2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782124

Pupil-Linked Arousal is Predictive of Team Performance in a Virtual Reality (VR) Sensory-Motor Task

Yinuo Qin[1], Weijia Zhang[1], Richard Lee[1], Xiaoxiao Sun[1], and Paul Sajda[1,2]

Abstract— **The performance of teams can be significantly affected by the physiological arousal of each team member. Pupil dynamics has been shown to index, among other things, arousal. In this paper, we describe a multi-person virtual reality (VR) based sensory-motor task, where individuals must cooperate with each other as co-pilots to navigate a virtual spacecraft. We analyze the relationship between team performance and pupil-linked arousal as a function of the task's difficulty level. We find that pupil dynamics across the team members is significantly correlated with team performance and task difficulty. We also find that the pupil dynamics of individual team members appear to correlate with situational awareness, where the pupil dynamics of these team members can be used to predict upcoming task success.**

I. Introduction

In the 1995 movie *Apollo 13* [1], there is a scene where three astronauts work together and manually adjust the re-entry angle of their spacecraft in order to land on Earth safely. In the scene, each astronaut controls a different degree of freedom of the spacecraft and has a different view of the environment. The astronauts are highly focused yet under tremendous stress during the maneuver. An interesting scientific question is how their arousal state influenced their performance and if there were potential physiological biomarkers that could indicate if they were headed toward a successful maneuver. In this paper, we set up a multi-person virtual reality platform that simulates the *Apollo 13* maneuver. We name this task **A**pollo **D**istributed **C**ontrol **T**ask (ADCT). While a team of three individuals (triad team) performs the re-entry task, we simultaneously record and analyze each individual's pupillometry.

Pupil dynamics has been directly linked to human arousal [2, 3]. An open research question in neuroscience is how activity in the cerebral cortex is modulated by subcortical nuclei that regulate arousal. Arousal is generally described as a psychological concept associated with cognitive and emotional states that depend upon the activation of the sympathetic nervous system, the autonomic nervous system, or the endocrine system [4]. The state of arousal can significantly affect a person's ability to make optimal decisions and actions in real-world dynamic environments [5].

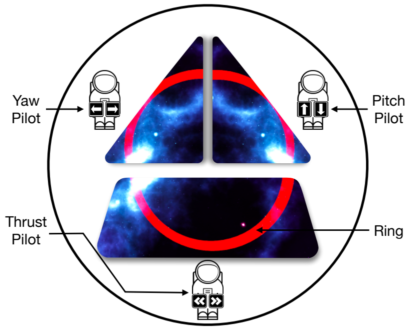

Fig. 1. ADCT environment and experimental setup. The experiment involves three participants, each assigned a distinct role and a corresponding viewing window. These viewing windows provide varied and partial views of the environment. The Yaw-Pilot is responsible for controlling the left-right movement of the spacecraft. The Pitch-Pilot manages the up-down movement, while the Thrust-Pilot controls the spacecraft's speed.

Previous work has shown that individual pupil-linked arousal is correlated with task performance [6, 7] and task difficulty levels [7–9]. However, most real-world tasks require multiple people to work together as a team. The question of how individual arousal changes can impact team performance and be influenced by task difficulty levels under immersive and heightened levels of arousal tasks remains unknown. In this paper, we investigate how multiple participants' pupil dynamics, as an index of arousal, predict task performance and task difficulty when performing the team-based sensory-motor task in VR.

II. Method

A. Experimental Setup and Data Collection

The main focus of this work is to study how each individual's pupil dynamics are predictive of team performance and task difficulty. To achieve this, we built a virtual reality environment that requires three participants to collaborate and cooperate with each other and navigate a spacecraft through the re-entry path marked by red rings. The triad team must reach a final destination in a limited time. Failing to pass any ring or running out of time will result in mission failure. As shown in Fig. 1, each participant is assigned a unique viewing window to observe the environment, with each window offering a distinct perspective. This experiment was approved by Columbia University's Institutional Review Board (IRB) for human subjects research.

Each participant wore an HTC VIVE Pro Eye Head Mounted Display (HMD) with a built-in Tobii eye-tracking system to view the task environment. Participants used the VIVE Pro Controller to control the spacecraft and external microphones to communicate with each other. We

This work was supported by funding from the Army Research Laboratory's STRONG Program (W911NF-19-2-0139, W911NF-19-2-0135, W911NF-21-2-0125) the National Science Foundation (IIS-1816363, OIA-1934968) and a Vannevar Bush Faculty Fellowship from the US Department of Defense (N00014-20-1-2027).

