Multi-modal Data Collection and Preliminary Statistical Analysis for Cognitive Load Assessment

Ana Krstevska
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
ana.krstevska2001@gmail.com

Sebastjan Kramar
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
sebastjan.kramar@ijs.si

Hristijan Gjoreski
Faculty of Electrical Engineering and Information Technologies
Skopje, Macedonia
hristijang@feit.ukim.edu.mk

Martin Gjoreski
Università della Svizzera italiana (USI)
Lugano, Switzerland
martin.gjoreski@usi.ch

Junoš Lukan
Department of Intelligent Systems
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
junos.lukan@ijs.si

Sebastijan Trojer
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
st5804@student.uni-lj.si

Mitja Luštrek
Department of Intelligent Systems
Jožef Stefan Institute
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Gašper Slapničar
Department of Intelligent Systems
Jožef Stefan Institute
Ljubljana, Slovenia
gasper.slapnicar@ijs.si

Abstract

To mitigate distractions during complex tasks, ubiquitous computing devices should adapt to the user's cognitive load. However, accurately assessing cognitive load remains a significant challenge. This study aims to present sophisticated, multi-modal data collection, which can enable accurate estimation of cognitive load using wearable and contact-free devices. A total of 25 participants participated in six cognitive load-inducing tasks, each presented at two levels of difficulty. Simultaneously, physiological and behavioral data were collected from a multi-modal sensory setup, including: Empatica E4 wristband, Emteq OCOsense glasses, an eye tracker, a thermal camera, a depth camera and an RGB video camera. Additionally, participants provided subjective measures of cognitive load by completing standardized NASA Task Load Index (NASA TLX) and Instantaneous Self-Assessment (ISA) questionnaires following each cognitive task. Preliminary statistical analyses were conducted on participant demographics, performance metrics, and the perceived difficulty of tasks, as reported in the completed questionnaires.

Keywords

cognitive load inference, wearable sensors, contact-free unobtrusive sensors

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.6961

1 Introduction

Human attention is a critical resource that is increasingly targeted by mobile applications, online services, and other forms of digital engagement. In an era of constant connectivity, capturing and retaining user attention has become a primary objective for many technologies. However, as users engage in cognitively demanding tasks, distractions can lead to performance degradation and increased stress. Therefore, to minimize interruptions and maintain productivity, ubiquitous computing systems must become capable of recognizing and adapting to the user's cognitive load in real time.

Cognitive load, defined as the mental effort required to process information and perform tasks, triggers a series of physiological responses in the human body. These responses are largely governed by the activation of the sympathetic nervous system. When cognitive load increases, measurable changes can be observed in physiological markers, including blood pressure, brain activity, eye movements, electrodermal activity (EDA), respiration rate, heart rate variability, etc. Furthermore, changes are also reflected in facial expressions, posture, and other behavioural patterns.

This study seeks to offer a unique multi-modal dataset with a rich set of wearable and unobtrusive sensors to capture the subtle changes that occur with the gradual activation of the sympathetic nervous system. Rather than solely focusing on maximizing data accuracy through the use of numerous devices, this approach also aims to identify the minimum set of sensors required to achieve reliable cognitive load assessment. To that end, rich multi-modal data was collected from a myriad of sensors, including wearables

(OCOsense glasses and Empatica E4 wristband) and contact-free unobtrusive sensors such as an advanced eye tracker, a thermal camera, a depth camera, and an RGB video camera. To the best of our knowledge, no prior dataset exists containing such rich multimodal data obtained with such an elaborate sensory setup.

