PUPILLARY RESPONSE AS AN INDICATOR OF PROCESSING DEMANDS WITHIN A SUPERVISORY CONTROL SIMULATION ENVIRONMENT

Ciara Sibley
Joseph Coyne
Naval Research Laboratory
Washington, DC
Akshith Doddi
Georgia Institute of Technology
Atlanta, GA
Phillip Jasper
Clemson University
Clemson, SC

Current Unmanned Aerial Vehicle (UAV) operator task demands are highly variable and unbalanced across team members, resulting in sub-optimal operator utilization which leads to mishaps. This has driven the Department of Defense's desire for more flexible team structures and task allocation tools. Unobtrusive and continuous measures of operator state are needed to effectively allocate tasking to operators and prevent errors. Twenty participants completed two twenty minute supervisory control sessions where task load was manipulated by varying event frequency (e.g., information requests) and eye tracking data was collected. Pupillometry data revealed increased mean and maximum pupil sizes with increased task load and larger pupil size standard deviation in participants who performed poorly, compared to those who performed well. These results suggest that increased pupil size is indicative of increased processing demands and could be predictive of task performance within a complex environment where performance measures can be challenging to obtain.

An increasing number of military aviation missions are being performed by unmanned systems, reducing the risk to Warfighters while increasing mission capabilities. Ironically though, unmanned systems have high manpower costs associated with conducting operations, partially due to specialized operator roles and highly variable levels of tasking throughout missions. Numerous Department of Defense (DoD) roadmaps have been promoting reductions in UAV manning via increased automation and more effective tasking tools (DoD, 2013); in particular, the 2015 Navy S&T plan highlights the need for improvements in "task allocation/assignment, planning, and coordination and control for heterogeneous systems" (ONR, 2015).

Enhanced planning and task assignment tools necessitate knowledge of the unmanned system's state and availability as well as the operator's state and availability. Ensuring that an operator's workload is balanced is critical, given that extremes in workload have been demonstrated to cause reductions in performance and human error (Yerkes & Dodson, 1908; Kahneman, 1973). This paper presents how pupillometry data collected from remote eye tracking systems can help provide insight into an operator's mental state and possibly aid in future task allocation.

Research conducted over the last several decades has established that pupil size varies as a function of cognitive processing and that the magnitude of the dilation correlates with the amount of mental effort exerted (Kahneman, 1973; Beatty & Lucero-Wagoner, 2000; Andreassi, 2007). The vast majority of these studies, however have been conducted within highly controlled and simple environments where participants are asked to perform tasks such as digit sequence recall, mental arithmetic, or verbal processing (Klingner, Tversky, & Hanrahan, 2011; Johnson, Miller Singley, Peckham, Johnson, & Bunge, 2014). For the purposes of this initial study, the authors were interested in investigating whether pupillometry data collected in a realistic UAV supervisory control environment could serve as a continuous metric of user state and be predictive of task performance.

Method

Supervisory Control Operations User Testbed

Human subject data collection was conducted using the Supervisory Control Operations User Testbed (SCOUT) which enables the investigation of the impact of scenarios of varying levels of difficulty on UAV operator task performance. SCOUT was developed by the Naval Research Laboratory to represent the tasking that a future UAV supervisory controller will likely perform assuming advancements in automation. The tasking within SCOUT was developed through interaction with current UAV operators and involved an iterative design and feedback process. SCOUT contains a pre-mission planning phase as well as mission execution phase.

During planning, the operator must determine the best initial route for sending their three heterogeneous vehicles, with different speed and sensor range capabilities, to search for seven stationary targets with varying priority levels, deadlines, and location uncertainties. Once a vehicle arrives at a target search area, the search is automated such that a target is found once the vehicle's sensor comes within range of the target. After the planning phase, mission execution begins and the operator is responsible for responding to incoming information requests (via chat), performing flight parameter updates (i.e. changing altitude and speed), managing airspace (requesting access to restricted operating zones) and re-planning vehicle routes as either new targets or new intelligence on existing targets (i.e. more precise location information) becomes available. All events within SCOUT are pre-scripted to occur at specific times.