[1] All authors are with Department of Biomedical Engineering, Columbia University. yinuo.qin, wz2540, rtl2118, xiaoxiao.sun@columbia.edu [2] Paul Sajda is with Department of Electrical Engineering, Department of Radiology, and Data Science Institute, Columbia University. psajda@columbia.edu

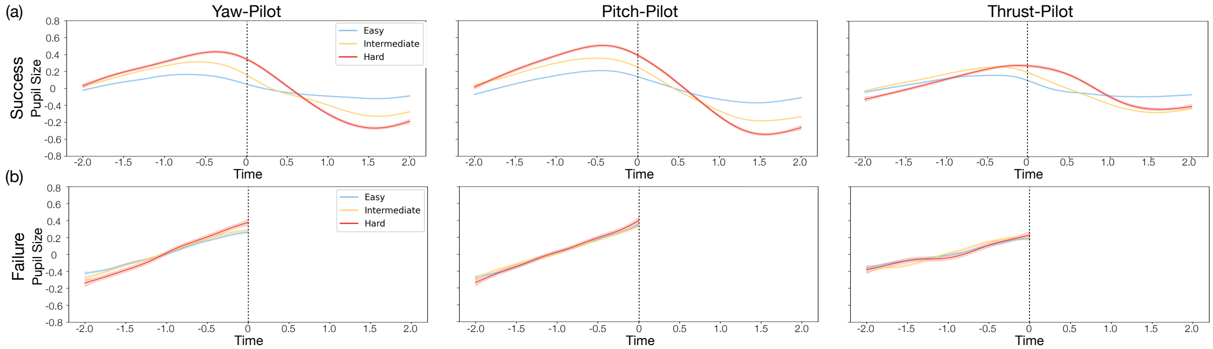

Fig. 2. The predicted Z-scored pupil sizes (with 95% *CI*) from the results of all the GAM analyses (both fail and pass situations). Results are plotted as a function of role (*yaw, pitch, thrust*) from left to right. Within each role, curves having different colors indicate different difficulty levels (i.e., easy: red; medium: blue; hard: purple) with different line styles indicating different communication modes (i.e., no: solid; single-word: dash; free: dotted). The solid-filled dot on each fitting curve marks the maximum of that curve.

use LabStreamingLayer [10] to collect pupillometry of all participants in each team. Each team participates in a series of 135 experimental trials involving 15 rings per trial. The difficulty level of the task correlates with the radius of the rings. In tasks categorized as easy, the radius of a ring is 2.5 times greater than the radius of the spacecraft. For medium or hard tasks, the respective ring radii are 2 and 1.5 times the radius of the spacecraft.

B. *Participants*

There were 33 participants (17 males, 16 females) between 19 to 28 years old ($M = 22.91$, $SD = 2.63$) in our study. Participants were split into groups of three, and each group participated in the experiment three times on distinct days. All participants in this experiment were recruited through voluntary participation forms.

C. *Data Processing*

We first filtered out invalid pupil data from each participant due to blinking or tracking loss. Then, we applied linear interpolation for brief gaps not exceeding one second. After Z-scoring the data, we segmented it into epochs based on ring interaction outcomes. Successful passes were analyzed in four-second epochs, capturing 2 seconds before and after passing the ring. For failed attempts, we used two-second epochs, focusing on the period immediately before the missed ring, as trials ended instantly after a failure.

D. *Statistical Analysis*

We used a generalized additive model (GAM) on data from both pass and fail situations to model the changes in participants' pupil sizes across time. A GAM offers greater flexibility and sensitivity relative to generalized linear models (GLM) and ANOVA to detect the underlying nonlinear trend between variables [11]. By using smooth nonlinear functions, a GAM can capture the dynamic changes of the relationship between variables [12]. Our GAM analysis is computed using the *Python* package *pyGAM* [13].

$$g(\mu|t) = \beta_0 + f(t) \tag{1}$$

Here, $g(.)$ is the link function. μ is the pupil size of one participant and t is the time. β_0 is the intercept term. $f(t)$ is a function of time.

