ORIGINAL ARTICLE

Pupillometry in telerobotic surgery: A comparative evaluation of algorithms for cognitive effort estimation

Paola Ruiz Puentes[1,#], Roger D Soberanis-Mukul[1,#,*], Ayberk Acar[2], Iris Gupta[1], Joyraj Bhowmick[1], Yizhou Li[2], Ahmed Ghazi[3], Peter Kazanzides[1], Jie Ying Wu[2], Mathias Unberath[1]

[1]Laboratory for Computational Sensing and Robotics (LCSR), Department of Computer Science, School of Engineering, Johns Hopkins University, Baltimore, MD 21205, USA

[2]Machine Automation, Perception and Learning (MAPLE) Lab, Department of Computer Science, Vanderbilt Institute for Surgery and Engineering, School of Engineering, Vanderbilt University, Nashville, TN 37235, USA

[3]Division of Minimally Invasive and Robotic Surgery, Johns Hopkins Brady Urological Institute, School of medicine, Johns Hopkins University, Baltimore, MD 21205, USA

ABSTRACT

Background: Eye gaze tracking and pupillometry are emerging topics in telerobotic surgery as it is believed that they will enable novel gaze-based interaction paradigms and provide insights into the user's cognitive load (CL). Further, the successful integration of CL estimation into telerobotic systems is thought to catalyze the development of new human-computer interfaces for personalized assistance and training processes. However, this field is in its infancy, and identifying reliable gaze and pupil-tracking solutions in robotic surgery is still an area of ongoing research and high uncertainty. **Methods:** Considering the potential benefits of pupillometry-based CL assessments in telerobotic surgery, we seek to better understand the possibilities and limitations of contemporary pupillometry-based cognitive effort estimation algorithms in telerobotic surgery. To this end, we conducted a user study using the da Vinci Research Kit (dVRK) and performed two experiments where participants were asked to perform a series of N-Back tests, either while (i) idling or (ii) performing a peg transfer task. We then compare four contemporary CL estimation methods based on direct analysis of pupil diameter in the spatial and frequency domains. **Results:** We find that some methods can detect the presence of cognitive effort in simple scenarios (*e.g.*, when the user is not performing any manual task), they fail to differentiate the different levels of CL. Similarly, when the manual peg transfer task is added, the reliability of all models is compromised, highlighting the necessity of more robust methods that consider different factors that complement the pupil diameter information. **Conclusion:** Our results offer a quantitative perspective of the limitations of the current solutions and highlight the necessity of developing tailored designs for the telerobotic surgery environment.

Key words: Pupillometry, eye gaze tracking, cognitive load, calibration, telerobotic surgery

INTRODUCTION

Pupil diameter and gaze tracking are emerging research topics in computer-assisted interventions and telerobotic surgery. It is hypothesized that knowledge of gaze position and pupil dilation will open new avenues for research in different fields. For example, in developing

[#]These authors contributed equally to this work.

***Corresponding Author:**

Roger D Soberanis-Mukul, Laboratory for Computational Sensing and Robotics (LCSR), Department of Computer Science, School of Engineering, Johns Hopkins University, Baltimore, MD 21205, USA. E-mail: rsobera1@jhu.edu; https://orcid.org/0000-0003-0712-0534

Received: 15 July 2023; Revised: 24 July 2023; Accepted: 7 August 2023; Published: 31 August 2023

https://doi.org/10.54844/mr.2023.0420

new intuitive human-computer interaction paradigms[1–3] and in skill assessment for the development of personalized learning experiences.[4–6] This is possible because eye-related metrics, like pupil diameter, gaze, and blink-rate, are correlated to the level of cognitive load (CL) a user is experiencing in a given situation.[7–9]

CL theory was developed in 1988,[10,11] and has three main components: Sensory, working, and long-term memories.[10,11] Information flows from the sensory memory into the working memory, where it is processed. Then, the brain categorizes it and moves it into the long-term memory, where it is stored in knowledge structures called schemas. The more practiced those schemas, the more effortless that behavior becomes. CL relates to the amount of unfamiliar information the working memory is currently processing. When the CL exceeds the working memory capacity, it leads to mental fatigue and decreases learning speed.[12–14] In an optimal scenario, CL should be small for surgeons with high expertise while in an optimal learning point for surgeons in training.[15,16] From the telerobotic surgery perspective, advancing in eye-related CL estimation can allow the development of tools like safety assistance models able to effectively detect and track fatigue during surgery, and personalized surgical training programs for an enhanced learning experience.[17,18]

