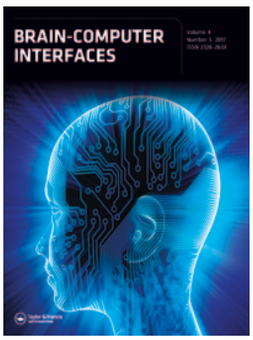

Brain-Computer Interfaces

ISSN: 2326-263X (Print) 2326-2621 (Online) Journal homepage: http://www.tandfonline.com/loi/tbci20

Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy

Yichuan Liu , Hasan Ayaz & Patricia A. Shewokis

To cite this article: Yichuan Liu , Hasan Ayaz & Patricia A. Shewokis (2017) Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy, Brain-Computer Interfaces, 4:3, 175-185, DOI: 10.1080/2326263X.2017.1304020

To link to this article: http://dx.doi.org/10.1080/2326263X.2017.1304020

BRAIN-COMPUTER INTERFACES, 2017
VOL. 4, NO. 3, 175–185
https://doi.org/10.1080/2326263X.2017.1304020

Taylor & Francis
Taylor & Francis Group

Mental workload classification with concurrent electroencephalography and functional near-infrared spectroscopy

Yichuan Liu[a,b], Hasan Ayaz[a,b,c,d] and Patricia A. Shewokis[a,b,e]

[a]School of Biomedical Engineering, Science & Health Systems, Drexel University, Philadelphia, PA, USA; [b]Cognitive Neuroengineering and Quantitative Experimental Research (CONQUER) Collaborative, Drexel University, Philadelphia, PA, USA; [c]Department of Family and Community Health, University of Pennsylvania, Philadelphia, PA, USA; [d]The Division of General Pediatrics, Children's Hospital of Philadelphia, Philadelphia, PA, USA; [e]Nutrition Sciences Department, College of Nursing and Health Professions, Drexel University, Philadelphia, PA, USA

ABSTRACT

A brain-computer interface that measures the mental workload level of operators has applications in human-computer interactions (HCI) for reducing human error and improving work efficiency. In this study, concurrently recorded electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) were combined at the decision fusion stage for the classification of three mental workload levels induced by an *n*-back working-memory task. An average three-class classification accuracy of 42, 43, and 49% has been achieved across 13 participants for the fNIR-alone, EEG-alone, and EEG-fNIRS combined approach, respectively. The current study demonstrated a multimodality-based approach to decode human mental workload levels that may potentially be used for adaptive HCI applications.

ARTICLE HISTORY
Received 25 November 2016
Accepted 6 March 2017

KEYWORDS
FNIRS; EEG; *n*-back; mental workload; multimodal fusion

1. Introduction

The human brain has a limited capacity to maintain and process imminent information in working memory [1]. Persistently high workload experienced by human operators such as aircraft pilots and automobile drivers can be the basis of life-threatening accidents resulting from decreased human performance [2,3]. Conversely, extremely low workload levels may cause boredom and low vigilance which can, in turn, decrease human performance and lead to fatal errors [4]. A reliable assessment technique to monitor human mental workload in real time is needed for improving performance and reducing human error [3,5–7].

There are four indicator types for mental workload assessment: (1) task performance indicators such as time spent, accuracy, and error rate; (2) subjective rating scales such as NASA-TLX [8] for self-assessment; (3) physiological signals such as functional near-infrared spectroscopy (fNIRS), electroencephalography (EEG), electrooculography (EOG), and electrocardiogram (ECG); and (4) hybrid approaches using two or more of the aforementioned indicators together [9,10]. With respect to real-time application, neurophysiological indicators are the preferred choice because of their non-intrusive nature to the task and the continuous availability of the signals. Above all, as neurophysiological indicators are measurements of brain activity, they are also more objective and can potentially reveal direct and finer details of a user's mental state compared to the systemic physiological measures, performance indicators, and subjective rating scales [11]. There is a substantial body of evidence on the real-time evaluation of a user's mental states using neurophysiological signals [12–14]. The phenomenon has recently been referred to as *passive brain-computer interfaces* [15,16]. Although there are many neuroimaging techniques available for acquiring signals from the human brain, EEG and fNIRS are the most suitable for real-life application because of the lower cost and portability of these devices.

1.1. EEG-based workload assessment

EEG, despite the issues regarding artifacts, is the most studied and a very popular physiological measurement for mental workload assessment [17]. As early as 1998, Gevins, Smith and colleagues investigated EEG-based *n*-back classification using spectral features with high (up to 95%) accuracy reported for classifying two workload

CONTACT Patricia A. Shewokis shewokis@drexel.edu

levels across eight subjects [12]. Furthermore, the classifications were found to be stable across days, participants, and two different tasks (verbal/spatial n-back). However, the data were recorded in three sessions of 6–8 hours time span with two-thirds of the data randomly chosen for training the classifier. In addition, the EEG data were visually inspected for artifact rejection before classification. All of these methodological assessments could be prohibitive for practical use. Grimes, Tan, and colleagues investigated the effect of several factors that might affect the classification accuracy for the *n*-back task [14]. These factors include the time window size of a trial, the amount of available training data, and the number of EEG channels. Although all of these factors were found to affect classification accuracy, the time window size was particularly sensitive. Using a combination of spectral and phase-coherence features, they achieved classification accuracy comparable to Gevins et al. using as few as 20 min worth of training data and a 30 s time window size. Brouwer and colleagues investigated *n*-back classification across 35 subjects using event-related potential (ERP) features in addition to the traditionally adopted spectral features which resulted in improved classification accuracy [18]. Muhl et al. found that EEG-based *n*-back classification can be performed across two affective contexts – relaxed and stressed – without much performance loss [19].

1.2. fNIRS-based workload assessment

Recently, fNIRS-based mental workload assessment has gained attention among researchers. fNIRS is an optical brain imaging technology for monitoring the concentration changes of oxygenated hemoglobin (ΔHbO) and deoxygenated hemoglobin (ΔHbR) in the cortex. It is easy to set up and particularly suitable for monitoring prefrontal activation due to its resistance to eye-movement artifacts and it is a cost-effective device for measuring cortical activation from the hairless forehead areas. fNIRS has been used to investigate cognitive activation patterns under different task load conditions. For example, Ayaz et al. compared the average fNIRS activation under different workload levels for *n*-back tasks [10,20], air traffic control tasks [21], and UAV piloting tasks [11]. Significant fNIRS activation was found across the different workload levels. The potential of fNIRS for workload level classification is important and the results could establish a framework to enable future real-time applications. Herff et al. (2014) investigated single-trial *n*-back classification using only prefrontal fNIRS [13] and ~78% accuracy was achieved for discriminating between 1-back and 3-back conditions. Hong et al. in 2015 investigated fNIRS-based classification of three tasks: mental arithmetic, left-hand motor imagery, and right-hand motor imagery [22]. An average classification accuracy of 75.6% was achieved across 10 subjects. Naseer and Hong (2015) further investigated fNIRS-based classification of four tasks: right-hand motor imagery, left-hand motor imagery, mental arithmetic, and mental counting [23], which resulted in an average classification accuracy of 73.3%.