E. *Multilayer Perceptron Neural Network*

We implemented a multi-layer perceptron (MLP) model to test whether pupil dynamics can be used to predict team performance. Team performance is measured by whether a team can successfully pass a ring. We focused on pupil dynamics leading up to the ring, specifically in 1 second before the ring. The train test validation split is $0.9, 0.05, 0.05$, respectively, employing a subject-independent approach where data from all subjects are mixed together, ensuring that the model's performance is not biased towards any specific individual's data. We used a network consisting of eight fully connected layers with 256 neurons for each layer and ReLU as the activation function for hidden layers. We can write the layer l and row i of the MLP model as,

$$f_i(x) = ReLU(\sum_{j=1}^{n} w_{ij}^{(l)} x_j + b_i^{(l)}), \tag{2}$$

where ReLU

$$ReLU = max(x, 0). \tag{3}$$

Here, the n is the length of the input. w_j are the weights, and x_j are the input pupil dynamics of the MLP model. From the second to the last layer, the inputs of the model are the outputs from the previous layer. The b_i is the bias.

Subsequently, a similar Multilayer Perceptron (MLP) was constructed to predict the task difficulty level based on pupillometry data from participants. This MLP aims to ascertain whether pupil size correlates with task difficulty in scenarios where the team successfully completes the task. Therefore, only epochs where the team successfully navigated through the rings were selected for analysis. These epochs include pupil size data from 2 seconds before to 2 seconds after passing a ring. The difficulty-focused MLP comprises eight fully connected layers, each with 512 neurons, to accommodate the larger input size associated with this task.

Note that we did not try to optimize the network. Instead, we used it as a prediction engine to evaluate the predictive information in the pupil dynamics across the individuals and integrate as a team. We used the accuracy on a left-out test set to evaluate performance, resampling the data multiple times to estimate confidence intervals on our predictions.

III. Results

A. *Pupil-Linked Arousal Predicts Team Performance*

Previous research has demonstrated that pupil size changes depend on task performance in multiple object tracking or short-term memory tasks [6, 14]. However, the relationship between individual pupil size variation and team performance remains an unresolved question. As Fig. 2 (a) shows, pupil size exhibits an increase followed by a decrease before successfully navigating through the ring object. In contrast, the pupil size of each subject demonstrates a monotonic increase before failing the task. The reduction in pupil size observed before successful navigation suggests that pupil-linked arousal begins to recover as individuals anticipate successfully passing through the ring. These results indicate that the pupil size of each team member is correlate with situational awareness and has the potential to predict the overall performance of the team.

Role	Prediction Accuracy
Yaw	0.5629±0.0284
Pitch	0.6274±0.0250
Thrust	0.6053±0.0357
Yaw + Pitch	0.6857±0.0312
Yaw + Thrust	0.6954±0.0288
Pitch + Thrust	0.6635±0.0288
All	**0.7113 ± 0.0241**

TABLE I

Task Performance Prediction Accuracy

TABLE I summarizes all MLP networks' prediction accuracies. In this study, the baseline accuracies are established at 0.5. This is because we have a balanced number of data in training, validation, and testing datasets. We utilized the pupil sizes of individual roles and combinations of multiple roles to train and test the MLP networks. The bold prediction accuracy indicates the best performance among all networks. As anticipated, the model's prediction accuracy reaches its peak when input includes the pupil sizes of all team members. The accuracy drops significantly ($p < 10^{-4}$) when the model relies on the pupil size of an individual or the combined pupil sizes of any two roles. These findings suggest that in team-based tasks, the ability of an individual to predict team performance is inherently limited.

B. Pupil-Linked Arousal Predicts Task Difficulty Levels

Fig.2.(a) illustrates the impact on pupil dynamics amplitude as a function of task difficulty level. Notably, the amplitude of the pupil dynamics is significantly different ($p < 10^{-8}$) for all roles. As tasks become more challenging, the participants' pupil size amplitude correspondingly rises. Under the hypothesis that pupil size reflects arousal levels, our findings indicate heightened arousal dynamics in more difficult tasks. This is evidenced by increased arousal when approaching a ring and a remarkable decrease or recovery in arousal upon successful navigation. Additionally, the participant's role distinctly affects pupil dynamics relative to the task difficulty level. Both Yaw-Pilot and Pitch-Pilot exhibit more significant changes in pupil-linked arousal compared to the Thrust-Pilot. This implies that managing the spacecraft's directional movement induces a higher level of arousal in an individual than controlling its speed.

Interestingly, when the task was not successfully completed, variations in pupil size amplitude across different levels of task difficulty for Pitch-Pilot and Thrust-Pilot did not exhibit statistical significance (PitchPilot: $p = 0.7532$, ThrustPilot: $p = 0.2348$). However, for Yaw-Pilot, a significant difference in pupil size amplitude was observed under varying task difficulties ($p < 0.05$), even in cases of task failure (Fig. 2 (b)). These outcomes suggest a role-dependent variation in situational awareness within team-based activities. Specifically, certain roles can distinguish task difficulty regardless of the team's overall success or failure in completing the task. In contrast, most roles are more likely to discern task difficulty levels primarily in the context of successful task execution. This distinction highlights the complex interplay between individual perception and team dynamics in task-oriented scenarios.