CL presents a direct relationship with the pupil diameter, which has led to the development of methods based on this metric.[8] They have been used is learning experiences,[19] and complex decision-making scenarios,[20] among other studies. However, approaches that make a direct use of the diameter are sensitive to illumination changes, and hence their application to unconstrained environment is limited. Multiple challenges arise in unconstrained environments. For instance, the changes in pupil diameter due to light variations is an order of magnitude larger than the response triggered by CL.[21,22] Additionally, the pupil has a maximum diameter from which it can no longer dilate, meaning the same source of CL can produce different changes in the diameter. Furthermore, since each person’s mental capacity and basal pupil diameter differ, the same activities can trigger different CL responses.[22,23]

Recent studies have proposed ways to overcome these limitations and propose to perform the analysis of the diameter in the spectral domain.[24,25] These works hypothesize that such analysis can separate changes due to reactions to light from changes related to effortful cognitive processes. These methods are based on the knowledge that changes in CL are related to high-frequency variations, while low-frequency changes are responses to luminescence. Additionally, they have been applied to comprehension studies,[26] direct and gradual video-speed adjustments for learning,[19] color visual short-term memory tasks,[27] among others. Considering the success of both pupil diameter-based and frequency analysis methods. This paradigm brings promising applicability to telerobotic surgery, however their current performance in this task has yet to be fully explored.

Contributions

Motivated by the recent advances in CL estimation, and their applications in telerobotic surgery, we evaluate a set of state-of-the-art models for CL estimation under a telerobotic environment. To this end, we perform a user study on the da Vinci Research Kit (dVRK)[28] where the participants are asked to perform a visual-manual task at the time they solve an auditive N-Back task to trigger different CL levels. Our results offer a quantitative perspective of the limitations of the current solutions and highlight the necessity of developing tailored designs for the telerobotic surgery environment.

MATERIAL AND METHODS

We compared four methods for CL estimation based on the analysis of the pupil diameter. In this section, we introduce the main ideas of each method and describe the data collection protocol employed to obtain pupil signals under different levels of CL.

Eye-tracker and pupillometry-based cognitive effort estimation

CL estimation has been of great interest in human center design, as it defines non-observable internal factors related to the user experience when performing a task. Different physiological measures like heart rate variability, electrodermal activity, pupil diameter, and blink rate reveal information about user’s CL. However, attending the necessities of users and use cases, CL estimation methods should not interfere with the natural workflow. Under this condition, eye-tracking measurements offer solution for non-invasive CL assessment, and hence, most of the methods employ eye-tracker information to quantify CL.

Pupil diameter has been a standard measurement to estimate CL, given its correlation with task difficulty.[9,24,25] Previous works show that complex problems cause an increment in pupil diameter motivating models that employ this fact to determine CL by comparing current measurements with a baseline diameter (basal diameter). The basal diameter describes the natural behavior of the pupil, and it is acquired during a baseline task performed independently (inter-trial change in pupil diameter called BCPD), or at the beginning of each step of the main task (intra-trial change in pupil diameter called CPD).[9] Both changes can be defined as:

$$CPD = \frac{k}{k+1}CPD + \frac{1}{k+1}(\bar{x}(t) - \mu_{T_b})$$

Where k starts at 0 and increases in steps of 1 during the temporal range CPD or BCPD is evaluated, $t \in [0, T_e]$ represents the temporal range of the trial, and μ_{Tb} is the mean diameter obtained during the baseline trial. $\bar{x}$ represents the pupil signal after applying a Butterworth filter.[9,29] Both CPD and BCPD represent the change in pupil diameter, which can be positive or negative and are presented as unitless in the paper.[9,29] These methods can reveal evidence of CL and have the advantage of being easy to implement. However, given their complete dependence on the raw pupil diameter their performance can be affected by changes in the lighting conditions.

Analysis in the spectral domain

To address these limitations, a second group of methods considers the fluctuations in the pupil diameter during the trial and perform an analysis of the rate of change of the diameter over time. Changes in CL are related to high-frequency variations, while low-frequency changes are responses to luminescence. Methods like the index of cognitive activity (ICA) analyze the rate of change in the pupil signal instead of directly comparing the pupil diameter with a baseline. According to Duchowski *et al.*, the ICA can separate light-related reflexes from CL-related responses (dilation reflex).[24] Even though the details of the implementation of ICA are not available, the authors propose an alternative measure based on the ratio of high-frequency responses of the pupil signal named the index of pupillary activity (IPA) and based on the wavelet decomposition of the pupil signal, and is presented as a unitless metric.[24] IPA performs a hard thresholding to filter the wavelet coefficients and analyses the remaining using the frequency of the count of coefficients per second. CL is quantified according to the number of counts, as they are directly proportional to the effort. The method was evaluated with 13 participants performing easy, difficult, and control tasks consisting of number-counting trials. Participants wore eye trackers with their heads stabilized by a chin rest in a room with limited ambient light. Even though the paper shows the sensibility of IPA to changes in CL, the constrained experimental environment is a factor to consider when applying the method to a different domain. Additionally, IPA might fail when estimating CL induced by the N-Back test, even when the gaze is fixed at a specific location.[25]