1.3. The concurrent EEG-fNIRS approach

In this study, we investigated workload classification with concurrent EEG and fNIRS. Our motivation is that EEG and fNIRS are highly complementary technologies each having its own advantages and disadvantages. EEG, for example, is precise in the time domain but lacks spatial specificity due to the volume conductive effect. fNIRS, on the other hand, is spatially specific yet lacks the time precision due to the delay in the hemodynamic response. EOG contamination is still a concern for EEG especially in the prefrontal areas despite substantial efforts to either detect or reduce the effect [24–26]. In contrast, fNIRS is not affected by eye activities and is particularly suitable for recording in the usually hair-free prefrontal sites. These qualities further highlight the compatibility for integration.

Concurrent EEG and fNIRS has been previously investigated and shows potential for brain-computer interfacing (BCI) [27–35]. Pfurtscheller et al. (2010) described a hybrid BCI that used fNIRS as a brain switch to turn on/off a steady-state visual evoked potentials (SSVEP) BCI [24], while Leamy found that using simultaneous EEG-fNIRS was able to better differentiate two mental states of the subjects: motor imagery and rest [28]. Fazli and colleagues investigated enhancing the performance of EEG-based real-time motor imagery BCI with fNIRS using two approaches. In the first approach, they directly used the fNIRS features in classification [30], while the second approach, fNIRS was used to predict the performance of an EEG-based BCI that informs a meta-classifier [29]. Both approaches improved the BCI accuracy by around 3–5% from a baseline of about 78% achieved by an EEG-only approach. Morioka and colleagues reported the enhancement of an EEG-based decoder in a spatial attention task using fNIRS prior to solving the cortical current estimation problem [31]. They found that using EEG cortical sources estimated with prior fNIRS sources significantly improved decoding accuracy (~79%) compared to the traditional EEG-only approach (~71%). Khan et al. in 2014 proposed a four-command BCI that used fNIRS to decode mental-counting/arithmetic and used EEG to decode left/right hand tapping [34], while Tomita et al. in 2014 showed that an improved classification can be achieved by adding fNIRS to detect the on and off state for an SSVEP-based BCI [33]. Blokland et al. investigated

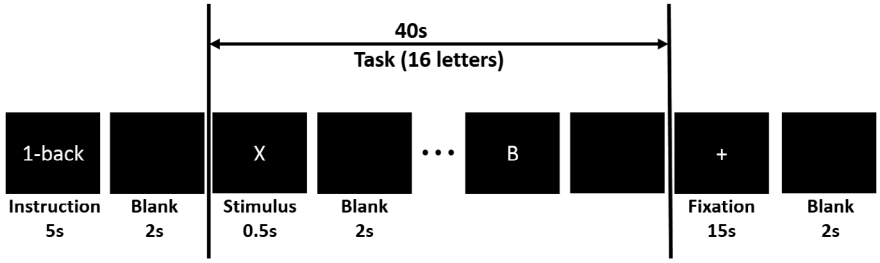

Figure 1. Time line of an *n*-back block.

differentiating motor execution, attempted execution, and motor imagery from rest for healthy subjects and tetraplegia patients [32]. They demonstrated that adding fNIRS improved classification accuracy approximately 6% for attempted movement and 7% for imagined movement. Yin et al. in 2015 extracted EEG time-phase-frequency features and combined them with fNIRS for the classification of a motor imagery task. Adopting a joint mutual information criterion for feature selection and an extreme learning machine for classification, a 1–5% improvement in decoding accuracy was found with the inclusion of fNIRS [35].

In the area of mental workload classification, the potency of EEG-fNIRS is yet to be shown. To our knowledge, there are three preliminary studies thus far addressing EEG-fNIRS mental workload classification. Hirshfield et al. investigated the classification of mental workload levels induced by a plane counting task and reported a binary classification accuracy of 53–55% (EEG) and 70–72% (fNIRS) [36]. Coffey et al. investigated the classification of *n*-back induced workload levels and reported a 2- vs. 0-back accuracy of 73% (EEG), 61% (fNIRS), and 61% (EEG+fNIRS) [37]. The authors of the two studies suggested a number of factors that might have affected accuracies, including but not limited to sensor layout/placement, signal processing, and the task adopted to elicit workload changes. Herff and colleagues recently reported the preliminary results of memory load level classification with concurrent EEG-fNIRS and showed that feature-level fusion of the two modalities increased the robustness of classification using the data recorded from 3 EEG channels, 4 EOG channels, 28 fNIRS sources, and 15 fNIRS detectors [38].

The main objective of this study is to investigate the fusion of EEG and fNIRS to discriminate mental workload levels. We adopted the *n*-back working memory task, which has been used by numerous working memory classification studies [12–14,18,19,36,37] to induce three controlled workload levels: 0-back, 1-back, and 2-back. A novel approach was employed to combine EEG and fNIRS at the decision fusion stage by taking into consideration the different temporal resolutions of the two modalities. The classification performance of the proposed approach was compared to those achieved by each single modality alone.

2. Materials and methods

2.1. Participants

Sixteen volunteers (six female) aged between 18 and 30 (mean = 22, standard deviation [SD] = 3) from Drexel University participated in the study. The Edinburgh Handedness Inventory showed that participants were all right-handed and the average Laterality Quotient (L.Q.) and Decile were 80 (SD = 18) and 6.4 (SD = 3.1), respectively. Participants self-reported that they had their vision corrected to 20/20, did not have any history of neurological or psychiatric disorders, and did not take any medication known to affect brain activity. Participants further self-reported to be naïve to the *n*-back paradigm. One participant was rejected from the study due to missing fNIRS data. Two more participants were excluded from the study due to excessive motion artifact in the fNIRS and EEG measures. As a result, the data from a total of 13 participants were used in the analyses. Prior to the experiment, participants gave written informed consent for their participation in the study. The protocol was approved by the Institutional Review Board of Drexel University (IRB protocol: 1409003112).

