Predicting Cognitive Failures in Virtual Reality Using Pupillometry

Sahar Niknam
VR/AR Lab
University of Luxembourg
Esch-sur-Alzette, Luxembourg
sahar.niknam@uni.lu

Jean Botev
VR/AR Lab
University of Luxembourg
Esch-sur-Alzette, Luxembourg
jean.botev@uni.lu

***Abstract*—Pupil dilation has consistently been investigated and confirmed as a reliable measure of cognitive load. In this study, we aim to explore the possibility of predicting cognitive failures in Virtual Reality by monitoring variations in pupil dilation during cognitive processing. To this end, we collected eye-tracking data from an individual performing a mental arithmetic task over two months, totaling 700 minutes. We achieved promising prediction results by training a neural network on the collected data, particularly considering the dataset's imbalanced nature. The ability to predict impending cognitive failures generally holds significant implications across various domains, including education, delegating decision-making tasks to autonomous systems, or self-adaptive virtual environments and user interfaces.**

***Index Terms*—cognitive load, pupillometry, neural networks, virtual reality**

I. Introduction

Estimating cognitive load with pupil dilation has been a well-established practice over the past half-century, yielding reliable results. In this project, we initiate a novel investigation, exploring the potential of pupil dilation as a predictor of cognitive failure.

The radial and circular muscles in the iris, which are responsible for pupil dilation and constriction, are controlled by sympathetic and parasympathetic nervous systems, respectively. Therefore, pupil size varies under the influence of non-visual stimuli, including emotional arousal and cognitive efforts [1]. Consequently, in a visually static environment without emotionally arousing stimuli, an observer can measure cognitive load robustly by monitoring pupil diameters. This has been repeatedly confirmed in recent decades, thanks to advancements in eye-tracking technology facilitating its use in everyday tasks and activities [2]. Furthermore, various studies have examined the variation in pupil dilation as a time series, indicating its validity as an indicator of attention [3], [4] and decision-making conflicts [5].

On the other hand, Cognitive Load Theory (CLT) [6] suggests that cognitive performance is constrained by the capacity of working memory, which is spatially and temporally limited. This implies that individuals have a limited cognitive load capacity. Based on these two premises, we hypothesize that there are patterns in pupil dilation variations as an individual reaches their cognitive load limit. By detecting these patterns, we can predict when cognitive load exceeds working memory capacity. Whether the working memory is overfilled by relevant information (i.e., an overwhelming cognitive demand) or irrelevant cognitive input (i.e., a lapse of attention), it causes a failure. In this project, we use *cognitive failure* as an umbrella term covering both cases and investigate potential signatures of such failures in pupil dilation changes.

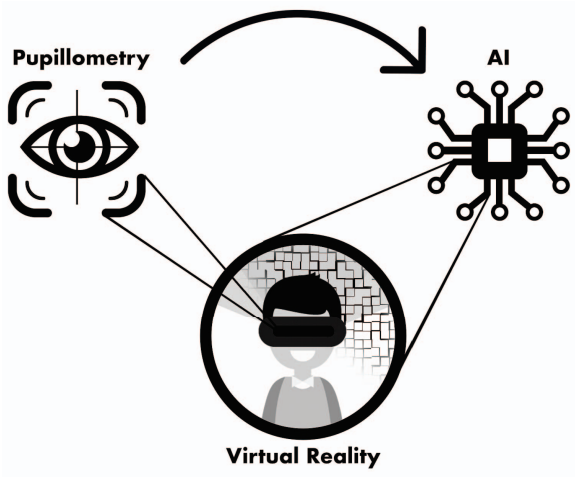

Fig. 1. Employing AI to instantly and continuously adapt the VR environment to the user's cognitive needs in response to pupil variation patterns.

The ability to predict cognitive failures would be highly valuable, with extensive applications spanning across diverse domains, from tailored learning pace in education to intelligent user interface design with the capacity to instantly re-strategize the flow of information communication.

Such predictive power can be especially an invaluable asset for self-adaptive virtual reality (VR) environments. The VR community is increasingly interested in self-adaptive technologies, such as deep learning, for designing environments capable of sensing data and automatically making adjustments to meet users evolving needs in a closed-loop design [7]–[11]. These technologies are being researched and implemented in different domains, such as medical training, commercial VR, collaborative and remote workspaces, rehabilitation, and serious games [7]. Cognitive load is one of the most effective types of data fueling self-adaptive technologies [9], [12]–[16]. This project is built on the same perspective with the feedback

2771-7453/24/$31.00 ©2024 IEEE
DOI 10.1109/AIxVR59861.2024.00043

loop depicted in Fig. 1, schematically illustrating how AI can adapt the VR environment to the user's needs in response to pupil variation patterns, for example, to reduce cognitive load or focus attention.