Duchowski *et al.* also proposed the Low/High Index of Pupillary Activity (LHIPA) to overcome the limitations of IPA. Instead of taking the maxima of the frequency signal, LHIPA takes the ratio between low and high frequency bands.[25] The high-frequency response is expected to increase under high CL, and hence, LHIPA will decrease under the presence of CL. LHIPA is also presented as a unitless metric. To test LHIPA, participants were asked to perform number counting, N-Back, and a modified text-copy tasks. The first task was performed with a fixed gaze, the N-Back was performed by fixing the gaze point at one of five positions on the screen, and the third problem required an unrestricted environment. Experiments were more complex than the presented by IPA, the users could change the position of the gaze point in a controlled or unrestricted way, allowing for more realistic use cases. The room's illumination during the experiments was kept reduced and constant.

In-house dataset for cognitive effort evaluation in telerobotic surgery

We defined a data collection protocol to record pupil information of users while performing manual and auditive tasks with the dVRK.[28] Users were asked to perform a teleoperated peg transfer and an auditory N-Back tasks (Table 1). We collected pupil information from 18 participants, which all were graduate computer science students from the Johns Hopkins University with basic knowledge about the dVRK. The participants were divided in two groups as follows: (i) 13 users performed peg transfer and N-Back tasks simultaneously, and (ii) 5 users performed the N-Back tasks without peg transfer. The first group replicates a real surgical scenario, where the surgeon is performing the handling of the robot arms with additional cognitive tasks like communication with its team and evaluating the next steps of the surgery. The second group enables us to assess the effect on each CL-detection method of performing multiple tasks. Data was collected using the Pupil Labs Invisible frame with adapted camera mounts (Figure 1) under an approved protocol (HIRB00014648). Personalized mounts were needed due to the angle of the head inside the dVRK that did not allow accurate tracking of the Pupil Labs Core headset.

Users are first introduced to the basic use of the dVRK, followed by a practice session to get familiar with the device. No recording is performed during this first step. After this training session, users are introduced to the protocols. Once all the introductory information is completed, the users are asked to wear the eye tracker and perform an eyeball and gaze calibration (see section *Calibration*). Then, they are asked to start the peg transfer task or to fix the gaze at a central point of the console, according to the group. This task continues during the whole study. Details are described below.

Calibration

We perform two kinds of calibrations. Before running the experiments, users perform an eyeball calibration[30] employing the eye-trackers' software (Pupil Capture 3.5.8). The process requires the users to fix the gaze in a point of the scene and move the head in circles while keeping the gaze fixed.

The second calibration aims to define a gaze model. In contrast with eyeball calibration, gaze calibration is repeated several times during the experiment. The calibration is based on the single point active alignment method (SPAAM), as shown in Figure 2.[31] In this method, a target is sequentially presented at nine different positions, and users are asked to fix the gaze at each position of the sequence.

N-Back test

The N-Back test consists of a random sequence of ten numbers read to the users. The task is to repeat the number in the sequence read *N* steps before the current digit. Table 1 introduces an example of the N-Back test for different values of *N*. The task requires the user to remember a portion of the sequence, while at the same time memorize new information. It has been showed this exercise triggers different levels of CL relative to the value of *N*.[32]

Data collection protocol

We employed two protocols to acquire data. The first protocol follows an initial eyeball calibration and introduction to the tasks. After this introduction, the user enters the telerobotic surgery console. After a dead-time period (no task is performed), we included two sequences of gaze calibration followed by an N-Back block. The N-Back block of the first sequence is composed of two dead-time separated by a 1-Back test. The second sequence's N-Back block contains up to three dead-time periods separated by a 1-, 2-, or 3-Back test. No additional tasks are performed, and users are suggested to fix their gaze at the center of the screen.

The second protocol follows a less constrained environment that combines N-Back tasks with continuous peg transfer. After the introduction and an initial eyeball and gaze calibrations, the users enter the dVRK console and start performing a simple peg transfer task (Figure 3). At minutes 2, 6, 10, and 14 of the experiment, peg transfer is interrupted and the SPAAM calibration is re-run to ensure an appropriate gaze calibration is available before every N-Back test. At minutes 8, 13, and 16 of the procedure, an auditory N-Back period is initiated. The peg transfer task continues even during the N-Back tasks. Each N-Back period is composed of two or three blocks of dead-time zone, followed by a randomly selected N-Back test.