2.2. Experiment paradigm

A visual verbal *n*-back task was adopted to manipulate mental workload level. Subjects sat comfortably in front of an LED screen. Sequences of capitalized letter stimuli (~1.7°° visual angle) were shown at the center of the screen. We employed the BCI2000 software for stimulus delivery and for the recording of EEG and behavioral data [39]. Each letter was displayed for a duration of 480 ms and the inter-stimulus interval (ISI) was 2000 ms. Subjects were instructed to identify target letters and press the 'enter' key on a number keypad with their right index finger as fast as possible. There were three workload conditions. In

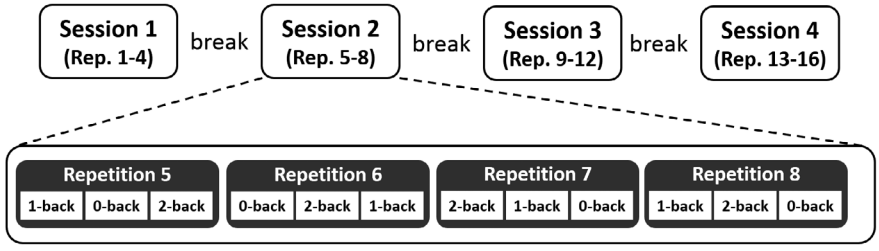

Figure 2. Experiment outline.

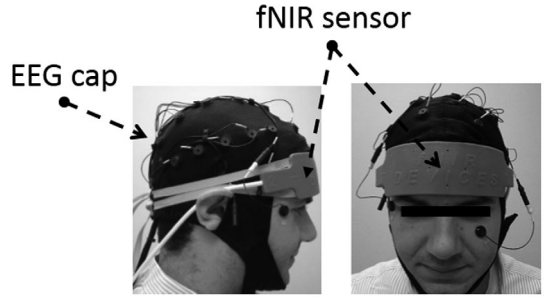

Figure 3. Recording setup.

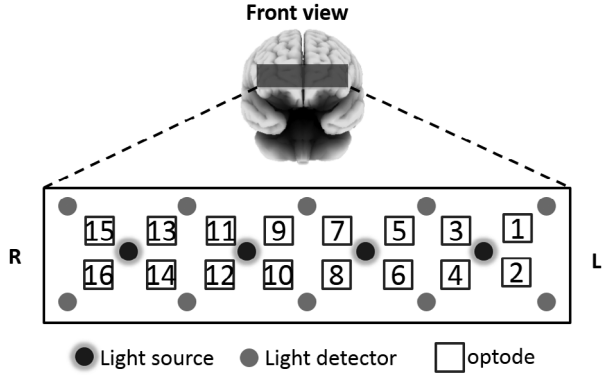

Figure 4. fNIRS sensor layout.

the 0-back condition, the letter 'X' was the target. In the 1-back condition, the current letter was the target if it was shown on the previous screen as well. In the 2-back condition, a letter was the target if it was shown two screens back. To successfully complete the 2-back task, the performer needed to keep updating a memory load of two letters, the letter on the previous screen and the letter two screens back.

The letter stimuli were grouped into *n*-back blocks. Each block included 7 s of instructions, 40 s of task execution, and 17 s of fixation. The instruction period informed the subject which task (0-, 1-, or 2-back) to perform. During the task period, 16 letters were shown to the participants on the screen in a pseudo-random order. Five of the letters were targets. No letters appeared more than twice in succession within a block. In the fixation period, subjects were instructed to focus their eye gaze on a white plus sign located at the center of the screen: this fixation allowed the fNIRS signals to return to the baseline. Figure 1 shows the time line of an *n*-back block.

There were four recording sessions. Each session included 12 *n*-back blocks, 4 from each condition. Hence, there were 48 *n*-back blocks for the entire experiment, 16 from each condition. To reduce the relationship between adjacent samples and to balance time induced experimental factors such as fatigue across the three workload conditions, the 48 *n*-back blocks were grouped into 16 repetitions. Each repetition included one block from each workload condition. The order of the blocks was further randomly shuffled so that no workload condition was repeated twice in succession within a session. Before the start of the first session, subjects practiced one block from each condition for familiarization with the procedure. Subjects took a break between the recording sessions for as long as requested. The entire recording time was about one hour. Figure 2 shows the outline of the experiment.

2.3. Data acquisition

EEG and fNIRS were simultaneously recorded from the subjects. Figure 3 illustrates the recording setup.

EEG was recorded using a Neuroscan Nuamp amplifier from 28 locations according to the International 10–20 system. Two additional electrodes, one placed below the left eye and the other placed at the right outer canthus, were used to record electrooculography (EOG) activities. All channels were grounded at the left mastoid, referenced to the right mastoid, digitally sampled at 500 Hz and low-pass filtered at 100 Hz for analysis.

Prefrontal fNIRS was recorded using a 16-channel continuous-wave fNIRS system developed at Drexel University [40,41] and manufactured by fNIRS Device LLC. The sensor included four light sources (LED) that can emit 730 nm and 850 nm wavelengths light and 10 photo detectors (see Figure 4). The distance between light sources and detectors is 2.5 cm, which allowed for a ~1.2 cm penetration depth. Data were sampled at 2 Hz and recorded using the Cognitive Optical

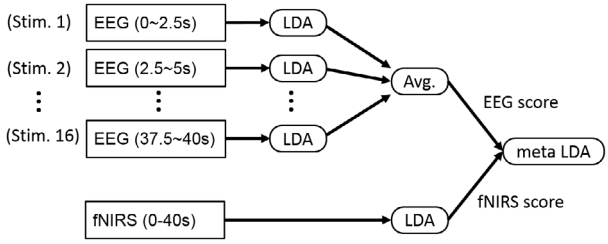

Figure 5. EEG-fNIRS workload classification. An LDA was trained to classify EEG band power features at single-stimulus level (2.5 s epoch). The output scores from the 16 stimuli (of a block) were averaged to produce the EEG score. Another LDA was trained to classify fNIRS features extracted from each block (40 s epoch) to produce the fNIRS score. A meta-LDA was then trained to optimally combine the EEG score and the fNIRS score for EEG-fNIRS classification. All of the above procedures were conducted on training data. All three LDA classifiers were then applied on testing data to evaluate the classification performance.

Brain Imaging (COBI) studio software [42]. To ensure repeatable sensor placement, the center of the sensor was aligned to the midline of the superior forehead and the bottom of the sensor was touching the participant's eye brow.

The EEG and fNIRS signals were synchronized using stimulus triggers that were sent from the stimulus presentation software BCI2000 [39] to the EEG and fNIRS data-acquisition devices during recording.

