Decoding Mental Workload in Virtual Environments: A fNIRS Study using an Immersive n-back Task

Decoding Mental Workload in Virtual Environments: A fNIRS Study using an Immersive n-back Task

Felix Putze[1], Christian Herff[2], Christoph Tremmel[3], Tanja Schultz[1] and Dean J. Krusienski[4]

***Abstract*— Virtual Reality (VR) has emerged as a novel paradigm for immersive applications in training, entertainment, rehabilitation, and other domains. In this paper, we investigate the automatic classification of mental workload from brain activity measured through functional near-infrared spectroscopy (fNIRS) in VR. We present results from a study which implements the established n-back task in an immersive visual scene, including physical interaction. Our results show that user workload can be detected from fNIRS signals in immersive VR tasks both person-dependently and -adaptively.**

I. INTRODUCTION

Virtual reality (VR) is a technology that allows the creation of highly immersive scenarios, which can be experiences through multiple senses - including depth vision and spatial sound - and physical interaction to create a feeling of presence in a scene. Usually VR employs a head-mounted display as output device, combined with motion controllers and tracking technology to situate the user in the virtual scene and allow for interaction. Through its unique capabilities, VR has been successfully applied in training, entertainment, rehabilitation, and many other domains. While VR technology has been around for several decades, a recent surge of affordable, high-performance VR headsets has lead to a sharp increase in such applications.

However, the high level of immersion does not come without downsides: First, a VR experience can be stressful and workload-intense to the user (e.g. during a virtual training episode [1]) and avoidance of the responsible scene is often not possible, except for exiting the VR completely. Thus, an adaptation of the VR scene is necessary to ensure the user's well-being and optimal productivity. Second, the communication bandwidth between the user and the VR environment is limited. As available input modalities are often occupied with an ongoing task, they are unavailable for the user to communicate their needs regarding interface adaptations.

An established method to provide such adaptation to the users' needs are brain computer interfaces (BCI) that are used to monitor a person's workload [2]. A BCI records and processes brain activity in real-time, usually based on machine learning technology. Several studies have shown that such a workload measurement can be successfully integrated into a closed-loop system that adapts to the measured workload level: For example, Putze et al. [3] showed how an information presentation system can be adapted in speaking style and content to optimize user error rate and throughput, while Afergan et al. [4] presented similar results for a simulated navigation task. Yuksel et al. [5] demonstrated that a workload-adaptive interface can help piano learners to improve their progress compared to a non-adaptive baseline.

The most viable sensor technologies for measuring brain activity for user interfaces are electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS). While the early literature on workload classification has been dominated by works on EEG, fNIRS has emerged in the last years as a serious alternative [6]. Benefits of fNIRS include a higher spatial resolution compared to EEG and a higher robustness to certain artifacts, such as eye movement or typing [7]. Additionally, as fNIRS optodes do not require the use of gel, an fNIRS headset can usually be mounted quickly. One of the drawbacks of fNIRS is its slow response to changes in the user state, which makes it more suited for detecting of slowly-evolving user states. fNIRS has been successfully used to create passive BCIs classifying user states such as emotions [8] and drowsiness [9]. Some studies also combine both EEG and fNIRS for the classification of workload level [10], [11] or workload type [12].

Numerous fNIRS studies have investigated the classification of mental workload. In recent years, more studies have emerged that move away from highly-controlled and abstracted scenarios, towards more complex and uncontrolled conditions. Examples of such applications involve the detection of workload for pilots [13], car drivers [14], or users of mobile Augmented Reality interfaces [15].

The use of fNIRS-based BCI technology in VR applications has been successfully demonstrated in several studies. Most of these studies concentrate on BCI for use in therapy and rehabilitation of patients. Example applications include the use of BCI using motor imagery for purposes of rehabilitation [16] or neurofeedback for patients with disorders of attention and impulse control [17], [18]. The application of BCI technology for user state monitoring has not been explored much in the literature.