Table 1: Examples of an N-Back recall task for *N* ∈ {0, 1, 2, 3}

Type of experiment	Sequence of numbers
Read	9 2 3 8 4 7 3 4 5 2 9 3 5 1 3
Recall	
0-Back	9 2 3 8 4 7 3 4 5 2 9 3 5 1 3
1-Back	- 9 2 3 8 4 7 3 4 5 2 9 3 5 1
2-Back	- - 9 2 3 8 4 7 3 4 5 2 9 3 5
3-Back	- - - 9 2 3 8 4 7 3 4 5 2 9 3

Participants must recall the number that was said *N* positions in the past. "-" indicates the user does not repeat a number in that moment of the experiment.

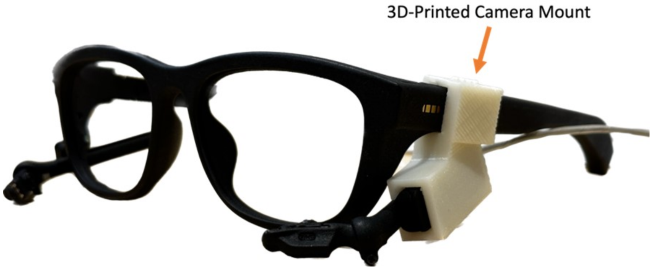

Figure 1. The Pupil Labs Invisible frame with personalized eye camera mounts. The camera mounts shape and orientation were designed considering the orientation of user's head and eyes in the dVRK console. dVRK, da Vinci Research Kit.

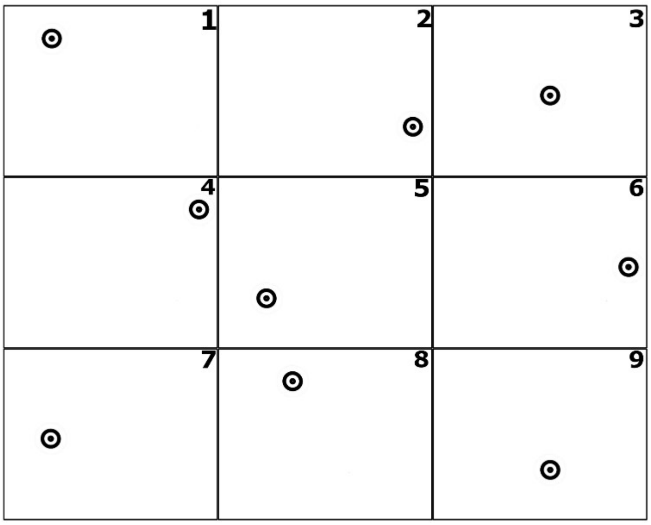

Figure 2. The nine positions of the target during the SPAAM gaze calibration. During the calibration, users observe each marker position for 30 seconds. Note that only one marker is displayed at the dVKR screen at a time. SPAAM, single point active alignment; dVKR, da Vinci Research Kit.

For both protocols, the ability of the methods to detect CL is evaluated for the N-Back periods. Pupillometry information, such as pupil diameter and gaze calibration are obtained using the Pupil Labs software, Pupil

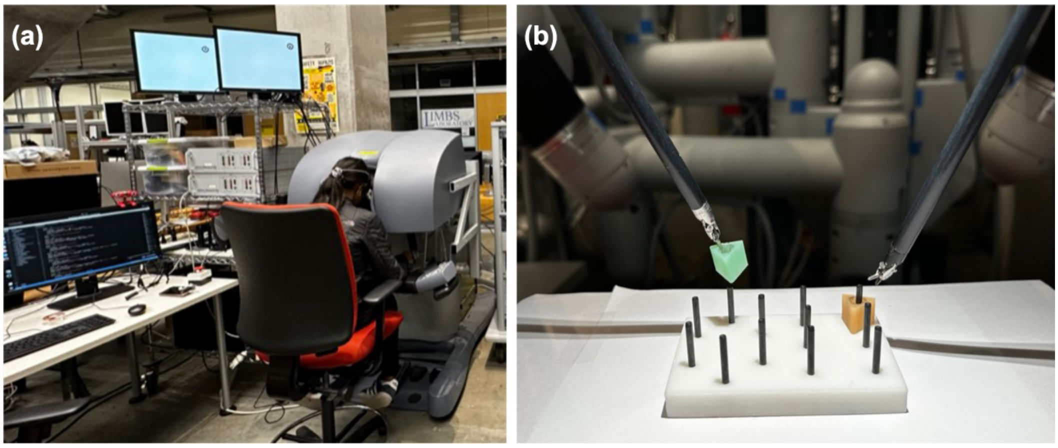

Figure 3. (a) The dVRK console with a user observing the gaze calibration routines (as displayed in the two upper monitors). (b) Detail of the peg transfer tasks, where users are asked to move the green and orange objects across the transfer plate. Once the peg transfer task starts, it continues during the entire experiment. dVRK, da Vinci Research Kit.