2.4. Signal processing and feature extraction

EEG. The raw EEG were high-pass filtered at 1 Hz to remove drift and notch-filtered at 58–62 Hz to suppress line noise using a finite impulse response (FIR) filter before subsequent analyses. The following 17 channels were used (according to the 10–20 system): Cz, Fz, Pz, F3, F4, P3, P4, O1, O2, F7, F8, T7, T8, C3, C4, P7, and P8. The EEG epochs were extracted from −500 ms to 2000 ms with respect to the onset of each single stimulus. Power spectral density (PSD) has been widely adopted for mental workload characterization [12,14,18,43,44]. The PSD of each epoch was estimated by adopting a periodogram with a Hann window. The log band power of δ (1–4 Hz), θ (4–8 Hz), α (8–13 Hz), β_1 (13–20 Hz), and β_2 (20–30 Hz) bands were used as features. EEG data included 17 [channels] × 5 [bands] = 85 features and 3 [workload levels] × 15 [blocks] × 16 [stimuli] = 720 samples.

fNIRS. Raw light intensities were first visually inspected to reject those optodes with inadequate contact or those positioned over the hairline. The Sliding-window Motion Artifact Rejection (SMAR) algorithm was adopted to reject motion-artifact-contaminated signal segments [45]. Raw light intensities were low-pass filtered at 0.08 Hz using the FIR filter for removing artifacts from physiological signals such as heartbeat and respiration and converted into concentration changes in oxy-hemoglobin (Δoxy-Hb) and deoxy-hemoglobin (Δdeoxy-Hb) using the modified Beer-Lambert law [46]. The Δoxy-Hb and Δdeoxy-Hb epochs of each *n*-back block were extracted from the onset of the first stimulus to 2.5 s after the onset of the last (the 16th) stimulus – a 40 s time window. The average activation difference between the first 5 s and the last 35 s of a block was calculated as a feature. The average activation amplitude with respect to a baseline has been adopted as the feature for characterizing the mental activities in many studies [20,40,47–49]. To further decrease feature size to reduce overfitting, fNIRS features in the following three areas were averaged (see Figure 4 for sensor layout): left lateral (optode 1–4), medial (optode 5–12), and right lateral (optode 13–16). fNIRS data included 3 [areas] × 2 (Δdeoxy-Hb / Δoxy-Hb) = 6 features and 3 [workload levels] × 15 [blocks] = 45 samples. Alternative feature options such as slope, peak, and range were explored but preliminary results suggested they are prone to outliers and do not outperform the most commonly adopted average amplitude.

2.5. Classification

We considered the binary classification problem of 2- vs. 0-back, 2- vs. 1-back, and 1- vs. 0-back and the three-class classification problem of 2- vs. 1- vs. 0-back. The following three approaches were compared:

- **EEG-alone**. A linear discriminant analysis (LDA) was trained to classify EEG features at a single trial level (2.5 s time windows with respect to a single stimulus). The LDA outputs from the 16 stimuli of a block were averaged to produce the EEG score, which determines the predicted workload levels.
- **fNIRS-alone**. An LDA was trained to classify fNIRS features at block level (40 s time windows, included 16 stimuli). The LDA output is termed the fNIRS score.
- **EEG-fNIRS**. A meta-LDA was trained to optimally combine the EEG score and the fNIRS score for EEG-fNIRS classification (see Figure 5).

For the EEG-based classification, an alternative method is averaging the band-powers within each block before LDA classification. However, there are only 15 sample blocks per condition with 85 features. Although alternative classifiers such as regularized-LDA and support vector machines can handle this situation, overfitting may become an issue. By training the LDA at the single stimulus level, the high temporal resolution of EEG is utilized and the sample size is increased to 240 per condition.

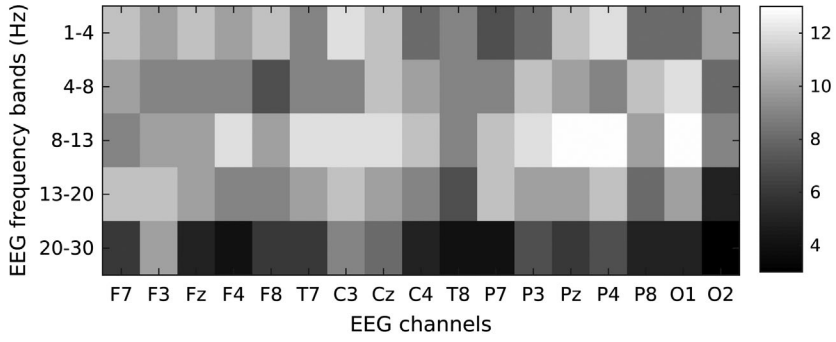

Figure 6. The number of participants (indicated by gray level) with better than chance-level (i.e. >55.4%) LDA accuracy.

2.6. Performance evaluation

Leave-one-out cross-validation (LOOCV) was adopted to evaluate classification performance. In LOOCV, all but samples from one block were used to train the three LDAs aforementioned. The LDAs were then applied on the block that was left out to evaluate classification performance. This procedure was repeated until all *n*-back blocks were left out exactly once.

A one-sided Wilcoxon signed-rank test was adopted to evaluate whether the median accuracies from the 13 participants were significantly better than 55.4% and 37.8%, the chance level, for the binary and three-class classification problems, respectively. The chance level is determined by the upper bound of the 95% confidence interval assuming the random-guess classification results follow a binomial/multinomial distribution [50].

2.7. Multiple comparisons

To correct for multiple testing, we adopted false discovery rate (FDR) control with the Benjamini-Hochberg procedure [51]. Without specification, we rejected null hypotheses for FDR $q < .05$.

3. Results

3.1. Key-press behavior

We evaluated the effect of mental workload on the key-press response time (RT) and target identification accuracy (TIC) with a repeated-measures one-way analysis of variance (ANOVA). The RT and TIC measures were normalized across the three workload conditions subject-wise before statistical tests.

For RT, the mean and SD were 450 ± 54 ms, 450 ± 88 ms, and 549 ± 145 ms for 0-back, 1-back, and 2-back conditions, respectively. A significant effect of mental workload level was found ($F_{(2,24)} = 24.7$, $p < .001$). Post-hoc tests revealed a significant difference in RT between 0-back and 2-back (FDR $q < .01$) and between 1-back and 2-back conditions (FDR $q < .001$).

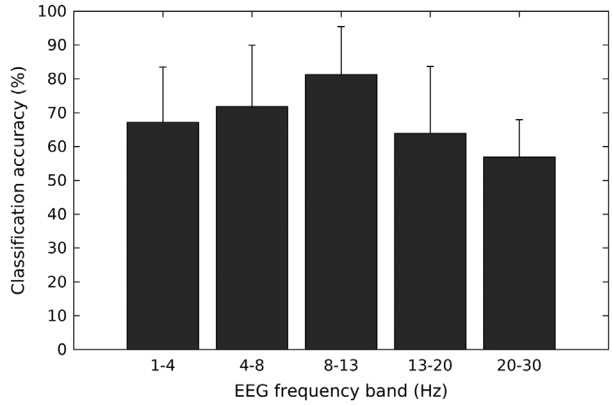

Figure 7. Average and standard deviation of LDA accuracy across 13 participants.

For TIC, the mean and SD were 98.0 ± 2.5%, 97.2 ± 2.9%, and 94.0 ± 5.8% for 0-back, 1-back, and 2-back conditions, respectively. ANOVA revealed a significant effect of mental workload ($F_{(2,24)} = 3.9$, $p < .05$). Post-hoc tests found only a significant difference between 0-back and 2-back conditions (FDR $q < .05$).