In order to promote participant motivation, SCOUT was designed to be game-like, where the user receives points for responding to chat messages and finding targets. Additionally, the user loses points if a vehicle enters a restricted operating zone without receiving authorization. SCOUT was designed to be used on two thirty inch monitors. Figure 1 shows SCOUT's left and right screens, where the left screen's primary functions involve communication, planning, and airspace management, and the right screen functions include monitoring and updating vehicle parameters. Lastly, SCOUT is integrated with the SmartEye Pro 6.1 eye tracking system, such that all simulation events, behavioral data and eye tracking data are synchronized and logged together.

Figure 1. SCOUT's left and right screens depicting three heterogeneous vehicles conducting a search mission.

Experimental Design

Twenty-three individuals voluntarily participated in this experiment. Three participants were excused from the study; two due to a lack of comprehension and one due to a simulation error. Analyses were conducted on the twenty remaining participants (7 women and 13 men) ranging in age from 18 to 48 years (M = 30, SD = 9.6).

Prior to data collection, participants received approximately thirty minutes of training, which included videos demonstrating how to perform tasks within SCOUT and interactive assessments to ensure comprehension. Following training, each participant engaged in two experimental sessions. The order in which participants conducted sessions was randomized to help prevent any order effects, where half the participants received Session A first, while the other half received Session B first. Table 1 illustrates the difficulty associated with each block, demonstrating that Block 1 was always of medium difficulty, followed by either an easy or hard block.

Each SCOUT session was pre-scripted and included four segments that always occurred in the following order: Planning, Block 1, Block 2, Block 3. Participants were given ten minutes to create an initial plan before the mission execution phase. Each mission execution block took approximately six minutes, for a total of approximately 18 minutes of mission execution per session. Each block had a different level of difficulty which was manipulated via the frequency of chat tasking and the frequency of new targets added (see Table 2). During the easy, medium and hard blocks, events were presented approximately every 75, 45 and 15 seconds, respectively. In particular, participants were tasked via chat message to update flight parameters, i.e. speed and altitude, on specific vehicles. This task required the operator to increase or decrease the current speed or altitude by a specific amount (e.g. “Decrease altitude of UH-28 by 117”). Performance on each of these requests was analyzed in terms of completion, reaction time and accuracy.

Table 1.
Sessions and difficulty blocks

	Difficulty Level		
	Block 1	**Block 2**	**Block 3**
Session A	Medium	Hard	Easy
Session B	Medium	Easy	Hard

Table 2.
Difficulty manipulations by block

Block Difficulty Level	Chat Task Frequency	New Targets Added
Easy	75 seconds	1
Medium	45 seconds	3
Hard	15 seconds	4

Results

Performance Results

The primary performance metric within difficulty blocks was success on flight parameter updates (i.e., altitude and speed updates). Success was measured by percent of tasks completed, reaction time, and percent error for the completed tasks within each block. A two-way (Session X Difficulty) repeated measures ANOVAs was conducted for each of the flight parameter update metrics (percent completed, reaction time, percent error). Table 3 shows the p-values associated with each two-way ANOVA test.

As shown in Table 3, there was a significant and large main effect of Difficulty $F(2,38) = 37.422$, $p = .000$, $\square_p^{\square} = 0.663$ for the percent of tasks completed. Additionally, there was a significant and medium main effect of Difficulty $F(2,38) = 4.095$, $p = 0.025$, $\square_p^{\square} = 0.177$ for the percent of error on completed tasks. See Figure 2 for a visual depiction of these effects.

No main effects were found for Session, however there was a significant and medium interaction effect between Difficulty and Session $F(2,38) = 3.288$, $p = .048$, $\square_p^{\square} = 0.148$ for the percent error, due to the fact that error slightly increased in the session two easy block and decreased in the hard and medium blocks. No differences were found in reaction times, indicating that participants responded at the same rate, both across sessions and difficulty blocks.

Table 3.
Significance values from two-way (Session X Difficulty) repeated measures ANOVA

	Performance Data ANOVA Significance Results		
	Percent Completed	**Reaction Time**	**Percent Error**
Difficulty	p = 0.000*	p = 0.860	p = 0.025*
Session	p = 0.335	p = 0.113	p = 0.883
Difficulty X Session	p = 0.074	p = 0.636	p = 0.048*
** Denotes significance at p < 0.05*			

