2026 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3195-0529-3/26/$31.00 ©2026 IEEE | DOI: 10.1109/VRW70859.2026.00312

Estimating Mental Effort in Virtual Reality Using Multimodal Biosignals

Riku Otono, University of Auckland
Jingruo Chen, Cornell University
Zane Karai, University of Auckland
Kunal Gupta, University of Auckland
Peta Murphy, Laura Fergusson Brain Injury Trust
Jo Nunnerley, University of Otago
Mark Billinghurst, University of Auckland

Figure 1: The dual N-back task in a virtual art gallery and shopping in a virtual supermarket

Abstract

We present a pilot study about estimating mental effort in Virtual Reality (VR) using multimodal physiological signals. A dual N-back task was implemented in VR to induce graded levels of mental effort, assessed using a subjective mental effort questionnaire (SMEQ) and physiological signals (EEG, ECG, and EDA). These signals were used for real-time estimation of mental effort. To examine general applicability, we conducted a preliminary evaluation in an interactive VR shopping environment. Our findings highlight both the feasibility of mental effort estimation in VR and the challenges of evaluating real-time inference in interactive VR contexts.

Index Terms: Mental Effort, Virtual Reality, Physiological Cues

1 Introduction

Mental effort is a core aspect of cognitive state, reflecting the amount of cognitive resources allocated to task demands. Prior work has shown that mental effort is associated with changes in physiological signals, including neural, cardiovascular, electrodermal activity, and pupil dilation [1, 5]. However, it is unclear whether physiological cues collected in controlled VR tasks reliably reflect mental effort, and how to calculate mental effort in real time. This uncertainty is amplified by VR-specific factors, including sensory load, interaction complexity, and embodiment.

To systematically induce graded levels of mental effort in VR, we employed an N-back task [2], which manipulates cognitive load by varying the number of items subjects must maintain in memory. Increasing the N level reliably increases task difficulty and perceived cognitive effort. In this study, a dual N-back task (combining spatial and auditory working memory demands) was implemented in VR and paired with a subjective mental effort questionnaire (SMEQ) to obtain self-reported effort levels [4]. During task performance, we collected multiple physiological signals; electroencephalogram (EEG), electrocardiogram (ECG), and electrodermal activity (EDA). This was to examine how different modalities reflect changes in mental effort under immersive conditions. These signals were used to explore an ML based estimation of graded mental effort in VR. To assess whether this approach extends beyond the N-back task, we evaluated it in a separate shopping activity in a virtual supermarket, where mental effort was induced through task demands and environmental distractions.

The main contribution of this work is that it shows that physiological cues can be used to reliably estimate mental effort in VR, and that it is possible to train a real-time classification system that can be used in a realistic VR task, with some limitations.

2 Data collection with Dual N-back Task

A dual N-back task was implemented in VR and presented within a virtual art gallery environment (Figure 1), viewed in an HTC Vive Pro Eye HMD. In each round, artwork appeared in one of the 2×2 grid locations accompanied by an auditory digit (1–9). Participants identified whether the visual location, the auditory stimulus, or neither matched the stimulus presented N trials earlier. Task difficulty was manipulated using 0-, 1-, 2-, and 3-back conditions presented in pseudorandom order across 36 blocks (12 rounds per block). To minimise emotional confounds, a neutral artwork image was used throughout the task [3]. Each stimulus was presented for three seconds, during which subjects responded, followed by a one-second interval. Subjects reported subjective mental effort using the SMEQ in VR, while EEG, ECG, and EDA were continuously recorded.

Five subjects (2 female, aged 21–34) completed a pilot study. EEG was collected using an eight-channel Unicorn Hybrid Black headset, and ECG/EDA were recorded using a Shimmer GSR+ sensor on the non-dominant hand. The protocol included a 30-second baseline and three test sessions, each of 12 task blocks followed by an SMEQ rating, with short breaks between sessions. Each study took about 90 minutes.

The mean SMEQ score (min: 0, max: 112) in each N(=0, 1, 2, 3) was 0.47 (SD=1.80), 11.62 (SD=7.95), 40.20 (SD=24.83), and 69.49 (SD=22.51), respectively. We ran a Shapiro-Wilk test and found that these scores were not normally distributed, except for N = 3 ($p = 0.47$). So we conducted a Wilcoxon test to compare the data in each N. The data in each N was significantly different from others ($p < 0.001$) (see figure 2), showing that the dual N-back task successfully produced different levels of mental effort.

979-8-3195-0529-3/26/$31.00 ©2026 IEEE
DOI 10.1109/VRW70859.2026.00312

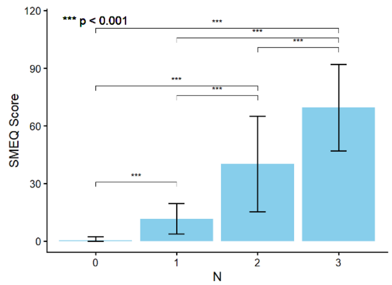

Figure 2: The SMEQ score in each N (0, 1, 2, 3)-back tasks.

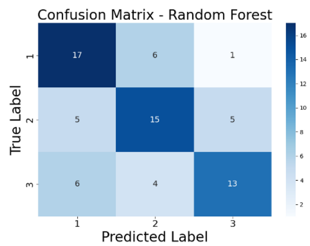

Figure 3: The Random Forest confusion matrix.