3.2. Reliability of EEG features

To show the spatial and spectral reliability of EEG features, LDA accuracies for discriminating 0-back and 2-back blocks were estimated by adopting subsets of EEG features and the results are shown in Figures 6 and 7. As expected, EEG power spectral density (PSD) features were most reliable in the alpha band over parietal areas (P3, Pz, and P4).

3.3. Reliability of fNIRS features

To show the reliability of fNIRS features, LDA accuracies for discriminating 0-back and 2-back blocks were estimated by adopting subsets of fNIRS features as shown in Figure 8. Overall, Δoxy-Hb provided better results

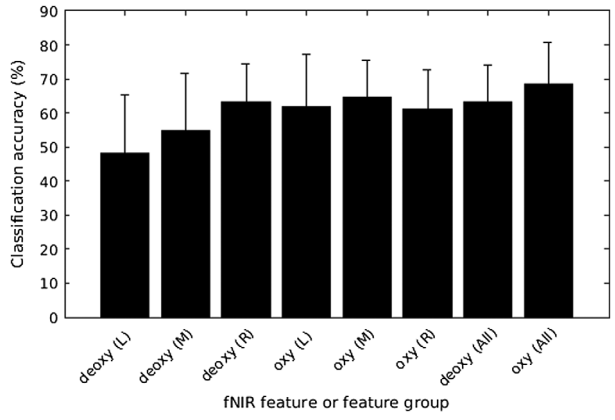

Figure 8. Average and standard deviation of LDA accuracy across 13 participants. L, left lateral PFC area; M,– medial PFC area; R, right lateral PFC area; All, all areas included.

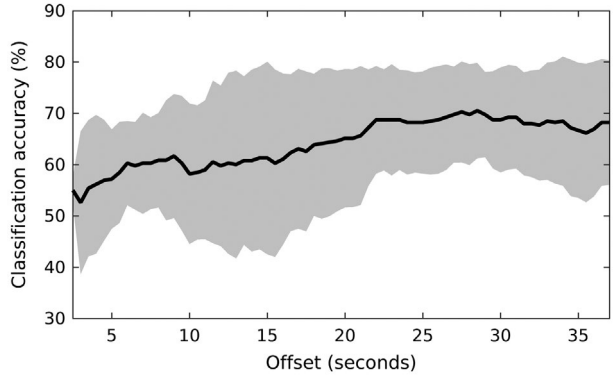

Figure 9. LDA accuracy using fNIRS adopting a moving time window of 5 s length. Solid line and the shaded area show average and standard deviation across 13 participants. Horizontal axis shows mid-point of time window from block start.

Table 1. Comparison of classification results. The results shown are mean ± standard deviation accuracy across 13 participants.

	n-back comparison			
Accuracy (%)	1 vs. 0	2 vs. 0	2 vs. 1	2 vs. 1 vs. 0
EEG-alone	32.8±11.5	79.2±15.2[#]	71.0±18.6[#]	42.7±12.8
fNIRS-alone	46.7±9.7	65.6±6.1[#]	59.5±15.4	41.7±7.9
EEG-fNIRS	38.2±8.8	83.1±12.6[#]	72.6±18.1[#]	48.7±12.7[#]

[#]Significantly better than chance level (FDR $q < .05$).

compared to Δdeoxy-Hb though no significant differences were found (Wilcoxon signed-rank test, FDR-adjusted).

We evaluated the temporal reliability of fNIRS features by estimating LDA accuracies using the averaged fNIRS activation within a 5 s moving time window as a feature, depicted in Figure 9. It can be seen that average accuracy increases with elapsed time from the block start to its stabilization at approximately 22 s. This can be explained by the time needed for the participants to start experiencing workload and the delay of the hemodynamic response.

3.4. Mental workload classification

Table 1 summarizes the average classification results achieved by the three approaches (EEG-alone; fNIRS-alone; EEG-fNIRS). For 1- vs. 0-back, none of the approaches achieved meaningful classification. Together with the behavioral performance, we found no obvious workload difference between 0-back and 1-back conditions. For 2- vs. 0-back, all three approaches achieved accuracies significantly better than 55.4% (the binary chance level). For 2- vs. 1-back, EEG-alone and the EEG-fNIRS performed significantly better than chance. Only EEG-fNIRS achieved accuracies significantly better than 37.8% (three-class chance level) for the 2 vs. 1 vs. 0 case.

The performances achieved by individual subjects using the three approaches are compared in Figure 10. EEG-fNIRS outperformed EEG-alone for the 1- vs. 0-back, 2- vs. 0-back, and 2- vs. 1 vs. 0-back (FDR $q < .05$). By including fNIRS, an equal or better accuracy was achieved for 12 out of 13 participants for 2- vs. 1-back and 2- vs. 1- vs. 0-back, whereas all 13 participants achieved better 2- vs. 0-back accuracy. EEG-fNIRS outperformed fNIRS-alone only for the 2- vs. 0-back case (FDR $q < .05$). Finally, compared to the fNIRS-alone approach, EEG-alone achieved better 2- vs. 0-back performance (FDR $q < .05$).

4. Discussion

Hemodynamic response signals may provide additional information to the commonly adopted electroencephalograph for the discrimination of cognitive tasks. In the current study, simultaneously recorded EEG-fNIRS was adopted for the classification of mental workload levels. The behavioral results suggested a higher mental workload level was experienced by subjects in the 2-back compared to 0-back and 1-back conditions. For fNIRS-alone classification, we achieved 65.6, 59.5, and 41.7% average accuracy for 2- vs. 0-back, 2- vs. 1-back, and 2- vs. 1- vs. 0-back, respectively. For EEG-alone classification, 79.2, 71.0, and 42.7% average accuracies were achieved respectively for the 2- vs. 0-back, 2- vs. 1-back, and 2- vs. 1- vs. 0-back tasks.