In this paper, we investigate whether methods for workload estimation based on fNIRS during performance of an n-back task can be transferred to complex, immersive VR settings

[1]Felix Putze and Tanja Schultz are with the Cognitive Systems Lab, University of Bremen, Germany {felix.putze|tanja.schultz}@uni-bremen.de

[2]Christian Herff is with the School for Mental Health and Neuroscience, Maastricht University, The Netherlands c.herff@maastrichtuniversity.nl

[3]Christoph Tremmel is with the Biomedical Engineerng Program, Old Dominion University (ODU), Norfolk (VA), USA ctrem003@odu.edu

[4]Dean J. Krusienski is with the ASPEN Lab, Biomedical Engineering Department, Virginia Commonwealth University (VCU), Richmond (VA), USA djkrusienski@vcu.edu

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

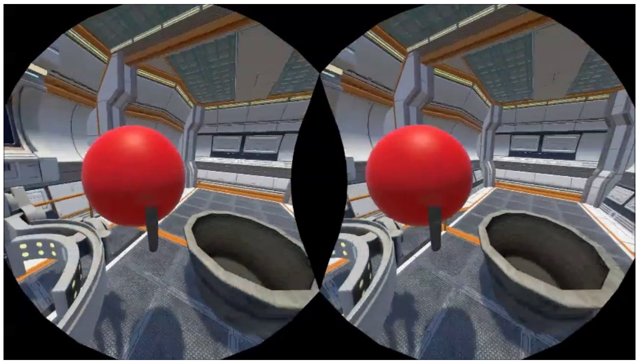

Fig. 1. Binocular first-person view of the n-back task, showing a participant hold a red ball to be placed in the "target" container.

involving physical interaction and visual distraction, which can lead to movement artifacts and other interfering cognitive processes. We investigate both person-dependent and person-adaptive classification modes.

II. MATERIAL AND METHODS

As there is little work on workload recognition from fNIRS in VR and no available data sets, we employed a custom VR scenario to induce different levels of workload in which we record fNIRS data.

A. Experiment Design & Data Acquisition

To induce different levels of workload, we employ the n-back paradigm with three levels of task difficulty. In the classic n-back task, participants are presented with a series of symbols and are asked to respond when the current symbol matched the symbol presented n symbols ago in the sequence [19]. Difficulty is modulated by adjusting the parameter n, in our study from 1 to 3. From previous research it is already established that different difficulty levels induce differences in experienced workload [20].

To transfer the n-back task to a VR setting, we ported it to a factory mock-up scene (Fig. 1). The participant was asked to pick up balls of different colors with the standard VIVE controller at a central "dispenser". Depending on whether a given ball is a "hit" (i.e. its color matches the color of the ball n steps ago) or a miss, the participant must physically turn 90° left or right while holding the ball and place it into the corresponding container. In case a ball was dropped or not placed in a container over the course of a trial, the trial was reset to its initial state. One trial consisted of 10 successfully-placed balls and lasted approximately 46 seconds. As fNIRS workload classification requires a minimum window length of 10 seconds due to the latency of the underlying physiological processes [21], we will regard each complete trial as one sample for classification. The same paradigm was previously used in a workload study with EEG, which demonstrates the general feasibility to induce measurable workload differences [22]. The software to reproduce the experiment is available in an Open Science Framework repository[1].

[1]https://osf.io/yhtz8/

The task was implemented using the Unity framework. For synchronization, we employed the Lab Streaming Layer[2] middleware, with a custom data source for the employed fNIRS device and the LSL4Unity[3] plugin.

Each participant completed a training trial, followed by 24 actual trials, 8 of each difficulty level. The order of the different n-back conditions was pseudo-randomized. Due to a technical error in the first six recordings, 2 of these trials were not transmitted and are thus discarded from analysis.

For the VR hardware, we employed the HTC VIVE. For capturing fNIRS signal, we employed an Oxymon Mark III by Artinis Medical Systems. The recording device used two wavelength of 765 and 856 nm and outputs concentration changes of HbO and HbR. To measure hemodynamic activity in the prefrontal cortex, we attached four transmitter and four receiver optodes to the forehead (Fig. 2). Each detector measures time-multiplexed from two sources, located at a distance of 3.5 cm, resulting in a total of 8 channels each of HbO and HbR. Our signals were sampled at 50 Hz and were downsampled to 10 Hz for processing.

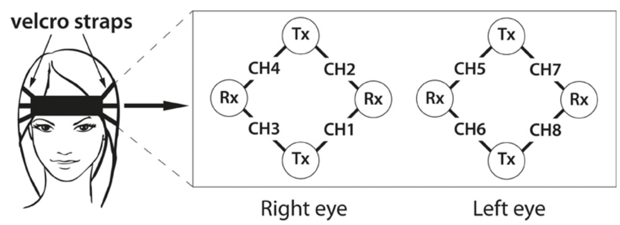

Fig. 2. fNIRS montage at the participant's forehead.