2 Related Work

The challenge of cognitive load estimation has been extensively studied across various fields. Significant emphasis has been placed on reducing cognitive load in dynamic environments, such as aviation [1]. Recent research has increasingly focused on transitioning from direct measurements, such as electroencephalography (EEG), to indirect methods of cognitive load assessment. For instance, ocular metrics, including pupil diameter and blink rate, have been shown to accurately estimate cognitive load [2, 3, 4]. Additionally, facial temperature variations have been widely correlated with cognitive workload, providing another non-invasive means of assessment [5, 6]. Novak et al. demonstrated that biometric indicators, such as galvanic skin response and skin temperature, can signal increased cognitive load; however, these measures are insufficient to distinguish between varying levels of cognitive load [7]. Wang et al. demonstrated that visual cues—including face pose, eye gaze, eye blinking, and yawn frequency—can serve as indicators of cognitive load [8].

This research aims to address the complexities of cognitive load estimation by integrating a wide range of psychophysiological signals, offering a more comprehensive approach to this task.

3 Experimental Setup

The objective of our data collection was to capture participants' cognitive load under varying levels of difficulty imposed by cognitive load-inducing tasks. The study was conducted in a quiet, temperature-controlled room, with participants tested individually. At the beginning of each session, participants were seated in a comfortable chair in front of a 24" monitor and given instructions about the experiment and their expected role. The Empatica E4 wristband was then fitted to the participant's non-dominant hand, and the OCOsense glasses for emotion recognition were equipped in line with product instructions.

Data collection was further enriched through the use of unobtrusive sensing technologies, including a Tobii Spark eye tracker (60 frames per second), an Intel RealSense Depth Camera D455 (providing depth data at 30 fps), a Logitech BRIO stream 4k webcam at 10 fps with HDR and noise-canceling microphones and a FLIR Lepton 3 thermal camera delivering a full 160x120 pixel thermal resolution with 8 fps. We used this set of devices to continuously monitor participants throughout the recording session. The experimental setup can be observed in Figure 1.

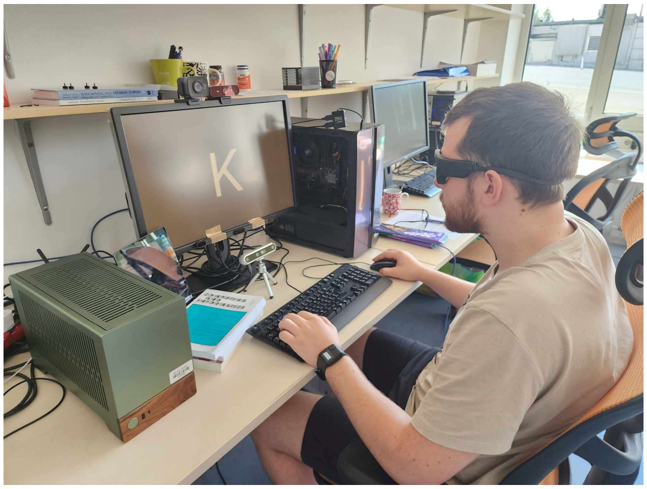

Figure 1: Experimental setup

4 Data Collection Protocol

Prior to the experiment, participants completed a brief sleep questionnaire to gather information about their sleep patterns (e.g., hours slept the night before and usual sleep duration) and rated their levels of fatigue and focus on a scale of 1 to 10.

Calibration data for the OCOsense glasses was then recorded by having participants replicate four facial expressions — smiling, frowning, brow raising, and eye squeezing — three times each. Calibration for the eye tracker followed, during which participants tracked a moving dot with their eyes. This calibration aimed to optimize participant's seating position for accurate eye-tracking.

The experiment's main phase involved participants completing cognitive load-inducing tasks that tested three aspects of cognition: attention, memory, and visual perception. For each cognitive domain, two widely recognized tasks were presented, each with two levels of difficulty (easy and difficult). This design allowed for the differentiation of cognitive load levels. Following each category of cognitive tasks, participants engaged in relaxation tasks that were not expected to induce cognitive load, such as meditation with open eyes, listening to music to relieve stress and passive viewing of aesthetically pleasing images. These tasks provided baseline data for periods of minimal cognitive load.