Role	Prediction Accuracy
Yaw	0.3781±0.0431
Pitch	0.3682±0.0456
Thrust	0.4040±0.0087
Yaw + Pitch	0.5274±0.0311
Yaw + Thrust	0.5505±0.0087
Pitch + Thrust	0.5505±0.0532
All	**0.6162 ± 0.0631**

TABLE II

Task Difficulty Prediction Accuracy

Next, we trained a series of MLP networks to assess the predictability of pupil-linked arousal on different task difficulty levels. As delineated in TABLE II, the model achieves the highest accuracy when inputs encompass pupil sizes from all team members (in bold text). All MLP networks in this study have a baseline accuracy of 0.3333. We selected an equal number of easy, intermediate, and hard trials for each role in the training, validation, and testing datasets. A significant reduction in accuracy ($p < 10^{-7}$) is evident when inputs are limited to the pupil size of an individual or a pair of participants. This result supports that although individual pupil dynamics are affected by task difficulty changes, the entire team's pupil dynamics are best used to yield the highest accuracy for predicting task difficulty levels.

IV. Discussion

A. The Effect of Task Performance on Pupil-Linked Arousal

As illustrated in Fig. 2, pupil dynamics exhibit distinct patterns in response to task success or failure. Specifically, in failure cases, there is a consistent increase in pupil-linked arousal before the task fails. In contrast, for successful tasks, pupil-linked arousal begins to decrease prior to task completion. Such decreases exist for all roles under all difficulty levels. These observations suggest that pupil-linked arousal is predictive of task performance. This inference is further supported by the neural network prediction results presented in TABLE I. The MLP model that takes pupil dynamics data as inputs can accurately classify team performance approximately 70% of the time. Therefore, it can be inferred that pupil-linked arousal is a reliable predictor of team performance in this virtual reality sensory-motor task. Notably, individuals within teams, each with different roles, exhibit varying degrees of predictive power regarding overall team performance. Previous works on single human tasks have shown similar results [6, 7]. However, a comprehensive and accurate prediction of team performance requires incorporating physiological data from all team members.

B. The Effect of Task Difficulty on Pupil-Linked Arousal

Our results demonstrate that the pupil dynamics of individuals within teams vary significantly across different task difficulty levels. Specifically, we observed an increase in the rate and magnitude of pupil dilation leading up to a task, with this effect being more pronounced as the task difficulty increases. Additionally, the subsequent constriction of the pupil size is more substantial with increasing task difficulty. These findings suggest that pupil dynamics serve as "windows" into internal brain states, particularly those related to arousal and cognitive load [2, 3]. This supports the statement that that pupil dynamics are a reliable indicator of task difficulty for individuals in team-based collaborative settings. The results from the MLP classification model, as detailed in TABLE II, further corroborate the predictive capability of individual pupil dynamics in relation to task difficulty. However, it is important to note that the model attains its optimal classification accuracy specifically when it includes pupil dynamics data from all team members. This highlights the crucial role of collective physiological responses in accurately understanding and forecasting the challenge of the tasks in intricate, team-based tasks.

C. The Effect of Role on Pupil-Linked Arousal

Our results depicted in Fig. 2 illustrate that the pupil dynamics of participants, specifically in terms of dilation and oscillation, are closely related to their assigned role in the task. Notably, the magnitude of these dynamics in individuals assigned as Thrust-Pilot is significantly lower (as indicated by Fig. 2 (a)) compared to Yaw-Pilots or Pitch-Pilots. Given that pupil dynamics are a recognized indicator of cognitive arousal [2, 3], these findings imply a comparatively lower level of arousal changes in Thrust-Pilots. This observation aligns with the task's differing demands of the control roles. Success in maneuvering the spacecraft relies more on the precise control of yaw and pitch than on thrust.

D. Pupil-Linked Arousal and the Effect of Illumination

We hypothesized that pupil dynamics index cognitive and physiologically relevant states. However, previous work has shown that luminance changes in a virtual environment strongly influence pupil sizes [15, 16]. These studies have shown that the human pupil constricts with increasing luminance in HMDs. In ADCT, the rings exhibit a greater luminosity compared to the space background. Thus, as participants approach a ring, the luminance gradually increases, and pupil dynamics should decrease in size. However, As shown in Fig. 2, the pupil dynamics of all roles are dilated when approaching the ring. Moreover, by analyzing the time of each individual focused on any objects in the virtual environment, we found the focusing time is less than 0.17‰ of the total time of each trial. Thus, the pupil dynamics we recorded reflect the arousal changes of participants instead of illumination change.

