vs. autonomy-enabled) over the course of the task (see Figure 1). Enablement of the robot's autonomy corresponded to classification of the NIRS data as *branching* (high workload), whereas disablement indicated the participant was experiencing *low* workload. Here we expected the behavior to show prolonged periods of autonomy enabled/disabled, but we found instead rapid classification-switching. This rapid oscillation between classifications was even more worriesome than Brainput's performance on the novel dataset. Specifically, the rapid sub-second oscillations were inconsistent with basic hemodynamics – that task-related hemodynamic changes occur over a period of several seconds [1]. These results thus indicated the Brainput classifier was not the primary factor contributing to performance improvements on the human-robot team task in [5], but rather, indicated the presence of a placebo effect or confounding factor (e.g., the mere presence of robot autonomy) in the protocol.

Placebo-controlled replication. To understand the performance improvements observed in [5] despite the unrealistic behavior of Brainput's classification, we performed a placebo-controlled replication of the original protocol ([5]), with the only modification being that the *cognitive state classifications from Brainput were replaced by random classifications* (generated based on the classification distributions from the original study). This design allowed us to explicitly mea-

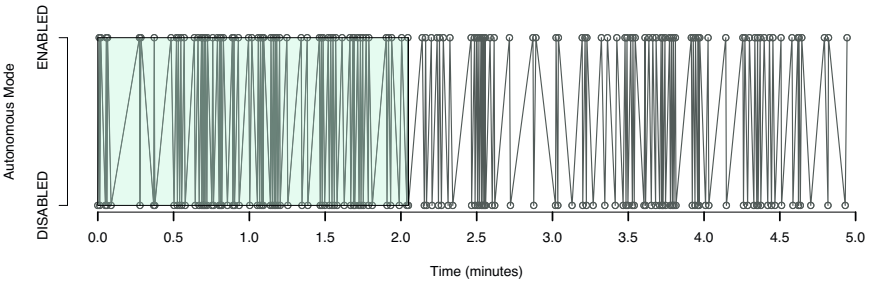

Fig. 1. Example classification log of *autonomy* toggling. Each line indicates a switch in autonomy (e.g., enabled to disabled). Figure reproduced from [1] with permission.

sure effects due to *autonomy* of the robot separate from Brainput's accuracy in recognizing multitasking states. Here we expected (random) *adaptive* and (random) *maladaptive* conditions to have equal effects on performance given that the adaptivity was drawn randomly from the *same* underlying classification distribution. Instead we found similar patterns of task success accross this placebo-replication and [5]. Specifically, subjects succeeded (in both experiments) more often at locating the goal location with the autonomous robot in the *adaptive* condition than in the *maladaptive* condition. This was surprising because the random classifications should have caused the success rates across the *adaptive* and *maladaptive* conditions here to be equal, given that the two conditions themselves were equal in all aspects in this replication.

Hence it was not Brainput that yielded the performance improvements observed initially, as, in the absence of the brain-based adaptivity, we still achieved these improvements. This result thus indicated a confound in the execution of the two conditions (adaptive vs. maladaptive). Upon inspection of the logs from this replication, we discovered a serious confounding factor: the goal location in the environment co-varied with the task condition. Specifically, the robot had significantly further to travel in the *maladaptive* condition (18.4m) than in the *adaptive* condition (9.4m). Hence, the task was strictly more time-consuming in the *maladaptive* condition. As success was determined by the team's ability to locate the goal in 5 min., it is clear that the coordinates of the goal location relative to the robot affected rate of success. Since no aspects of the underlying system were changed aside from the classification approach (from NIRS-based to random) in this placebo-controlled replication of [5], this affected the original study as well. Review of the logs from the original evaluation ([5]) confirmed that the goal locations in the adaptive versus maladaptive conditions of the original study were the same as what we found in this replication.

4 Discussion

4.1 Implications

Due to finding null results in an extension of [5], we were motivated to conduct a reinvestigation of the Brainput framework. We first revisited Brainput's signal processing to test the its extensibility to a novel dataset. There we found its classification approach to perform worse than expected (effective for only 25% of subjects), indicating low efficacy for a general population. Moreover, in looking at Brainput's realtime behavior (plots of the classification logs from [5]), we found that its behavior did not follow basic hemodynamic principles. With unrealistic behavior and worse-than-expected classification accuracy, we then revisited the original evaluation of Brainput's efficacy via a placebo-controlled experiment to find what was responsible for the performance improvements observed in [5]. There, by successful replication of the original performance improvements – despite the absence of NIRS-based adaptivity – we identified a confound within the experimental design. In analyzing logs from the replication and the original study, we found a disparity in starting locations between the experiment conditions (i.e., the robot was 2x closer to the target location in the adaptive versus maladaptive condition), which resulted in the task being easier in the *adaptive* condition. These results indicate that further work is necessary to achieve a robust framework for NIRS-BCIs.