II. Methodology

In this study, over a period of 2 months, we recorded the eye-tracking data of one of the authors engaging in a mental arithmetic task during 70 sessions, each lasting 10 minutes.

A. Implementation

The VR-based implementation of the experimental design allowed us to fully control the environmental parameters and ensure the scalability of the study for more complex tasks and settings. In particular, scene brightness remains constant so that confounding effects of pupil light response on the cognitive-emotional response can be avoided [17]–[19]. Eye-tracking data were collected with the Tobii eye-tracker[1] integrated into the HTC Vive Pro Eye VR headset with a 120 Hz sampling rate[2] (Fig. 2). Multiple studies investigated and confirmed the reliability of pupillometry in VR, particulary using the HTC Vive Pro Eye VR headset, which was used in this study [19]–[23]. Furthermore, the experiment environment was designed to keep the illumination constant and the visual content at a minimum level to avoid pupil diameter variation due to irrelevant parameters.

The task in this study is mental arithmetic, specifically multiplication, as it presents a cognitively demanding yet straightforward task. We compiled a list of 522 questions and updated it with more difficult questions when the mistake rate dropped stably below 5%. Our aim was to list questions that challenged the participant without being too difficult to discourage engagement. We sorted the questions into three levels of difficulty based on the magnitude of the multiplicands and other considerations, such as the presence of 11 or a product of 5 as a multiplicand. We gradually introduced more challenging questions after two consecutive correct answers to simpler ones or reset the difficulty level after an incorrect response. However, given the absence of a standardized grading system for the difficulty of the questions and the unobservable nature of individuals' tailored shortcuts or strategies for mental calculation, we did not include the difficulty as a parameter in our dataset features.

Each session consists of a series of trials, beginning with a question phase followed by an answer phase. During the question phase, the questions are displayed in large white font against a backdrop of solid blue. The participant has 15 seconds to provide the answer, and with no response within the time limit, a new question is shown. In this phase, the question is masked, and the participant can access a virtual keypad to enter the answer. The keypad is operated through hand gestures detected by a Leap Motion controller[3] attached to the headset (cf. Fig. 3).

[1]https://www.tobii.com/products/integration/xr-headsets/device-integrations/htc-vive-pro-eye

[2]https://www.vive.com/sea/product/vive-pro-eye/overview/

[3]https://www.ultraleap.com/product/

Fig. 2. Experiment setup.

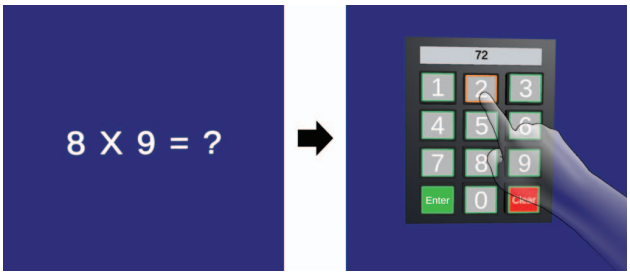

Fig. 3. User interface of the experiment in VR.

B. Data Preprocessing

In the preprocessing stage, we worked with the eye-tracking data collected during the question phase before the participant entered her answers. In this initial phase of the study, we only used the pupil dilation parameters for both eyes and their corresponding validity values. However, we chose not to remove invalid readings, as they may indicate blinks or squeezed eyes, which we hypothesized could be correlated with the task. Finally, we extracted the 3-second interval (360 eye-tracking readings) before the answer phase and dropped the data points with less than 3 seconds in the question phase.

Data points with correct answers were labeled 0, and those with wrong answers 1. To address the data scarcity, we excluded the third class, which represents instances where participants did not provide an answer within the given time limit. Out of the remaining 3055 samples (shape 4×360), only 299 samples were labeled 1.

C. Training

We randomly assigned 99 samples from each class to our test set and used the remaining for training. To overcome the imbalance, we applied the data weighting technique. Using `tf.keras` [24], we implemented a sequential model consisting of 3 LSTM layers {8, 32, 128} with Tanh activation, followed by 3 Dense layers {128, 32, 8} with linear activation, and a Dense output layer with Softmax activation. The model was trained using the Adam optimizer with a learning rate of 0.001 for 100 epochs.