V. Conclusion

Understanding team dynamics requires insight into both behavioral and physiological interactions among individuals. In virtual reality team environments, monitoring participant's arousal through pupil dynamics linked to autonomic and cognitive arousal systems proves to be a significant indicator of underlying team dynamics. This study demonstrates that pupil-linked arousal provides valuable insights into various performance aspects of a demanding triad-team sensorimotor task. Notably, pupil-linked arousal is a biomarker that reflects task difficulty and team performance. Hence, measuring such biomarkers may be useful for inferring the state of the team and expected performance. This potentially offering a way to use a closed-loop system to regulate arousal levels across team members for improving individual and collective performance. Further research is needed to test these biomarkers in a wider range of tasks for better generalizability.

References

[1] B. Grazer, *Apollo 13*. Universal City Studios, Inc., 1995.

[2] J. Bradshaw, "Pupil size as a measure of arousal during information processing," *Nature*, vol. 216, no. 5114, pp. 515–516, 1967.

[3] M. R. Nassar, K. M. Rumsey, R. C. Wilson, K. Parikh, B. Heasly, and J. I. Gold, "Rational regulation of learning dynamics by pupil-linked arousal systems," *Nature neuroscience*, vol. 15, no. 7, pp. 1040–1046, 2012.

[4] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology*, vol. 76, no. 5, p. 805, 1999.

[5] J. Faller, J. Cummings, S. Saproo, and P. Sajda, "Regulation of arousal via online neurofeedback improves human performance in a demanding sensory-motor task," *Proceedings of the National Academy of Sciences*, vol. 116, no. 13, pp. 6482–6490, 2019.

[6] B. Wahn, D. P. Ferris, W. D. Hairston, and P. König, "Pupil sizes scale with attentional load and task experience in a multiple object tracking task," *PloS one*, vol. 11, no. 12, p. e0168087, 2016.

[7] C. K. Foroughi, C. Sibley, and J. T. Coyne, "Pupil size as a measure of within-task learning," *Psychophysiology*, vol. 54, no. 10, pp. 1436–1443, 2017.

[8] S. Saproo, V. Shih, D. C. Jangraw, and P. Sajda, "Neural mechanisms underlying catastrophic failure in human–machine interaction during aerial navigation," *Journal of neural engineering*, vol. 13, no. 6, p. 066005, 2016.

[9] Y. Qin, W. Zhang, R. Lee, X. Sun, and P. Sajda, "Predictive power of pupil dynamics in a team based virtual reality task," in *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 2022, pp. 592–593.

[10] T. Stenner, C. Boulay, M. Grivich, D. Medine, C. Kothe, tobiasherzke, G. Grimm, xloem, A. Biancarelli, B. Mansencal, chausner, J. Frey, kyucrane, S. Powell, P. Clisson, and phfix, "sccn/liblsl: v1.16.0," Mar. 2022. [Online]. Available: https://doi.org/10.5281/zenodo.6387090

[11] T. Hastie and R. Tibshirani, "Exploring the nature of covariate effects in the proportional hazards model," *Biometrics*, pp. 1005–1016, 1990.

[12] D. Ruppert, M. P. Wand, and R. J. Carroll, "Semiparametric regression during 2003–2007," *Electronic journal of statistics*, vol. 3, p. 1193, 2009.

[13] daniel servén, C. Brummitt, H. Abedi, and hlink, "dswah/pygam: v0.8.0," Oct. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1476122

[14] M. T. Kucewicz, J. Dolezal, V. Kremen, B. M. Berry, L. R. Miller, A. L. Magee, V. Fabian, and G. A. Worrell, "Pupil size reflects successful encoding and recall of memory in humans," *Scientific reports*, vol. 8, no. 1, p. 4949, 2018.

[15] Y.-G. Cherng, T. Baird, J.-T. Chen, and C.-A. Wang, "Background luminance effects on pupil size associated with emotion and saccade preparation," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.

[16] S. Koorathota, K. Thakoor, L. Hong, Y. Mao, P. Adelman, and P. Sajda, "A recurrent neural network for attenuating non-cognitive components of pupil dynamics," *Frontiers in Psychology*, vol. 12, p. 12, 2021.