4.2 Challenges

This reinvestigation from [1], as well as the limitations to the interpretation of results, highlight several challenges for NIRS-based BCIs. Here we discuss three. First, questions of generizability were raised: whether results from small samples extend to larger populations, as well as whether models trained on one workload-induction strategy (e.g., the n-back task) extend to other workload-based variants (e.g., human-robot team tasks) [6]. As many NIRS studies are underpowered (e.g., [5]) and all BCIs must necessarily be trained on labeled data (i.e., *not* unconstrained and asynchronous tasks such as human-robot interactions), these questions are highly relevant to HCI. In particular, successful deployment of a BCI depends on how well it works within a general population and also how well the training tasks model the more realistic target tasks.

Second, although the reinvestigation revealed low reliability of the Brainput framework, subjective reports indicate some utility of the brain-based adaptivity [1]. This suggests that the current ways in which we measure and interpret the success or efficacy of a BCI may not capture its full or true utility. For instance, although Brainput's classification accuracy on the novel dataset was only slightly above chance level (55%), that minor improvement in understanding (and adaptive response to) the user's cognitive state may be sufficient to improve a interactions with a robot. Lastly, efforts to disseminate the results of this reinvestigation questionned what is an appropriate forum for (failures in) replication. While this is not unique to NIRS-BCIs, the reliability of signal

processing is an orthogonal endevor to developing effective interactions between people and computer systems. Hence it is not comparable to applications of NIRS-BCIs in terms of novelty. It is, however, complimentary in that the utility of NIRS-based BCIs is dependent on the robustness of the methods used to extract meaningful information from noisy signals. Thus, while growing accessibility of brain-based sensors allows for researchers to approach signal processing as a black box, it is important to consider how we can facilitate discussion on both fronts (novel applications and signal processing methods) as they are both invaluable to the advancement of NIRS-BCIs. Publication, in particular, allows discussion, feedback, and improvement of the work, but without a clear venue for replication, progress for NIRS-BCIs will be slowed.

5 Conclusions

NIRS-based BCIs have received considerable attention as a tool for HCI. However, in this series of reinvestigations of the Brainput NIRS-BCI, we found significant limitations of its efficacy. First, we found when we increased our sample size from 3 to 40, Brainput's performance was only effective for 1/4 of the population. Moreover, we observed that Brainput's realtime behavior (rapid state-switching) is not in accordance with basic hemodynamic principles (slow changes). Further investigation identified a major confounding factor (different goal locations) in our original evaluation of the system, which was likely responsible for the performance improvements (not the NIRS-based adaptive autonomy). Hence, it is important that we revisit our NIRS-BCI frameworks to consider the reliability of our systems. We hope that this systematic reinvestigation and discussion of related challenges will lead towards more robust NIRS-BCIs.

References

1. Strait, M., Canning, C., Scheutz, M.: Reliability of NIRS-based BCIs: a placebo-controlled replication and reanalysis of Brainput. In: Alt.CHI (2014)
2. Canning, C., Scheutz, M.: Function near-infrared spectroscopy in human-robot interaction. Journal of Human-Robot Interaction 2, 62–84 (2013)
3. Strait, M., Scheutz, M.: Building a literal bridge between robotics and neuroscience using functional near infrared spectroscopy. In: Human-Robot Interaction (HRI) Workshop on Bridging Robotics and Neuroscience (2014)
4. Strait, M., Canning, C., Scheutz, M.: Limitations of NIRS-based BCI for realistic applications in human-computer interaction. In: BCI Meeting (2013)
5. Solovey, E., Schermerhorn, P., Scheutz, M., Sassaroli, A., Fantini, S., Jacob, R.: Brainput: enhancing interactive systems with streaming fNIRS brain input. In: CHI, pp. 2193–2202 (2012)
6. Brouwer, A.-M., van Erp, J., Heylen, D., Jensen, O., Poel, M.: Effortless passive bCIs for healthy users. In: Stephanidis, C., Antona, M. (eds.) UAHCI 2013, Part I. LNCS, vol. 8009, pp. 615–622. Springer, Heidelberg (2013)