Capture 3.5.8 and Pupil Player version 3.5.7.

Data preprocessing

High confidence pupil diameter was obtained using Pupil Player version 3.5.7. Points detected with confidence lower than 0.9 were treated as outliers. Blinks were detected based on a slope detection approach, non-blink outliers were detected based on the standard deviation and mean pupil diameter of a 0.1 s window (12 data points). Both blinking regions and outliers were linearly interpolated between two high confidence zones. Finally, for CPD, BDPC and direct pupil diameter (PD) evaluation the signal was processed with a Butterworth filter with cutoff frequency of 12 Hz and an average filter over a 12 second window to reduce the noise of the pupil diameter signal.[29,33,34]

For the CPD algorithm, the first dead-time zone of each N-Back test was used as baseline. Remaining dead-time zones along each experiment were considered as regions without CL. For BCPD, the dead-time before the first N-Back experiment was used as the baseline measurement, and all the remaining of the test were considered as regions without CL.

RESULTS

We compare IPA,[24] LHIPA,[25] CPD,[9] and BCPD,[9] in addition to the direct PD as reference metric, as single diameter changes can be influenced by the CL level. Except for LHIPA, all metrics should increase together with the level of CL.

Table 2 and Table 3 show the average and standard deviation of the metrics under different N-Back levels, with and without peg transfer, respectively. We also evaluated the algorithms in neural zones (dead times), where no N-Back test was performed.

DISCUSSION

Pupil diameter metrics

N-Back only

From Table 2, we can observe that diameter-based metrics act close to their expected behavior for the N-Back-only protocol. Both metrics increment their output during the N-Back tasks. For the 1-Back and 2-Back experiments, these metrics reveal a consistent increment in their average corresponding to the N-Back level. This output is expected, as increasing the N-Back task will also increase the triggered CL. However, for the 3-Back test, we observed a reduction in the values of both metrics. This may derive from the level of difficulty of the N-Back task, which can cause an overload on the users, decreasing their CL response.[25]

Peg transfer with N-Back

Moving our attention to Table 3, the results for BCPD and CPD are less consistent with the theory. CPD values do not reveal enough information to suggest the presence of CL during the N-Back tests, and, in many cases, reveal smaller values compared with the neutral zone. BCPD average values are bigger for the N-Back tests than the neutral zone, showing a tendency to detect regions with CL. However, values between N-Back tests are not consistently increasing with difficulty. In general, it is possible that the addition of peg transfer adds additional challenges to the methods. Eye movement is unconstrained, and it is possible that as users engage in the task, their eye movements become more erratic and less predictable. Additionally, the light entering the eyes might have a focal component, and users may experience variations in focal brightness as they focus on different areas while grabbing and moving the peg, causing changes to the pupil diameter unrelated to the N-Back test. Unrestricted movements might cause

Table 2: Comparison of cognitive load algorithms on users performing N-Back experiments without the peg transfer task

Method	Neutral	1-Back	2-Back	3-Back
LHIPA (↓)	7.220 ± 2.200	6.846 ± 1.985	6.670 ± 1.709	7.257 ± 2.166
IPA (↑)	1.920 ± 0.495	1.942 ± 0.441	1.940 ± 0.460	1.902 ± 0.540
CPD (↑)	0.117 ± 0.570	0.283 ± 0.398	0.706 ± 0.380	0.456 ± 0.238
BCPD (↑)	-0.049 ± 0.323	0.144 ± 0.312	0.633 ± 0.463	0.395 ± 0.316
PD (mm) (↑)	3.756 ± 0.753	4.028 ± 0.698	4.650 ± 0.338	4.200 ± 0.776

LHIPA, Low/High Index of Pupillary Activity; IPA, index of pupillary activity; CPD, intra-trial change in pupil diameter; BCPD, inter-trial change in pupil diameter; PD, pupil diameter; ↓Value is expected to decrease with cognitive effort, ↑ Value is expected to increase with cognitive effort.