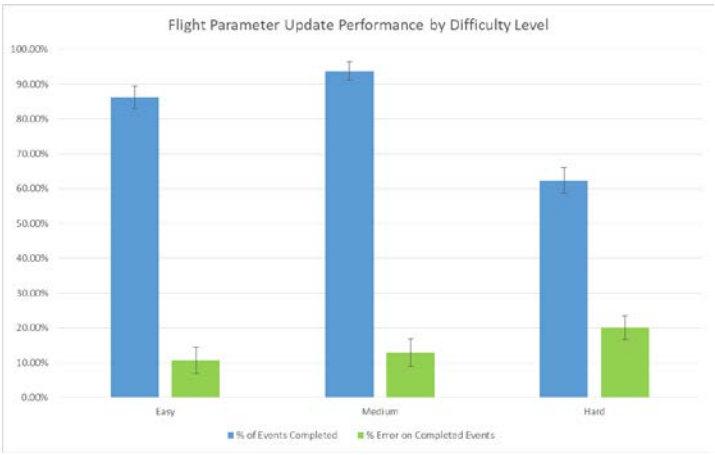

Figure 2. Performance on the Flight Parameter update task shown by block difficulty level

Pupillometry Results

Pre-processing was conducted prior to analyzing the pupillometry data. The first step involved filtering out all dropped and poor quality data, defined as having a value of zero, using SmartEye Pro's built-in quality measure. This initial step reduced the amount of data by 13%. Next, outliers were removed, which were defined as the top and bottom 1% of the entire twenty subject dataset. This ensured that erroneous data which were not caught by the quality measure (e.g. pupil measurements of 12 mm) were removed. Pupillometry data from both the left and right eye were consistent, but only data from the left eye are presented here.

Three pupillometry metrics were considered within this analysis: pupil size mean, pupil size standard deviation, and pupil size maximum. Each of these metrics was calculated for each individual, during each block, and within each session. A two-way (Session X Difficulty) repeated measures ANOVA was conducted on each of the three pupil size metrics (see Table 4). Analysis revealed a significant main effect of Difficulty for pupil mean $F(2,38) = 8.985$, $p = .001$, $\square_p^{\square} = 0.321$ and pupil maximum $F(2,38) = 3.577$, $p = .038$, $\square_p^{\square} = 0.158$.

Post hoc comparisons using a Least Significant Difference test indicated that the mean pupil size in the hard block (M = 3.60, SD = 0.55) was significantly greater than the mean pupil size in the easy block (M = 3.52, SD = 0.55). Additionally the maximum pupil size in the hard block (M = 5.84, SD = 0.25) was also greater than the maximum pupil size in the easy block (M = 5.73, SD = 0.41). Lastly, there was a significant main effect of Session $F(1,19) = 15.026$, $p = 0.001$, $\square_p^{\square} = 0.442$ for the pupil mean metric, where mean pupil size was significantly smaller in the second session than the first, which may reveal fatigue or learning. No interaction effects were found.

Table 4.
Significance values from two-way (Session X Difficulty) repeated measures ANOVA

	Pupillary Metrics ANOVA Significance Results		
	Pupil Mean	**Pupil Standard Deviation**	**Pupil Maximum**
Difficulty	p = 0.001*	p = 0.062	p = 0.038*
Session	p = 0.001*	p = 0.801	p = 0.586
Difficulty X Session	p = 0.301	p = 0.319	p = 0.301
** Denotes significance at p < 0.05*			

Pupillometry data were also collected immediately following a Situation Awareness (SA) probe that occurred within each block. During this time, known as the Regain SA phase, the screen and mission execution was paused so that the user could take as much time as needed to relax, regain situation awareness, and not have to attend or respond to any events before the block tasking resumed. Each of the three pupil metrics were computed during the easy, medium and hard block's Regain SA phases. A two-way (Session X Difficulty) ANOVA was conducted for each pupil metric (pupil mean, standard deviation and maximum) and revealed no significant main effects and no significant interaction effects; meaning no differences existed among the different Regain SA blocks.

Operator Performance and Pupillometry Data

Out of the twenty participants, the top four and bottom four performer's pupillometry data were analyzed to investigate whether those who performed well had different pupillary signatures than those who perform poorly. Performance was considered based upon a combination of overall mission score at the end of each session, and flight parameter update performance (i.e. high performance was defined as high percentage of events answered, low reaction times, and low error). No statistical analyses were conducted, given the low number of participants; however Figure 3 shows a comparison of pupil size standard deviations for the top and bottom performers in the easy, medium and hard blocks.