III. Results

On the balanced test set with a total number of 198 samples, we reached an accuracy of 70.71%. The confusion matrix (cf. Table I) shows a false-positive rate of 42.42% and a true-positive rate of 85.86%. We achieved 0.669 for the precision, 0.859 for recall, and an overall F1 score of 0.752.

TABLE I
Confusion Matrix

		Predicted	
		Positive	Negative
Actual	Positive	85	14
	Negative	42	57

IV. Discussion

Artificial neural networks have demonstrated exceptional predictive power thanks to their ability to find intricate patterns in data. Primarily, these networks are trained for population-level applications. However, in this study, we aim to develop personalized neural networks tailored to predict individual cognitive patterns. Training individuals for critical jobs, such as a pilot or a surgeon, is a rigorous and financially demanding process, typically spanning many years. Compared to such a training process, collecting tens of hours of an individual's data for training a network capable of predicting an impeding cognitive failure seems a reasonable investment.

Our model's performance on the evaluation metrics demonstrates promising results, particularly given the significant class imbalance and the small size of the dataset. These findings highlight the potential of our ongoing project in predicting cognitive failures through pupil dilation patterns. With adding more 'wrong answer' samples, we anticipate achieving a similar level of accuracy with positive sample classification since the model showed a substantially lower level of confidence in classifying these samples compared to the negative class.

Ongoing work focuses on collecting more data for the subject. With more data samples, we can include other eye-tracking features, such as null fixations or saccades, potentially improving the model's accuracy. Additionally, a larger, feature-richer dataset allows us to include the third class of unanswered questions, representing a distinct cognitive event.

The next step involves determining a prediction horizon. In this phase, we artificially separated the decision-making phase from cognitive processing by the left-click event. However, in real-life scenarios, it is crucial to have a horizon that provides warning of an impending cognitive failure within a specific time interval. Even a brief horizon, measured in milliseconds, can significantly enhance the practicality of such technology.

With its limitation to a specific cognitive task and individual, this work is preliminary in nature, and the positive response of deep learning to this case needs to be reinforced by demonstrating the generalizability of the results to other individuals and cognitive tasks. Therefore, we plan to continue this research with different and more complex tasks. Moreover, achieving comparable results with more participants is critical to ensure that the predictability observed in our preliminary findings is not solely attributed to the pupillometry characteristics of a single subject. By expanding the dataset, recruiting more participants, and using diverse tasks, we can thoroughly examine the generalizability of patterns in pupil dilation and the model's ability to adapt robustly to individual differences and task variations.

V. Conclusion

Customizing and training a neural network for an individual may initially seem inefficient due to the challenges of collecting enough data samples. However, personalized AI presents a strong argument when the individual in question is responsible for a critical task involving real-time decision-making. This work anticipates a promising path for applying neural networks in detecting cognitive patterns through pupillometry. With ever-evolving AI and VR technologies, predicting a cognitive failure even seconds in advance gives us a wide range of options, from delegating tasks to auto-pilot systems to re-rendering the virtual environment to modulate the cognitive load.

Acknowledgment

This work is a part of the larger EU-funded ChronoPilot project [25] and received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964464.

References

[1] K. G. Seeber and D. Kerzel, "Cognitive load in simultaneous interpreting: Model meets data," *International Journal of Bilingualism*, vol. 16, no. 2, pp. 228–242, 2012.

[2] P. van der Wel and H. van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: A review," *Psychonomic Bulletin & Review*, vol. 25, pp. 2005–2015, 2018.

[3] O. Kang and T. Wheatley, "Pupil dilation patterns reflect the contents of consciousness," *Consciousness and Cognition*, vol. 35, pp. 128–135, 2015.

[4] ——, "Pupil dilation patterns spontaneously synchronize across individuals during shared attention," *Journal of Experimental Psychology: General*, vol. 146, no. 4, pp. 569–576, 2017.

[5] J. W. de Gee, T. Knapen, and T. H. Donner, "Decision-related pupil dilation reflects upcoming choice and individual bias," *Proceedings of the National Academy of Sciences*, vol. 111, no. 5, pp. E618–E625, 2014.

[6] J. Sweller, J. J. G. Van Merrienboer, and F. G. W. C. Paas, "Cognitive Architecture And Instructional Design," *Educational Psychology Review*, vol. 10, pp. 251–296, 1998.