Table 3: Comparison of cognitive load algorithms on users performing N-Back experiments and peg transfer task

Method	Neutral	0-Back	1-Back	2-Back	3-Back
LHIPA (↓)	7.314 ± 2.177	7.092 ± 2.038	7.475 ± 2.268	7.694 ± 2.319	6.924 ± 1.965
IPA (↑)	1.817 ± 0.655	1.818 ± 0.677	1.617 ± 0.864	1.647 ± 0.842	1.832 ± 0.608
CPD (↑)	0.115 ± 0.259	0.147 ± 0.257	0.019 ± 0.544	0.096 ± 0.313	0.142 ± 0.309
BCPD (↑)	-0.086 ± 0.327	-0.024 ± 0.414	0.037 ± 0.514	0.017 ± 0.371	0.039 ± 0.322
PD (mm) (↑)	4.055 ± 0.999	4.489 ± 0.980	4.04 ± 1.173	4.077 ± 0.909	4.138 ± 1.048

LHIPA, Low/High Index of Pupillary Activity; IPA, index of pupillary activity; CPD, intra-trial change in pupil diameter; BCPD, inter-trial change in pupil diameter; PD, pupil diameter; ↓ Value is expected to decrease with cognitive effort; ↑ Value is expected to increase with cognitive effort.

additional noise to the eye tracker, and the focal brightness will affect the relative changes of the pupil since both BCPD and CPD do not account for changes in the incident light.

To evaluate the differences in eye movements between tasks with and without peg transfer, the Shannon gaze entropy is computed. The field of view is divided into nine areas of interest for analysis.[35] Figure 4b illustrates that the entropy is higher when performing the peg transfer task, indicating a more extensively eye movement across the field of view, resulting in a higher dispersion of gaze.[36] This dispersion correlates with the higher variation of the focal brightness, as show in Figure 4a highlighting the relevance of considering the focal component for pupil-based methods.

Frequency based methods

N-Back only

LHIPA presents overall lower values during the N-Back tests compared with the neutral task, suggesting the model can detect CL regions. However, we can also observe no pattern related to the levels of task difficulty. In contrast, IPA did not distinguish between neutral and CL steps, behaving inconsistently with their expected values. This behavior, however, aligns with Duchowski *et al.*, that mention IPA can fail when applied to N-Back tasks.[25]

Peg transfer with N-Back

Overall, the performance of the frequency-based methods when peg transfer is included (Table 3) does not present a clear pattern that could allow distinguish the presence of CL. The IPA presents inconsistent values, decreasing with higher levels of CL. LHIPA outputs also behave contrary to their expected value for the 1-Back and 2-Back tests. As discussed before, it is possible that the eye and head movements introduce additional noise to the signal negatively affecting the performance of these methods.

Final remarks

We evaluated the performance of four recent CL estimation methods on a telerobotic-surgical-like task with the dVRK. We induce CL by employing an auditive N-Back task, performing experiments when the N-Back task is combined with peg transfer. The task is unrestricted in the sense that the hand-eye coordination of the peg transfer task requires the user to focus on different regions of the screen.

Overall, while some metrics like LHIPA or BCPD allow identifying potential changes in CL, none of the metrics can be directly employed to assess the level of CL a user is experiencing. CL indices perform according to their underlying theory in some cases and fail in others, even if the induced CL is high. This behavior might result from not returning to the basal pupil diameter between N-Backs due to the multiple tasks being performed (*e.g.*, peg transfer, calibration, and N-Back). While frequency-based metrics are expected to be light insensitive, the results suggest that they fail to assess the presence of CL

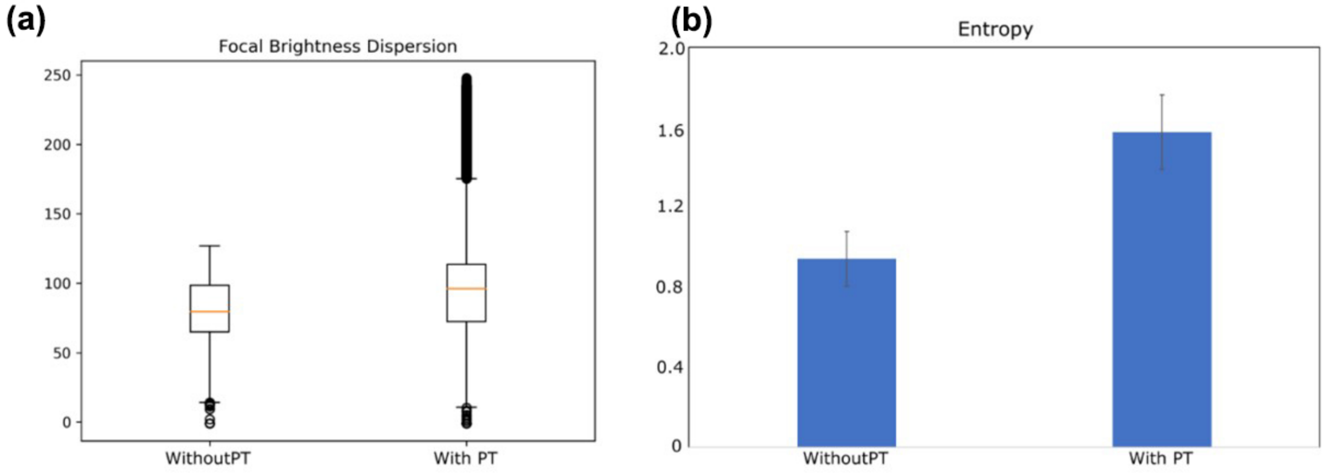

Figure 4. Lighting variations and entropy comparison between users performing N-Back task with and without peg transfer. (a) Dispersion of focal brightness without PT task and with PT task. (b) Gaze-entropy without peg transfer task and with peg transfer task. PT, peg transfer.