The recording setup on the forehead is very simple and needs less than 3 minutes to be fixed in place and to assess data quality. The VR device setup is then placed on top of the fNIRS headset. Slight individual adjustments of the setup were necessary to accommodate for the placement of the VR headset, which is also placed on the participant's forehead. This usually resulted in the fNIRS optodes being placed slightly above their standard position (Fig. 3). Participants were seated at a fixed position with respect to the VR tracking space to ensure a comparable perspective on the scene. We calibrated the VR tracking before each session. For calibration of the fNIRS device, we adjusted position and pressure of each optode until a minimum device-specific photon count was reached and a clear heart-beat was visible in the unfiltered signal.

Ten students (4 female/6 male, average age 34.6 with a span from 17 to 65 years) participated voluntarily in this study. All participants had normal or corrected to normal vision. The participants were informed prior to the experiment and gave written consent. The total duration of one session was 36 minutes. We chose not to further extend the session duration as multiple pilot participants found donning head-mounted equipment to be fatiguing. The recordings were

[2]https://github.com/sccn/labstreaminglayer
[3]https://github.com/xfleckx/LSL4Unity

Fig. 3. Combined fNIRS and VR montage.

performed at the Biosignal Lab of the Cognitive Systems Lab in Bremen, Germany.

B. Data Processing, Feature Extraction & Classification

Before classification, the fNIRS data were segmented based on the logged timestamps from the n-back task. Windows were 12 s long and locked at the last onset of a trial, to ensure that no further restarts (due to loss of the ball) were included during the window duration. As the workload effects for different values of n begin after the initial trials (i.e., ball presentations), the window was extracted at an offset of two seconds after the start of an experimental run.

The fNIRS signal is contaminated with biological and technical artifacts [23]. To attenuate slow signal trends, we used a moving average filter, which subtracted the mean of the 20 s before and after every sample from every HbO and HbR datapoint. Heart-beat and faster frequency signals are attenuated using an Butterworth IIR low-pass filter with cutoff frequency of 0.5 Hz and filter order of 6. It should be noted that the chosen preprocessing imposes a minimum classification latency of 40 s after the beginning of the trial.

A prototypical hemodynamic response to an increase of mental workload is an increase for HbO in the prefrontal cortex and a return to baseline afterwards. The HbR signals responds inversely and decreases upon stimulus onset and increase back to baseline after the end of the stimulus. To exploit this behavior as features for classification, we calculate the signal mean for all HbO and HbR channels. Additionally, we approximate the signal via linear regression and use the resulting slope and coefficient of determination as features for each channel. As classifier, we employed shrinkage LDA with least squares solver, using empirically-optimized shrinkage coefficients.

III. RESULTS

For evaluation, we investigated two different modes, person-dependent and person-adaptive classification of different workload levels. In each mode, we compare the two binary conditions $n = 1$ vs. $n = 2$ and $n = 1$ vs. $n = 3$ as well as the three-class task (note that the accuracy baseline is different for different classification tasks). For person-dependent classification, we perform a 10-fold cross-validation on the data of each participant individually. This is a challenging setup as in the three-class task, only 22 to 24 trials are available for each person (even less in the two binary tasks), i.e. this analysis gives an estimate of how well workload classification works with little training data. In Figure 5, we report classification accuracy for each participant of the three-class task. The plot shows that for eight of ten participants, the classification accuracy is above the majority baseline of $\frac{1}{3}$ (although in some cases only slightly). The average classification accuracy is 41%. For the binary conditions, mean classification accuracy is 62% ($n = 1$ vs. $n = 2$) and 49% ($n = 1$ vs. $n = 3$), respectively.

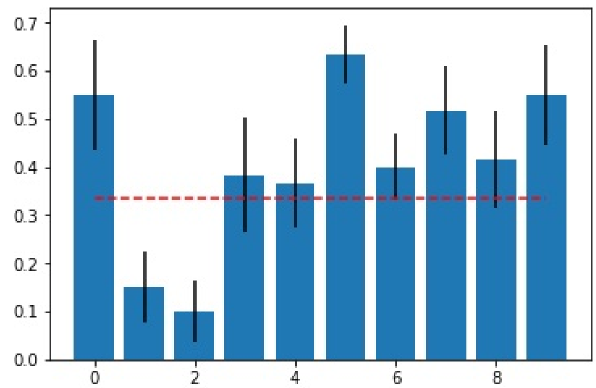

Fig. 4. Mean accuracy (whiskers denote standard deviation) for person-dependent 10-fold cross-validation for all individual participants.