3 Mental Effort Classifier Development

A key goal of our work was to develop a real-time mental effort classifier suitable for VR. To do this, we used empirically selected SMEQ thresholds of 7 and 40 to define low, intermediate, and high mental effort. Physiological data from EEG, ECG, and EDA were segmented into 20-second windows with 5-second overlap (sampling rate: 256Hz). From each window, standard statistical features were extracted (EEG: 97, ECG: 21, EDA: 6, 124 features in total), including Delta, Theta, Alpha, Beta, and Gamma EEG bands. To develop the classification model, Random Forest (RF), Gradient Boost (GB), Support Vector Machine (SVM), and Logistic Regression (LR) were compared. The RF model achieved the highest mean F1 score (0.624), followed by GB (0.527), LR (0.513), and SVM (0.459). Cross-validation variance remained substantial across models (e.g., RF: 0.465 ± 0.194), reflecting the small sample size, overlapping windows, and inter-participant variability.

4 Evaluation using Supermarket Environment

In order to see if our approach could be generalized beyond the dual N-back task, we evaluated the RF model in a VR shopping task. This involved subjects interacting with a virtual character to complete a shopping task. Subjects used conversational interaction (enabled by Convai [1]) to talk to the character and request help searching for six items within a budget. Once the subjects found an item, they asked the character to pick it up and put it in the shopping trolley, and where they had to go next. While shopping, we induced cognitive load using visual and auditory distractions (noise, aisle blockages, baby crying) and occasional mental arithmetic prompts.

We recruited three participants and collected SMEQ responses at predefined task events, such as upon reaching an aisle or after picking up an item. While subjects engaged in the task, we estimated mental effort in real time using 20-second physiological windows with a 15-second step size (5-second overlap). The model's output was 0:low, 1:intermediate, and 2:high. We averaged predicted effort levels between SMEQ prompts, determined whether the model's prediction is low, intermediate, or high mental effort, and compared the model's output with the corresponding SMEQ responses. For the three participants, the minimum and the maximum SMEQ scores were 5, 55 respectively (mean: 19.17, SD:12.47). We used the same SMEQ thresholds of 7 and 40 as the model training to define low, intermediate, and high mental effort. Although the task was designed to elicit a range of effort levels, SMEQ responses were biased toward the intermediate level, limiting evaluation of low and high effort states. The classifier's output was also biased, with the intermediate class accounting for 86.1% of all predictions (low: 8.2%, high: 5.7%). Under these conditions, we observed a weighted F1 score of 0.696, which is indicative only due to the small sample size and label imbalance.

[1] https://www.convai.com/

5 Discussion

This pilot study shows that perceived mental effort can be systematically induced and differentiated in VR using a controlled dual N-back task, and that multimodal physiological signals reflect these graded effort levels under immersive conditions. This suggests the feasibility of using VR-based paradigms to study mental effort in a manner that goes beyond traditional desktop settings.

At the same time, there are important challenges in extending mental effort estimation beyond controlled VR tasks. For example, in the VR supermarket scenario, while effort-related patterns learned during the dual N-back task were detectable, most of the results were in the intermediate rate. This could be due to the nature of the task, or the low subject numbers, both causing restricted variability in subjective mental effort.

6 Conclusions and Future Work

In this work, we presented a pilot investigation of mental effort estimation in VR, combining a dual N-back task with multimodal physiological sensing and a preliminary evaluation in an interactive VR environment. The novelty was in exploring the use of multiple physiological cues to identify mental workload, and developing a system for real time classification of workload. Our results show that graded mental effort can be reliably induced in VR under controlled conditions, and that physiological signals can support real-time estimation of mental effort in VR. At the same time, extending mental effort estimation beyond controlled VR tasks remains challenging. In interactive VR environments, task structure and cognitive demand play a crucial role in both inducing and evaluating mental effort. Consequently, models trained under controlled conditions may not directly translate to more naturalistic scenarios without careful task design. These results emphasize that mental effort estimation in VR is not only a sensing or modeling problem, but also a task and evaluation design challenge.

Future work will focus on designing more systematically varied and cognitively demanding VR tasks, collecting data from many more subjects, improving the classification model, and conducting longitudinal evaluations to assess the robustness of the findings. Beyond evaluation, real-time mental effort estimation offers opportunities for adaptive VR systems, such as virtual characters that adjust task difficulty, or feedback based on users' cognitive state.

References

[1] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple task performance*, pp. 279–328, 2020. 1

[2] L. Lawlor-Savage and V. M. Goghari. Dual n-back working memory training in healthy adults: a randomized comparison to processing speed training. *PloS one*, 11(4):e0151817, 2016. 1

[3] S. Mohammad and S. Kiritchenko. WikiArt emotions: An annotated dataset of emotions evoked by art. European Language Resources Association (ELRA), 2018. 1

[4] J. Sauro and J. S. Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 1599–1608, 2009. 1

[5] Wiediartini, U. Ciptomulyono, and R. S. Dewi. Evaluation of physiological responses to mental workload in n-back and arithmetic tasks. *Ergonomics*, 67(8):1121–1133, 2024. 1