We next showed that classification performance can be improved by combining fNIRS and EEG. The 2- vs. 1- vs. 0-back three-class classification accuracy was increased by an average of 6% and 7% respectively compared to using only EEG or only fNIRS. By integrating EEG and fNIRS, the workload classification was significantly higher than 37.8%, the upper bound of the 95% confidence interval of chance-level accuracy. Previously, EEG-fNIRS has improved classification over the EEG-alone approach for motor imagery tasks [29,30,32], spatial attention tasks [31], and the on/off state of SSVEP-based BCIs. Our study

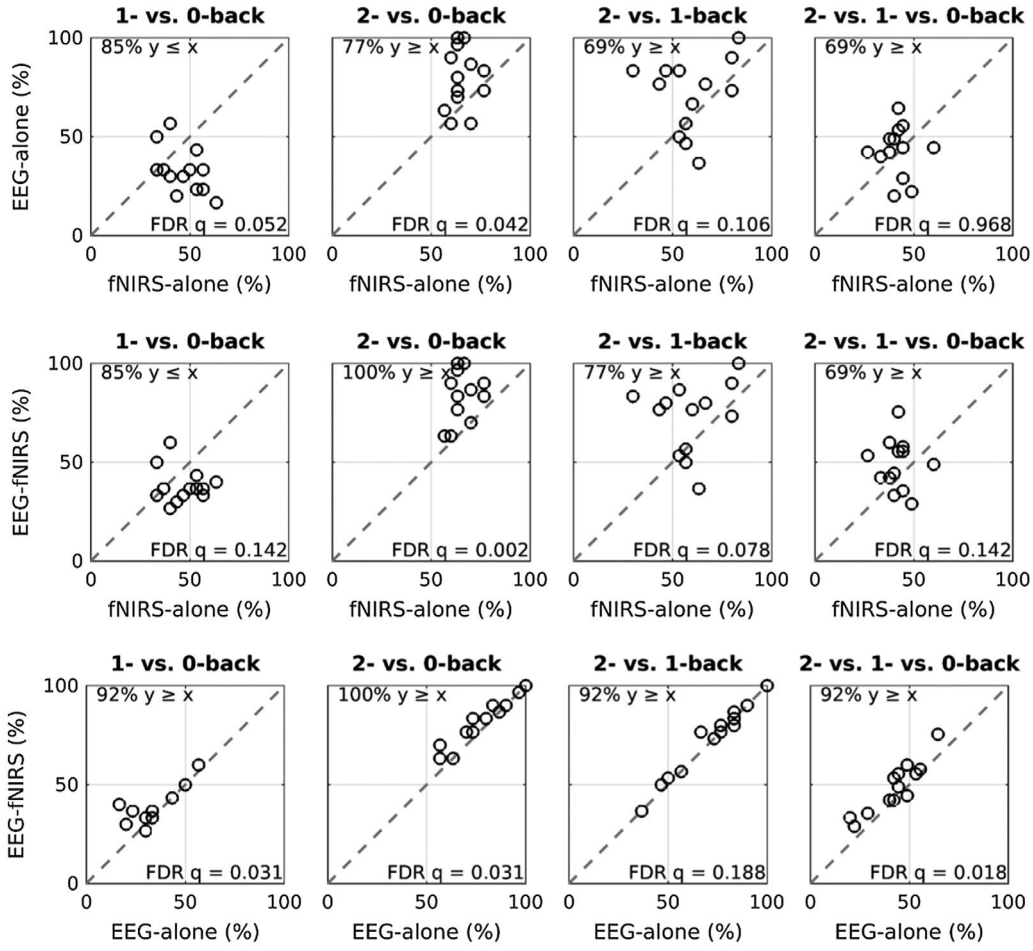

Figure 10. Scatter plot comparing classification accuracies. Top row: fNIRS-alone (*x*-axis) and EEG-alone (*y*-axis); middle row: fNIRS-alone (*x*-axis) and EEG-fNIRS (*y*-axis); bottom row: EEG-alone (*x*-axis) and EEG-fNIRS (*y*-axis). Each circle represents the results from one participant. The percentage at top left indicates the percent of subjects with equal or improved accuracy by adopting the better approach. The results from the Wilcoxon signed-rank test with H_0:$x = y$ and FDR correction are shown at the bottom of each sub-figure.

suggested an improvement may also be obtained for a working-memory task.

EEG data are often contaminated by EOG artifacts, particularly at the frontal sites. To investigate whether reducing EOG contamination may help improve classification, we employed an independent component analysis (ICA) based EOG reduction technique to spatially filter EEG data and repeated the classification analysis. EOG reduction resulted in a deterioration of classification performance, which may be caused by a removal of some useful information along with the EOG components. This analysis, however, is preliminary. Future work may be conducted investigating alternative EOG-reduction techniques [25,52–54].

In this study, a three-class classification accuracy of 48.7% was achieved by integrating EEG and fNIRS. Various approaches may be adopted to further improve the workload classification accuracy. First of all, techniques other than linear discriminant analysis may be investigated for classification. Naseer et al. in 2016 compared the fNIRS-based binary classification of mental arithmetic vs. rest using linear discriminant analysis, quadratic discriminant analysis, *k*-nearest neighbor, naïve Bayes, support vector machine, and artificial neural networks (ANN). They found that the ANN approach provided the best classification. Second, alternative features may be adopted to characterize mental workload. In our study, we extracted the average activation changes and the band powers from fNIRS and EEG, respectively. Herff et al. adopted the slopes of straight lines fitted into a moving window as features for fNIRS characterization [13]. Grimes et al. adopted the phase-locking values between EEG channels as features [14]. The size of these features, however, is large. Hence more sample blocks per subject are needed in order to prevent overfitting.

The present study demonstrated the potential of the EEG-fNIRS multimodality approach for measuring

mental workload level. Despite the promising results, our study was limited primarily in three aspects. First, the current experiment was conducted in a controlled experimental environment and we assumed a synchronized BCI scenario in which the start and end of a task block are known before signal processing and feature extraction. In real-life applications, the measuring of mental workload needs to be able to operate in an asynchronous manner. This could be particularly challenging for continuous-wave fNIRS considering the feature-extraction procedure involved in baseline correction. A potential method to solve this problem is to remove low-frequency trends via high-pass digital filtering or to apply wavelet-based de-trending [55]. Second, due to the non-stationary nature of fNIRS and EEG signals, the test-retest reliability of the proposed approach across different sessions and days needs careful investigation. In a real-world application of BCI, a change in the joint distribution of input features with time – formally known as the covariate shift – is usually unavoidable. This covariate shift effect may be alleviated by adaptive algorithms, recently proposed in the literature, that weight training samples according to their deviation from the distribution of testing data [56,57] or by on-line updating an estimation of the covariance matrix when new data are acquired [58,59]. Third, we recruited participants aged between 18 and 30 years from universities. Since hemodynamic responses may be different for young and elderly subjects [60], how age may affect workload classification requires further investigation. Finally, due to limitations in simultaneous EEG-fNIRS acquisition hardware, we recorded fNIRS only from prefrontal sites and the two prefrontal EEG channels Fp1 and Fp2 were not recorded. We expect including the Fp1 and Fp2 is unlikely to have an important impact on EEG classification results since the prefrontal activities are distributed to other sites because of the volume conductive effect. Evidence regarding the Fp1 and Fp2 influence on EEG classification could be found in [14], where using only 16 channels and in some cases 8 channels (both with the two prefrontal sites excluded) resulted in nearly the same *n*-back classification accuracy compared to the accuracy level using all 32 channels. The impact of including central-parietal fNIRS sites, however, is difficult to predict due to a lack of full-scalp fNIRS-based working-memory classification studies in the literature. Most of the current work only recorded fNIRS from the hairless prefrontal sites due to the portability and cost-effectiveness of the sensor. We expect fNIRS-based classification may be improved by including the central-parietal sites. Future work, addressing the aforementioned issues, is needed to improve our understanding of multimodal workload classification methods.