properly when peg transfer is combined with N-Back tests. It is possible that the movements that the users perform in the telerobotic console introduce additional noise to the pupil signal, which affects their performance, mainly for IPA, as previously reported.[25]

It is interesting to observe that the models based on the evaluation of pupil diameter, like BCPD, seem like a promising option to detect the potential presence of CL. However, their sensitivity to light and free movement of the eye, as mentioned by Krejtz *et al.*[9] is a factor to consider, especially in the surgical environment, where the specularities, the presence of tools, and smoke can lead to different changes in the focal brightness, leading to a larger range of pupil diameters. It is possible that including light models that consider the measured focal brightness based on gaze information could help minimize the effects of light.

As none of the evaluated methods success in giving information about the levels of CL, it is necessary to analyze the diverse factors that affect pupil diameter and the available complementary information (*e.g.*, brightness, gaze and blink) in the definitions of future models that allows predicting the presence and level of CL.

DECLARATION

Author Contributions

Puentes PR and Soberanis-Mukul RD: Designed the user study, analyzed the results, and wrote the paper. Acar A and Li Y: Designed the 9-point calibration routine. Soberanis-Mukul RD, Gupta I and Bhowmick J: Implemented and performed the user study. Ghazi A, Kazanzides P, Wu JY and Unberath M: Supervised the project and revise the manuscript. Kazanzides P, Wu JY and Unberath M: Obtained funding for the project.

Ethics Approval

The user study was approved under the HIRB00014648 at Johns Hopkins University.

Source of Funding

This research was funded by an Intuitive Surgical Technology Research Grant and Johns Hopkins University Internal funds.

Conflict of Interest

Mathias Unberath is an editorial board member of this journal. The article was subject to the journal's standard procedures, with peer review handled independent of the editor and the affiliated research groups.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

REFERENCES

1. Zhu H, Salcudean SE, Rohling RN. A novel gaze-supported multimodal human-computer interaction for ultrasound machines. *Int J Comput Assist Radiol Surg.* 2019;14(7):1107–1115.
2. Long Y, Wu JY, Lu B, *et al.* Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery. In: *2021 IEEE International Conference on Robotics and Automation (ICRA).* 2021: 13346–13353.
3. Wu JY, Tamhane A, Kazanzides P, Unberath M. Cross-modal self-supervised representation learning for gesture and skill recognition in robotic surgery. *Int J Comput Assist Radiol Surg.* 2021;16(5):779–787.
4. Bharathan R, Vali S, Setchell T, Miskry T, Darzi A, Aggarwal R. Psychomotor skills and cognitive load training on a virtual reality laparoscopic simulator for tubal surgery is effective. *Eur J Obstet Gynecol Reprod Biol.* 2013;169(2):347–352.