In summary, each recording session included six cognitive load-inducing tasks (with two levels of difficulty) and three relaxation tasks, totaling 15 tasks. After each task, participants completed the NASA Task Load Index (NASA TLX) questionnaire, a validated instrument for assessing cognitive load across six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration [9]. Each question was rated on a scale of 0 to 100. In this study, the unweighted version of the NASA TLX, known as the Raw NASA TLX, was used. Additionally, participants completed a single-item Instantaneous Self-Assessment (ISA) of workload, providing a subjective measure of the cognitive load induced by the task [10]. These questionnaires served as subjective assessments of cognitive load and as reference points for the difficulty of each task [11].

The tasks were implemented using PsychoPy, an open-source software package commonly used in neuroscience and experimental psychology research [12]. For attention-related tasks, participants completed the N-back and Stroop tests. In the N-back task, participants were presented with a sequence of letters and asked to determine whether the current letter matched the one presented N trials earlier (with task difficulty increasing as N increased) [13]. Participants completed both a 2-back and a 3-back task. In the Stroop test, participants identified whether the word matched the color in which it was written, with the easier version involving two colors (red and blue) and the more difficult version incorporating five colors [14].

Memory-related tasks included a memory game and a question-answering task based on a previously shown image. In the memory game, participants recalled as many words as possible from a set, with the easier version comprising seven words and the more difficult version consisting of 15 words. In the question-answering task, participants focused on an image and then answered questions about it (e.g., remembering the number of particular objects in the image), with the hard version using an image with greater detail.

The visual perception tasks included a "spot the difference" task and a pursuit test. In the "spot the difference" task, participants were presented with two images and were asked to identify as many differences as possible within a one-minute time frame. The difficulty of this task varied, with the more challenging version involving an image that contained greater detail compared to the simpler, easier version. The pursuit test required participants to visually track irregularly curved, overlapping lines. As with the "spot the difference" task, the pursuit test was administered at two levels of difficulty. The more difficult version featured a more intricate image, with longer and more tangled lines, as opposed to the less complex image used in the easier version of the task.

5 Statistics

In this section, we present some descriptive demographic and task-related statistics for the participants involved in the experiment. The average age of participants was 29.28 years, with a standard deviation of 8.31. In terms of educational background, the majority of participants (44 %) had obtained a Bachelor's degree (BSc), followed by those with a Master's degree (MSc), 28 %. A smaller portion had completed only high school (16 %) or had earned a PhD (12 %). Additionally, 60 % of the participants were male.

We then looked at the descriptive statistics derived from the performance of the participants in each task. These indicate that participants performed consistently well on tasks such as the 2-back task, both easy and difficult versions of the Stroop test, the easy memory task (where participants recalled an average of 5 out of 7 words), the easy version of the "spot the difference" task (with an average detection rate of approximately 90 % of all the differences), and both versions of the pursuit test. Notably, participants performed slightly better on the difficult version of the Stroop test, likely due to their increased familiarity with the task.

However, performance was lower on tasks such as the 3-back test (which most participants perceived as highly or extremely difficult), the difficult memory task (with an average recall rate of 39 %), and both the easy and difficult question-answering tasks. The difficult version of the "spot the difference" task also showed lower performance, with participants detecting only 25 % of the differences on average. Consistent performance among subjects (with low standard deviation) was observed across all tasks except for the N-back tasks. Notably, the N-back tasks were always presented first to participants, suggesting that they may have required additional time to adjust to the testing environment and fully engage with the task.

Next, an inferential statistical analysis was performed on the relationship between task scores and various variables of the sleep pattern. To investigate the potential influence of tiredness on performance, responses from the sleep patterns questionnaire were analyzed. A non-parametric Kruskal-Wallis test was performed to determine whether there was a statistically significant difference in overall scores across different levels of tiredness (low, medium, and high). The resulting *p*-value (0.91) indicated no significant difference in performance between these groups. Thus, tiredness levels did not show a statistically significant impact on performance within a 95 % confidence interval.