5. Conclusion

In summary, we sought to improve mental workload classification with simultaneously recorded EEG and fNIRS. By including fNIRS in addition to EEG, a significant improvement in classification accuracy was found despite the relatively low classification accuracy achieved using only fNIRS. The current study presents a promising application of the simultaneous EEG-fNIRS approach in enhancing the performance of a passive BCI.

Acknowledgements

We thank Dr. Anna Rodriguez for her meaningful discussions on workload classification. This study is made possible, in part, by a research award from the National Science Foundation (NSF) IIS: 1065471 (Shewokis, PI). The content of the information herein does not necessarily reflect the position or the policy of the sponsors and no official endorsement should be inferred.

Disclosure statement

The optical brain imaging instrumentation utilized in the present research was manufactured by fNIR Devices, LLC. Dr. Hasan Ayaz was involved in the development of the technology and thus offered a minor share in fNIR Devices, LLC. All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORCID

Yichuan Liu http://orcid.org/0000-0002-1899-2082
Hasan Ayaz http://orcid.org/0000-0001-5514-2741
Patricia A. Shewokis http://orcid.org/0000-0001-9246-9495

References

[1] Cowan N. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. Behav Brain Sci. 2001;24:87–114; discussion -85. Epub 2001/08/23.
[2] Byrne EA, Parasuraman R. Psychophysiology and adaptive automation. Biol Psychol. 1996;42:249–268.
[3] Parasuraman R. Neuroergonomics: research and practice. Theor Issues Ergonomics Sci. 2003;4:5–20.
[4] Pattyn N, Neyt X, Henderickx D, et al. Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue? Physiol Behav. 2008;93:369–378.
[5] Parasuraman R, Bahri T, Deaton JE, et al. Available from: theory and design of adaptive automation in aviation systems. http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA254595: Warminster PNAWC, Aircraft Division; 1992.
[6] Durkee K, Geyer A, Pappada S, et al. Real-time workload assessment as a foundation for human performance augmentation. In: D Schmorrow, C Fidopiastis, editors. foundations of augmented cognition. lecture notes in computer science. 8027. Las Vegas (NV): Springer Berlin Heidel Berg; 2013. p. 279–288.

[7] Johnson AW, Oman CM, Sheridan TB, et al. Dynamic task allocation in operational systems: issues, gaps, and recommendations. Aerospace Conference, 2014, IEEE; 2014.
[8] Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: AH Peter, M Najmedin, editors. Advances in Psychology. Volume 52: North-Holland; 1988. p. 139–183.
[9] Shewokis PA, Ayaz H, Panait L, et al. Brain-in-the-loop learning using fNIR and simulated virtual reality surgical tasks: hemodynamic and behavioral effects In: D Schmorrow, C Fidopiastis, editors. Foundations of augmented cognition. 9183. Los Angeles (CA): Berlin Heidel Berg Springer; 2015. p. 324–335.
[10] H Ayaz, S Bunce, P Shewokis, et al., editors. Advances in brain inspired cognitive systems. lecture notes in computer science. 7366. Shenyang, China: Berlin Heidel Berg Springer; 2012; p. 147–155.
[11] Ayaz H, Cakir MP, Izzetoglu K, et al. Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging. Aerospace Conference, 2012 IEEE 2012. p. 1–11.
[12] Gevins A, Smith ME, Leong H, et al. Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. Hum. Factors: J Hum Factors Ergonomics Soc. 1998;40:79–91.
[13] Herff C, Heger D, Fortmann O, et al. Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS. Front Hum Neurosci. 2014;7:935. Epub 2014/01/30.
[14] Grimes D, Tan DS, Hudson SE, et al. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; Florence, Italy; 2008.
[15] Zander T, Kothe C, Jatzev S, et al. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. In: Tan DS, Nijholt A, editors. Brain-computer interfaces. London: Human-Computer Interaction Series Springer; 2010. p. 181–199.
[16] Zander TO, Kothe C. Towards passive brain–computer interfaces: applying brain–computer interface technology to human–machine systems in general. J Neural Eng. 2011;8:025005.
[17] Berka C, Levendowski DJ, Lumicao MN, et al. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviation, space, and environmental medicine. 2007;78:B231–44. Epub 2007/06/06.
[18] Brouwer AM, Hogervorst MA, van Erp JB, et al. Estimating workload using EEG spectral power and ERPs in the n-back task. J Neural Eng. 2012;9:045008. Epub 2012/07/27.
[19] Muhl C, Jeunet C, Lotte F. EEG-based workload estimation across affective contexts. Front Neurosci. 2014;8:114. Epub 2014/06/28.
[20] Ayaz H, Izzetoglu M, Bunce S, et al. Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy. Proceedings of the 3rd international IEEE EMBS Conference on Neural Engineering. Hawaii: IEEE; 2007. p. 342–345.
[21] Ayaz H, Willems B, Bunce S, et al. Estimation of Cognitive Workload during Simulated Air Traffic Control Using Optical Brain Imaging Sensors. In: D Schmorrow, C Fidopiastis, editors. Foundations of augmented cognition directing the future of adaptive systems. lecture notes in computer science. 6780. Orlando (FL): Berlin Heidel Berg Springer; 2011. p. 549–558.
[22] Hong K-S, Naseer N, Kim Y-H. Classification of prefrontal and motor cortex signals for three-class fNIRS–BCI. Neurosci Lett. 2015;587:87–92.
[23] Naseer N, Hong K-S. Decoding answers to four-choice questions using functional near infrared spectroscopy. J Near Infrared Spectrosc. 2015;23:23–31.
[24] Joyce CA, Gorodnitsky IF, Kutas M. Automatic removal of eye movement and blink artifacts from EEG data using blind component separation. Psychophysiology. 2003;41:313–325.
[25] Schlögl A, Keinrath C, Zimmermann D, et al. A fully automated correction method of EOG artifacts in EEG recordings. Clin Neurophysiol. 2007;118:98–104.
[26] Croft RJ, Barry R. Removal of ocular artifact from the EEG: a review. Neurophysiol Clin Clin. Neurophysiol. 2000;30:5–19.
[27] Pfurtscheller G, Allison BZ, Brunner C, et al. The hybrid BCI. Front Neurosci. 2010;4:30.
[28] Leamy D, Collins R, Ward T. Combining fNIRS and EEG to Improve Motor Cortex Activity Classification during an Imagined Movement-Based Task. In: D Schmorrow, C Fidopiastis, editors. Foundations of augmented cognition directing the future of adaptive systems. lecture notes in Computer Science. 6780. Orlando (FL): Berlin Heidel Berg Springer; 2011. p. 177–185.
[29] Fazli S, Mehnert J, Steinbrink J, et al. Using NIRS as a predictor for EEG-based BCI performance. Engineering in Medicine and Biology Society (EMBC). Annual International Conference of the IEEE 2012; 2012. p. 4911–4914.
[30] Fazli S, Mehnert J, Steinbrink J, et al. Enhanced performance by a hybrid NIRS–EEG brain computer interface. NeuroImage. 2012;59:519–529.
[31] Morioka H, Kanemura A, Morimoto S, et al. Decoding spatial attention by using cortical currents estimated from electroencephalography with near-infrared spectroscopy prior information. NeuroImage. 2014;90:128–139. Epub 2014/01/01.
[32] Blokland Y, Spyrou L, Thijssen D, et al. Combined EEG-fNIRS decoding of motor attempt and imagery for brain switch control: an offline study in patients with tetraplegia. IEEE Trans Neural Syst Rehabil Eng. 2014;22:222–229.
[33] Tomita Y, Vialatte FB, Dreyfus G, et al. Bimodal BCI using simultaneously NIRS and EEG. IEEE Trans Biomed Eng. 2014;61:1274–1284.
[34] Khan MJ, Hong MJ, Hong K-S. Decoding of four movement directions using hybrid NIRS-EEG brain-computer interface. Front Hum Neurosci. 2014;8:244.
[35] Yin X, Xu B, Jiang C, et al. A hybrid BCI based on EEG and fNIRS signals improves the performance of decoding motor imagery of both force and speed of hand clenching. J Neural Eng. 2015;12:036004.