[7] N. Vaughan, B. Gabrys, and V. N. Dubey, "An overview of self-adaptive technologies within virtual reality training," *Computer Science Review*, vol. 22, pp. 65–87, 2016.

[8] K.-C. Siu, B. J. Best, J. W. Kim, D. Oleynikov, and F. E. Ritter, "Adaptive Virtual Reality Training to Optimize Military Medical Skills Acquisition and Retention," *Military Medicine*, vol. 181, no. 5 Suppl, pp. 214–220, 2016.

[9] C. I. Aguilar Reyes, D. Wozniak, A. Ham, and M. Zahabi, "Design and evaluation of an adaptive virtual reality training system," *Virtual Reality*, vol. 27, no. 3, pp. 2509–2528, 2023.

[10] M. S. Aquino, F. F. De Souza, and A. C. Frery, "Vepersonal: an infrastructure of virtual reality components to generate web adaptive environments," in *Proc. 11th Brazilian Symposium on Multimedia and the Web*, 2005, pp. 1–8.

[11] C. Baker and S. H. Fairclough, "Adaptive virtual reality," in *Current Research in Neuroadaptive Technology*. Elsevier, 2022, pp. 159–176.

[12] C. McDonald, M. Davis, and C. Benson, "Using Evidence-Based Learning Theories to Guide the Development of Virtual Simulations," *Clinical Social Work Journal*, vol. 49, no. 2, pp. 197–206, 2021.

[13] P. Ramakrishnan, B. Balasingam, and F. Biondi, "Cognitive load estimation for adaptive human-machine system automation," in *Learning control*, D. Zhang and B. Wei, Eds. Elsevier, 2021, pp. 35–58.

[14] U. Lahiri, E. Bekele, E. Dohrmann, Z. Warren, and N. Sarkar, "Design of a virtual reality based adaptive response technology for children with autism," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, no. 1, pp. 55–64, 2012.

[15] C. Tremmel, C. Herff, T. Sato, K. Rechowicz, Y. Yamani, and D. J. Krusienski, "Estimating Cognitive Workload in an Interactive Virtual Reality Environment Using EEG," *Frontiers in Human Neuroscience*, vol. 13, 2019.

[16] J. T. Doswell and A. Skinner, "Augmenting Human Cognition with Adaptive Augmented Reality," in *Proc. International Conference on Augmented Cognition (AC)*. Springer, 2014, pp. 104–113.

[17] B. John, P. Raiturkar, A. Banerjee, and E. Jain, "An Evaluation of Pupillary Light Response Models for 2D Screens and VR HMDs," in *Proc. 24th ACM Symposium on Virtual Reality Software and Technology (VRST)*, 2018.

[18] H. Chen, A. Dey, M. Billinghurst, and R. W. Lindeman, "Exploring Pupil Dilation in Emotional Virtual Reality Environments," in *Proc. International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*, 2017.

[19] J. Lee, N. De Jong, J. Donkers, H. Jarodzka, and J. Van Merriënboer, "Measuring Cognitive Load in Virtual Reality Training via Pupillometry," *IEEE Transactions on Learning Technologies*, pp. 1–7, 2023.

[20] P. Stark, T. Appel, M. J. Olbrich, and E. Kasneci, "Pupil Diameter during Counting Tasks as Potential Baseline for Virtual Reality Experiments," in *Proc. Symposium on Eye Tracking Research and Applications (ETRA)*, 2023, pp. 1–7.

[21] A. Hebbar, S. Vinod, A. K. Shah, A. Pashilkar, and P. Biswas, "Cognitive load estimation in VR flight simulator," *Journal of Eye Movement Research*, vol. 15, no. 3, 2022.

[22] J. Orlosky, B. Huynh, and T. Hollerer, "Using Eye Tracked Virtual Reality to Classify Understanding of Vocabulary in Recall Tasks," in *Proc. IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2019, pp. 66–73.

[23] L. J. Zheng, J. Mountstephens, and J. Teo, "Four-class emotion classification in virtual reality using pupillometry," *Journal of Big Data*, vol. 7, pp. 1–9, 2020.

[24] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," 2015. [Online]. Available: https://www.tensorflow.org/

[25] J. Botev, K. Drewing, H. Hamann, Y. Khaluf, P. Simoens, and A. Vatakis, "ChronoPilot – Modulating Time Perception," in *Proc. IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, 2021, pp. 215–218.