Visual inspection suggests that the bottom performers had higher variability in their pupil sizes, compared to the top performers. The minimum and maximum pupil sizes were assessed for each performer group to ensure there wasn't an inherent pupil size difference between the two groups. No difference existed: the maximum pupil size for the low performers was 5.966 and 5.961 for the high performers; the minimum pupil size was 1.997 for the low performers and 1.997 for the high performers. Mean pupil size and maximum pupil size were also compared between the two performance groups, but did not reveal anything conclusive.

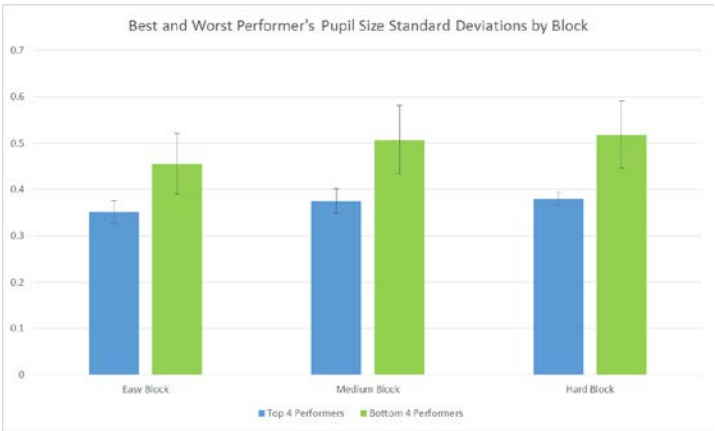

Figure 3. Pupil Size Standard Deviations for top and bottom four performers during easy, medium and hard difficulty blocks

Discussion

The results of this initial study are consistent with previously mentioned research conducted within simple task environments, and demonstrates that pupillometry data is also effective at discriminating between extremes in task load levels within a complex supervisory control environment. In particular, the mean and maximum pupil size metrics were significantly larger in high task load blocks of time compared to low task load blocks. Furthermore, the lack of difference among the metrics within the RegainSA phases, which took place in the middle of each difficulty block, provides additional evidence that pupillometry metrics are in fact indicative of a user's mental state (i.e. cognitive processing) and not some other factor such as fatigue.

While pupil size standard deviation was not statistically different among block levels at the aggregate level, it did appear to be trending towards significance at p=0.062. More interestingly, though, is the data in figure 3 which suggests that individuals who are struggling with tasking have greater variability in their pupil size. This finding requires further investigation but could be promising for potentially identifying operators who are overloaded and susceptible to making an error. None of the twenty participants were able to complete all tasking during either of the difficult blocks without making an error. Future analysis will involve a finer grained investigation into this performance data to assess whether it is possible to identify a signature in the pupillometry metrics indicating a user has become overloaded and is susceptible to making errors.

These results suggest that pupillometry data could be a useful metric for determining how to allocate tasking within an actual supervisory control environment, where operator state and task performance data is very rarely captured or available. Pupillometry data provides a continuous stream of information about an individual's mental state, which could be used to reveal when a user is overloaded. This information could be used to shed tasking either to another team member or automation before an error occurs. Future research will focus on building individual performance models and investigate the utility of embedding pupillometry data within a task allocation tool.

References

Andreassi, J.L. (2007). "Pupillary response and behavior," in *Psychophysiology: Human behavior and physiological response.* 5th ed (Mahwah, NJ: Lawrence Erlbaum Associates), 289-307.

Beatty, J., and Lucero-Wagoner, B. (2000). "The pupillary system," in Handbook of Psychophysiology, eds. J.T. Cacioppo, L.G. Tassinary & G.G. Berntson. 2 ed (Cambridge, UK: Cambridge University Press), 142-162.

Department of Defense (2013). "Unmanned Systems Integrated Roadmap FY2013-2038". (Washington, DC: Department of Defense). Retrieved from:http://www.defense.gov/pubs/DOD-USRM-2013.pdf

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Developmental Psychology, 5*, 218. http://doi.org/10.3389/fpsyg.2014.00218

Kahneman, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice-Hall.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, 48(3), 323-332.

Office of Naval Research. (2015). Naval S&T Strategic Plan. Arlington, VA. Retrieved from: http://www.onr.navy.mil/About-ONR/~/media/Files/About-ONR/2015-Naval-Strategy-final-web.ashx

Yerkes, R.M., and Dodson, J.D. (1908). The relation of strength of stimulus to rapidity of habit-formation. Journal of comparative neurology and psychology 18, 459-482.