For person-adaptive classification, we pooled the data of all participants together and again performed 10-fold cross-validation. This approach can be applied when a multi-person data set is available and then adapted by a number of person-dependent training trials. In this case, training data fits less well to the specific testing data but may offset this with a larger overall training data set. Figure 5 summarizes the results for the three different classification tasks: For all three tasks, the classifier outperforms the baseline, although standard deviation is relatively high. This observation is in line with the results of the person-dependent classification which shows similar differences between the performance scores for different participants. The average classification accuracy for the three conditions is 66%, 64%, and 42%, respectively. This score is significantly better than the majority baseline for the $n = 1$ vs. $n = 3$ task ($p = 0.029$, $t = 2.58$ for a 1-sample t-test), but not for the other two, likely due to the small sample size. The performance is higher than the corresponding results for the person-dependent classification (on average as well as for 80% of individual participants), indicating that the additional number of training samples offsets the individual differences.

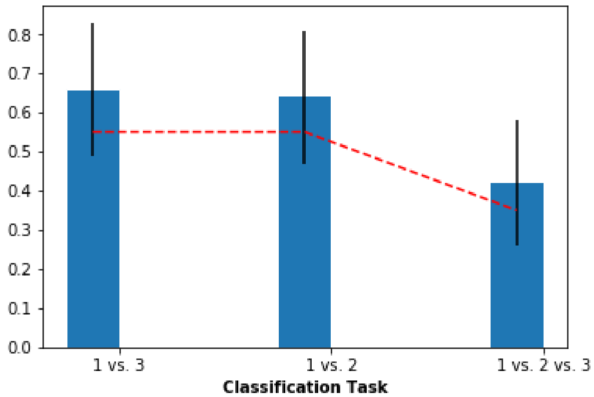

Fig. 5. Mean accuracy (whiskers denote standard deviation) for person-adaptive 10-fold cross-validation for three different classification tasks.

IV. CONCLUSION

In this study, we showed that person-adaptive fNIRS workload classification can be applied to immersive VR tasks, despite the involved physical interaction and visual distraction. This result can act as a baseline for further experiments on adaptive VR environments or classification of other cognitive and affective states. It should be noted that this initial study is limited in a number of ways: First, the number of participants and the number of trials per participant is low. The result that person-adaptive classification outperforms person-dependent classification indicates that more trials per participant could translate to better individual performance. Second, while the study showed promising results for most participants, high standard deviation for person-adaptive classification hint at the possibility of sub-optimal signal acquisition for the low-performing sessions. For future studies, we will thus explore additional optode placements which minimize the interaction between the VR and fNIRS headsets.

V. ACKNOWLEDGEMENT

For the development of the n-back task, we would like to thank Yusuke Yamani from Old Dominion University, Norfolk, VA, USA as well as Krzysztof Rechowicz, Hector Garcia, and John Shull from the Virginia Modeling, Analysis and Simulation Center (VMASC), Suffolk, VA, USA. We also thank Jovana Kaurin for support in recording the data.

REFERENCES

[1] G. Makransky, T. S. Terkildsen, and R. E. Mayer, "Adding immersive virtual reality to a science lab simulation causes more presence but less learning," *Learning and Instruction*, 2017.

[2] A. Girouard, E. T. Solovey, L. M. Hirshfield, E. M. Peck, K. Chauncey, A. Sassaroli, S. Fantini, and R. J. K. Jacob, "From Brain Signals to Adaptive Interfaces: Using fNIRS in HCI," in *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*, ser. Human-Computer Interaction Series, D. S. Tan and A. Nijholt, Eds. London: Springer London, 2010, pp. 221–237.

[3] D. Heger, F. Putze, and T. Schultz, "An EEG Adaptive Information System for an Empathic Robot," *International Journal of Social Robotics*, vol. 3, no. 4, pp. 415–425, 2011.

[4] D. Afergan, E. M. Peck, E. T. Solovey, A. Jenkins, S. W. Hincks, E. T. Brown, R. Chang, and R. J. Jacob, "Dynamic Difficulty Using Brain Metrics of Workload," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014, pp. 3797–3806.

[5] B. F. Yuksel, K. B. Oleson, L. Harrison, E. M. Peck, D. Afergan, R. Chang, and R. J. Jacob, "Learn Piano with BACh: An Adaptive Learning Interface That Adjusts Task Difficulty Based on Brain State," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2016, pp. 5372–5384.

[6] M. Strait and M. Scheutz, "What we can and cannot (yet) do with functional near infrared spectroscopy," *Frontiers in Neuroscience*, vol. 8, 2014.

[7] E. T. Solovey, A. Girouard, K. Chauncey, L. M. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. J. Jacob, "Using fnirs brain sensing in realistic hci settings: experiments and guidelines," in *Proceedings of the 22nd annual ACM symposium on User interface software and technology*. ACM, 2009, pp. 157–166.