Similarly, the effect of focus level (low vs. high) on overall performance was examined using a non-parametric Mann-Whitney test. The *p*-value was 0.12, indicating no statistically significant difference in performance between low and high focus groups at the 5 % significance level.

Furthermore, the relationship between hours of sleep the night before the experiment and participant performance was examined using Spearman's correlation. The *p*-value was 0.42, indicating no statistically significant correlation between overall performance scores and hours of sleep the night before the experiment.

The potential influence of participants' highest education level on overall performance was also investigated. To assess this, a non-parametric Kruskal-Wallis test was conducted. The results (*p*-value of 0.33) indicated no statistically significant difference in performance scores across different education levels among the participants.

Overall, the small sample size may have constrained the ability to detect significant effects. The limited variability in the sample's educational background and other factors likely contributed to the lack of observed differences, emphasizing the need for a larger, more diverse sample to better understand the impact of these variables on cognitive load performance.

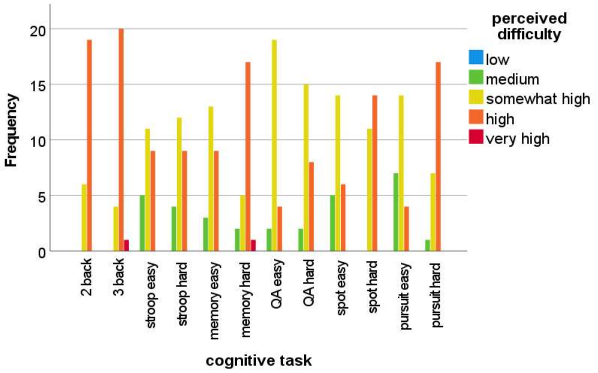

Figure 2: Reported perceived difficulty per cognitive task

As shown in Figure 2, participants consistently perceived the difficulty of the two N-back tasks and the difficult version of the "spot the difference" task as somewhat high or high. This suggests a general consensus regarding the difficulty of these tasks. In contrast, the NASA TLX-based perceived difficulty of remaining tasks, exhibited significant variability among participants.

To assess differences in performance across task difficulties and evaluate the potential for differentiating cognitive load using machine learning models, we conducted additional inferential statistical analyses. The Wilcoxon signed-rank test was used to compare participant performance on the easier and more difficult versions of each cognitive task.

Statistically significant differences in performance were found between the two difficulty levels for all tasks. For the N-back, "spot the difference", and pursuit tasks, participants performed significantly better on the easier versions, indicating that increased difficulty negatively impacted performance. Conversely, for the Stroop, memory, and question-answering tasks, participants performed better on the more difficult versions.

The statistical analysis conducted in this study provides initial evidence supporting the validity of the data collection protocol, particularly with respect to the selection of tasks and task difficulty levels. The tasks chosen for this experiment varied significantly in terms of their cognitive demands, as reflected by the substantial differences in performance between the easier and more difficult versions of each task. These results indicate that cognitive load and performance are task-specific, and the significant differences observed support the feasibility of using machine learning models to differentiate between varying levels of cognitive load.

6 Conclusion and Future Work

This study employs a novel approach to data collection for cognitive load inference by combining psychophysiological data obtained from multi-modal sensory setup, including wearable and unobtrusive contact-free sensors. The decision to utilize a diverse set of devices was motivated by the hypothesis that integrating data from multiple sources could provide a more accurate assessment of cognitive load, while also aiming to identify the minimal sensor configuration required to achieve reliable results. This is particularly relevant in dynamic and high-stakes environments, such as driving, where accurate cognitive load assessment could have life-saving implications. To the best of our knowledge, no prior research has incorporated such a comprehensive and multifaceted setup for cognitive load evaluation.