5. Chen IA, Ghazi A, Sridhar A, *et al.* Evolving robotic surgery training and improving patient safety, with the integration of novel technologies. *World J Urol.* 2021;39(8):2883–2893.
6. Zakeri Z, Mansfield N, Sunderland C, Omurtag A. Physiological correlates of cognitive load in laparoscopic surgery. *Sci Rep.* 2020;10(1):12927.
7. Joseph AW, Murugesh R. Potential Eye Tracking Metrics and Indicators to Measure Cognitive Load in Human-Computer Interaction Research. *J Sci Res.* 202064,168–175.
8. Skaramagkas V, Giannakakis G, Ktistakis E, *et al.* Review of Eye Tracking Metrics Involved in Emotional and Cognitive Processes. *IEEE Rev Biomed Eng.* 2023;16:260–277.
9. Krejtz K, Duchowski AT, Niedzielska A, Biele C, Krejtz I. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLoS One.* 2018;13(9):e0203629.
10. Atkinson RC, Shiffrin RM. Human Memory: A Proposed System and its Control Processes. *Psycho Learn Motiv.* 1968;2:189–195.
11. Sweller J. Cognitive Load During Problem Solving: Effects on Learning. *Cogn Sci.* 1988;12(2):257–285.
12. Sweller J. Cognitive load theory. In: Mestre JP, Ross BH, eds. *The psychology of learning and motivation: Cognition in education.* Elsevier Academic Press; 2011: 37–76.
13. Larsen HH, Scheel AN, Bogers T, Larsen B. Hands-free but not Eyes-free: A usability evaluation of SIRI while driving. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval.* 2020: 63–72.
14. Strayer DL, Cooper JM, Turrill J, Coleman JR, Hopman RJ. The smartphone and the driver's cognitive workload: A comparison of Apple, Google, and Microsoft's intelligent personal assistants. *Can J Exp Psychol.* 2017;71(2):93–110.
15. Tokuno J, Carver TE, Fried GM. Measurement and Management of Cognitive Load in Surgical Education: A Narrative Review. *J Surg Educ.* 2023;80(2):208–215.
16. Wilson RC, Shenhav A, Straccia M, Cohen JD. The Eighty Five Percent Rule for optimal learning. *Nat Commun.* 2019;10(1):4646.
17. Sridhar AN, Briggs TP, Kelly JD, Nathan S. Training in Robotic Surgery-an Overview. *Curr Urol Rep.* 2017;18(8):58.
18. Brook NR, Dell'Oglio P, Barod R, Collins J, Mottrie A. Comprehensive training in robotic surgery. *Curr Opin Urol.* 2019;29(1):1–9.
19. Chung YJ, Hsu CW, Chan MH, Cherng FY. Enhancing ESL Learners' Experience and Performance through Gradual Adjustment of Video Speed during Extensive Viewing. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* 2023: 1–7.
20. Krejtz K, Żurawska J, Duchowski AT, Wichary S. Pupillary and Microsaccadic Responses to Cognitive Effort and Emotional Arousal During Complex Decision Making. *J Eye Mov Res.* 2020;13(5):10.
21. Pfleging B, Fekety DK, Schmidt A, Kun AL. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 2016: 5776 –5788.
22. Wang CA, Munoz DP. A circuit for pupil orienting responses: implications for cognitive modulation of pupil size. *Curr Opin Neurobiol.* 2015;33:134–140.
23. Aminihajibashi S, Hagen T, Andreassen OA, Laeng B, Espeseth T. The effects of cognitive abilities and task demands on tonic and phasic pupil sizes. *Biol Psychol.* 2020;156:107945.
24. Duchowski AT, Krejtz K, Krejtz I, *et al.* The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.* 2018: 1–13.
25. Duchowski AT, Krejtz K, Gehrer NA, Bafna T, Bækgaard P. The low/high index of pupillary activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 2020: 1–12.
26. Abbad-Andaloussi A, Burattin A, Slaats T, Kindler E, Weber B. Complexity in declarative process models: Metrics and multi-modal assessment of cognitive load. *Expert Syst Appl.* 2023;233:120924.
27. Bacchin D, Gehrer NA, Krejtz K, Duchowski AT, Gamberini L. Gaze-based Metrics of Cognitive Load in a Conjunctive Visual Memory Task. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* 2023: 1–8.
28. Kazanzides P, Chen Z, Deguet A, Fischer GS, Taylor RH, DiMaio SP. An open-source research kit for the da Vinci® Surgical System. In: *2014 IEEE international conference on robotics and automation (ICRA).* 2014: 6434–6439.
29. Klingner J, Kumar R, Hanrahan P. Measuring the task-evoked pupillary response with a remote eye tracker. In: *Proceedings of the 2008 symposium on Eye tracking research & applications.* 2008: 69–72.
30. Kassner M, Patera W, Bulling A. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication.* 2014: 1151–1160.
31. Tuceryan M, Navab N. Single point active alignment method (SPAAM) for optical see-through HMD calibration for AR. In: *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000).* 2020: 149–158.
32. Mehler B, Reimer B, Dusek JA. MIT AgeLab delayed digit recall task (N-Back). Massachusetts Institute of Technology, Cambridge, MA, USA. 2011.
33. Memar Ardestani M, Yan H. Noise Reduction in Human Motion-Captured Signals for Computer Animation based on B-Spline Filtering. *Sensors (Basel).* 2022;22(12):4629.
34. Kret ME, Sjak-Shie EE. Preprocessing pupil size data: Guidelines and code. *Behav Res Methods.* 2019;51(3):1336–1342.
35. Lee Y, Jung KT, Lee HC. Use of gaze entropy to evaluate situation awareness in emergency accident situations of nuclear power plant. *Nucl Eng Technol.* 2022;54(4):1261–1270.
36. Di Stasi LL, Diaz-Piedra C, Rieiro H, *et al.* Gaze entropy reflects surgical task load. *Surg Endosc.* 2016;30(11):5034–5043.