[8] D. Heger, C. Herff, F. Putze, R. Mutter, and T. Schultz, "Continuous affective states recognition using functional near infrared spectroscopy," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 113–125, 2014.

[9] M. J. Khan and K.-S. Hong, "Passive BCI based on drowsiness detection: an fNIRS study," *Biomedical Optics Express*, vol. 6, no. 10, pp. 4063–4078, 2015.

[10] C. Herff, O. Fortmann, C.-Y. Tse, X. Cheng, F. Putze, D. Heger, and T. Schultz, "Hybrid fnirs-eeg based discrimination of 5 levels of memory load," in *Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on*. IEEE, 2015, pp. 5–8.

[11] H. Aghajani, M. Garbey, and A. Omurtag, "Measuring Mental Workload with EEG+fNIRS," *Frontiers in Human Neuroscience*, vol. 11, 2017.

[12] F. Putze, S. Hesslinger, C.-Y. Tse, Y. Huang, C. Herff, C. Guan, and T. Schultz, "Hybrid fnirs-eeg based classification of auditory and visual perception processes," *Frontiers in neuroscience*, vol. 8, p. 373, 2014.

[13] T. Gateau, G. Durantin, F. Lancelot, S. Scannella, and F. Dehais, "Real-Time State Estimation in a Flight Simulator Using fNIRS," *PLOS ONE*, vol. 10, no. 3, p. e0121279, 2015.

[14] A. Unni, K. Ihme, H. Surm, L. Weber, A. Ldtke, D. Nicklas, M. Jipp, and J. W. Rieger, "Brain activity measured with fNIRS for the prediction of cognitive workload," in *2015 6th IEEE International Conference on Cognitive Infocommunications*, 2015, pp. 349–354.

[15] R. McKendrick, R. Parasuraman, R. Murtza, A. Formwalt, W. Baccus, M. Paczynski, and H. Ayaz, "Into the Wild: Neuroergonomic Differentiation of Hand-Held and Augmented Reality Wearable Displays during Outdoor Navigation with Functional Near Infrared Spectroscopy," *Frontiers in Human Neuroscience*, vol. 10, 2016.

[16] L. Holper, T. Muehlemann, F. Scholkmann, K. Eng, D. Kiper, and M. Wolf, "Testing the potential of a virtual reality neurorehabilitation system during performance of observation, imagery and imitation of motor actions recorded by wireless functional near-infrared spectroscopy (fNIRS)," *Journal of NeuroEngineering and Rehabilitation*, vol. 7, no. 1, p. 57, 2010.

[17] F. Blume, J. Hudak, T. Dresler, A.-C. Ehlis, J. Khnhausen, T. J. Renner, and C. Gawrilow, "NIRS-based neurofeedback training in a virtual reality classroom for children with attention-deficit/hyperactivity disorder: study protocol for a randomized controlled trial," *Trials*, vol. 18, no. 1, p. 41, 2017.

[18] J. Hudak, F. Blume, T. Dresler, F. B. Haeussinger, T. J. Renner, A. J. Fallgatter, C. Gawrilow, and A.-C. Ehlis, "Near-Infrared Spectroscopy-Based Frontal Lobe Neurofeedback Integrated in Virtual Reality Modulates Brain and Behavior in Highly Impulsive Adults," *Frontiers in Human Neuroscience*, vol. 11, 2017.

[19] K. M. McMillan, A. R. Laird, S. T. Witt, and M. E. Meyerand, "Self-paced working memory: validation of verbal variations of the n-back paradigm." *Brain research*, vol. 1139, pp. 133–142, 2007.

[20] F. Putze, J.-P. Jarvis, and T. Schultz, "Multimodal recognition of cognitive workload for multitasking in the car," in *20th International Conference on Pattern Recognition*. IEEE, 2010, pp. 3748–3751.

[21] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, "Mental workload during n-back taskquantified in the prefrontal cortex using fnirs," *Frontiers in human neuroscience*, vol. 7, p. 935, 2014.

[22] C. Tremmel, C. Herff, and D. Krusienski, "EEG Movement Artifact Suppression in Interactive Virtual Reality," in *Proceedings of the 41st Engineering Medicine and Biology Conference*, Berlin, Germany, 2019.

[23] R. Cooper, J. Selb, L. Gagnon, D. Phillip, H. W. Schytz, H. K. Iversen, M. Ashina, and D. A. Boas, "A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy," *Frontiers in neuroscience*, vol. 6, p. 147, 2012.