The statistical analyses conducted thus far offer promising validation for the data collection protocol. The selection of tasks and task difficulty levels proved effective in eliciting a range of cognitive load levels, as evidenced by the significant performance differences between task difficulties.

To further enhance the validity of the data collection protocol, several changes could be implemented in potential subsequent collections. Refining task difficulty levels could offer more granularity in cognitive load differentiation, ensuring a clearer distinction between varying levels of cognitive load. Furthermore, increasing the diversity of participants in terms of age, educational background, and other demographic factors is desirable to enhance the generalizability of the findings.

In future work, the collected data will be processed and utilized to train machine learning models aimed at estimating cognitive load. Ground truth for the machine learning models can be derived from various sources, including perceived task difficulty reported through the standardized questionnaires, the designed difficulty level of the tasks or the participants' performance on the tasks. These machine learning models will leverage sophisticated ML techniques to effectively integrate and analyze multi-modal data, aiming to enhance the accuracy of cognitive load predictions. We also plan to further expand the dataset with another phase of data collection, offering a rich dataset both in terms of modalities, as well as in terms of participants. The collected dataset will serve as a stepping stone towards robust multi-modal cognitive load assessment, allowing for creation and benchmarking of ML models and will be made available to general public after the collection is finalized.

Acknowledgements

This work was supported by the Jožef Stefan Institute and Università della Svizzera italiana (funded by SNSF through the project XAI-PAC (PZ00P2_216405)).

References

[1] Jonathan Mead, Mark Middendorf, Christina Gruenwald, Chelsey Credlebaugh, and Scott Galster. 2017. Investigating Facial Electromyography as an Indicator of Cognitive Workload. In *19th International Symposium on Aviation Psychology*, 377–382.

[2] Muneeb Imtiaz Ahmad, Ingo Keller, David A. Robb, and Katrin S. Lohan. 2020. A framework to estimate cognitive load using physiological data. *Personal and Ubiquitous Computing*, 27, 2027–2041.

[3] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject workload classification using pupil-related measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 4, 1–8.

[4] Tobias Appel, Natalia Sevcenko, Franz Wortha, Katerina Tsarava, Korbinian Moeller, Manuel Ninaus, Enkelejda Kasneci, and Peter Gerjets. 2019. Predicting Cognitive Load in an Emergency Simulation Based on Behavioral and Physiological Measures. In *Proceedings of the 21st ACM International Conference on Multimodal Interaction (ICMI)*, 154–163.

[5] Fangqing Zhengren, George Chernyshov, Dingding Zheng, and Kai Kunze. 2019. Cognitive load assessment from facial temperature using smart eyewear. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*, 657–660.

[6] Yomna Abdelrahman, Eduardo Velloso, Tillman Dingler, Albrecht Schmidt, and Frank Vetere. 2017. Cognitive Heat: Exploring the Usage of Thermal Imaging to Unobtrusively Estimate Cognitive Load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 33, 1–20.

[7] Klemen Novak, Kristina Stojmenova, Grega Jakus, and Jaka Sodnik. 2017. Assessment of cognitive load through biometric monitoring. In *7th International Conference on Information Society and Technology (ICIST)*.

[8] Zixuan Wang, Jinyun Yan, and Hamid Aghajan. 2012. A framework of personal assistant for computer users by analyzing video stream. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 1–3.

[9] Sandra G. Hart, and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, 52, 139-183

[10] Andrew J. Tattersall, and Penelope S. Foord. 2007. An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39, 740-748.

[11] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. 2023. A Survey on Measuring Cognitive Workload in Human-Computer Interaction. *ACM Computing Surveys*, 55, 1–39.

[12] Jonathan Peirce, Rebecca Hirst, and Michael MacAskill. 2022. Building Experiments in PsychoPy. Sage Publications

[13] Michael J. Kane, and Andrew Conway. 2016. The invention of n-back: An extremely brief history. *The Winnower*

[14] John Ridley Stroop. 1992. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 121, 15-23