[36] Hirshfield LM, Chauncey K, Gulotta R, et al. Combining Electroencephalograph and Functional Near Infrared Spectroscopy to Explore Users' Mental Workload. Proceedings of the 5th International Conference on Foundations of Augmented Cognition Neuroergonomics and Operational Neuroscience: Held as Part of HCI International 2009; San Diego, CA: Springer-Verlag; 2009. p. 239–247.
[37] Coffey EBJ, Brouwer A-M, van Erp JBF. Measuring workload using a combination of electroencephalography and near infrared spectroscopy. Proc Hum Fact Ergon Soc Annu Meet. 2012;56:1822–1826.
[38] Herff C, Fortmann O, Tse C, et al. Hybrid fNIRS-EEG based discrimination of 5 levels of memory load. Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on; 2015; 22–24 April; 2015.
[39] Schalk G, McFarland DJ, Hinterberger T, et al. BCI2000: a general-purpose brain-computer interface (BCI) system. IEEE Trans Biomed Eng. 2004;51:1034–1043.
[40] Ayaz H, Shewokis PA, Bunce S, et al. Optical brain monitoring for operator training and mental workload assessment. NeuroImage. 2012;59:36–47.
[41] Ayaz H, Onaral B, Izzetoglu K, et al. Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: Empirical examples and a technological development. Front Hum Neurosci. 2013;7:871.
[42] Ayaz H, Shewokis PA, Curtin A, et al. Using MazeSuite and functional near infrared spectroscopy to study learning in spatial navigation. J Vis Exp: JoVE. 2011:3443.
[43] Wilson GF, Russell CA. Operator functional state classification using multiple psychophysiological features in an air traffic control task. Hum. Factors: J Hum Factors Ergonomics Soc. 2003;45:381–389. Epub 2004/01/02.
[44] Wilson GF, Russell CA. Real-Time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum. Factors: J Hum Factors Ergonomics Soc. 2003;45:635–643. Epub 2004/04/02.
[45] Ayaz H, Izzetoglu M, Shewokis PA, et al. Sliding-window motion artifact rejection for functional near-infrared spectroscopy. Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE: IEEE; 2010. p. 6567–6570.
[46] Cope M, Delpy DT. System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. Med Biol Eng Comput. 1988;26:289–294.
[47] Merzagora AC, Izzetoglu M, Polikar R, et al. Functional near-infrared spectroscopy and electroencephalography: a multimodal imaging approach. In: D Schmorrow, IEstabrooke, M Grootjen, editors. Foundations of augmented cognition neuroergonomics and operational neuroscience. lecture notes in computer science. 5638. San Diego (CA): Berlin Heidel Berg Springer ; 2009. p. 417–426.
[48] Liu Y, Ayaz H, Curtin A, et al. Towards a hybrid P300-Based BCI using simultaneous fNIR and EEG. In: D Schmorrow, C Fidopiastis, editors. Foundations of augmented cognition. lecture notes in computer science. 8027. Las Vegas (NV): Berlin Heidel Berg Springer; 2013. p. 335–344.
[49] Herff C, Putze F, Heger D, et al. Speaking mode recognition from functional near infrared spectroscopy. Proc IEEE Eng Med Biol Soc. 2012;2012:1715–1718.
[50] Mueller-Putz G, Scherer R, Brunner C, et al. Better than random: a closer look on BCI results. Int J Bioelectromagn. 2008;10:52–55.
[51] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodological). 1995;57:289–300.
[52] Romero S, Mañanas MA, Barbanoj MJ. Ocular reduction in EEG signals based on adaptive filtering, regression and blind source separation. Ann Biomed Eng. 2009;37:176–191.
[53] Halder S, Bensch M, Mellinger J, et al. Online artifact removal for brain-computer interfaces using support vector machines and blind source separation. Comput Intell Neurosci. 2007;7:82069.
[54] Winkler I, Haufe S, Tangermann M. Automatic classification of artifactual ICA-components for artifact removal in EEG signals. Behav Brain Funct. 2011;7:1–15.
[55] Jang KE, Tak S, Jung J, et al. Wavelet minimum description length detrending for near-infrared spectroscopy. J Biomed Opt. 2009;14:034004. Epub 2009/07/02.
[56] Shimodaira H. Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of Statistical Planning and Inference. 2000;90:227–244.
[57] Sugiyama M, Krauledat M. M K-R. Covariate shift adaptation by importance weighted cross validation. J Mach Learn Res. 2007;8:985–1005.
[58] Vidaurre C, Schloogl A, Cabeza R, et al. A fully on-line adaptive BCI. IEEE Trans Biomed Eng. 2006;53:1214–1219.
[59] Blumberg J, Rickert J, Waldert S, et al. Adaptive classification for brain computer interfaces. Engineering in Medicine and Biology Society, 2007 EMBS 2007 29th Annual International Conference of the IEEE; 2007 22-26 Aug; 2007.
[60] Schroeter ML, Zysset S, von Cramon DY. Shortening intertrial intervals in event-related cognitive studies with near-infrared spectroscopy. NeuroImage. 2004;22:341–346.