Towards inferring cognitive state changes from pupil size variations in real world conditions

Naga Venkata Kartheek Medathati
Facebook Reality Labs
Redmond, Washington, USA
mnvhere@gmail.com

Ruta Desai
Facebook Reality Labs
Redmond, Washington, USA
rutadesai@fb.com

James Hillis
Facebook Reality Labs
Redmond, Washington, USA
jmchillis@fb.com

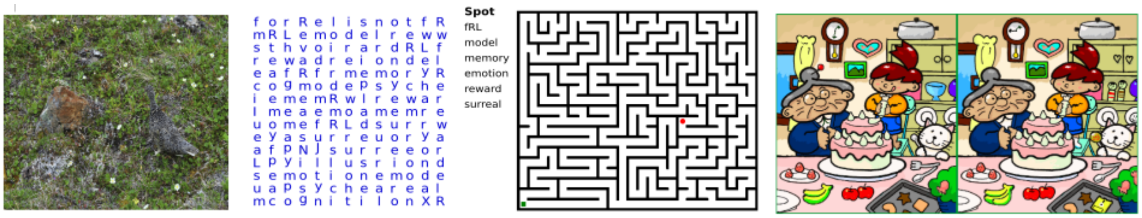

f o r R e l i s n o t f R
m R L e m o d e l r e w w
s t h v o i r a r d R L f
r e w a d r e i o n d e l
e a f R f r m e m o r y R
c o g m o d e P s y c h e
i e m e m R w l r e w a r
l m e a e m o a m e m r e
u o m e f R L d s u r r w
e y a s u r r e u o r y a
a f P N J s u r r e e o r
L P y i l l u s r i o n d
s e m o t i o n e m o d e
u a P s y c h e a r e a l
m c o g n i t i l o n X R
Spot
fRL
model
memory
emotion
reward
surreal

Figure 1: To enable the inference of cognitive state changes from pupillary variations in the real-world, we perform a series visual search experiments under normal indoor lighting conditions. Tasks associated with each experiment from left to right: Spot all the animals in the image where participants had no knowledge about the animals they were looking for (bird and its chicks) [Commons 2017], find words hidden in a matrix, path search to a destination through a maze, spot the differences between a pair of similar images where the number of differences were not known to the participants [Commons 2018].

ABSTRACT

The ability to infer cognitive state from pupil size provides an opportunity to reduce friction in human-computer interaction. For example, the computer could automatically turn off notifications when it detects, using pupil size, that the user is deeply focused on a task. However, our ability to do so has been limited. A principal reason for this is that pupil size varies with multiple factors (e.g., luminance and vergence), so isolating variations due to cognitive processes is challenging. In particular, rigorous benchmarks to detect cognitively-driven pupillary event from continuous stream of data in real-world settings have not been well-established. Motivated by these challenges, we first performed visual search experiments at room scale, with natural indoor conditions with real stimuli where the timing of the detection event was user-controlled. In spite of the natural experimental conditions, we found that the mean pupil dilation response to a cognitive state change (i.e., search target detected) was qualitatively similar and consistent with more controlled laboratory studies. Next, to address the challenge of detecting state changes from continuous data, we fit discriminant models using Support Vector Machine (SVM) computed on short epochs of 1-2 seconds extracted using rolling windows. We tested three different features (descriptive statistics, baseline corrected pupil size, and local Z-score) with our models. We obtained best performance using local Z-score as a feature (mean Area under the Curve (AUC) of 0.6). Our naturalistic experiments and modeling results provide a baseline for future research aimed at leveraging pupillometry for real-world applications.

CCS CONCEPTS

• **Computing methodologies** → *Cognitive science*; • **Human-centered computing** → Empirical studies in HCI.

KEYWORDS

pupillary response, visual search, user state estimation

ACM Reference Format:
Naga Venkata Kartheek Medathati, Ruta Desai, and James Hillis. 2020. Towards inferring cognitive state changes from pupil size variations in real world conditions. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3379155.3391319

1 INTRODUCTION

Variation in pupil size has been shown to be correlated with changes in luminance, vergence, and higher-level cognitive processes [Korn and Bach 2016; Mathôt 2018]. The discovery that higher-level cognitive processes affect pupil size [Hess and Polt 1960] leads to the promise of using pupil measurements to infer cognitive state changes. Recent advances in affordable eye-tracking systems [Coyne and Sibley 2016] also enhance the possibility of using this signal in real world applications. However, there are at least two significant challenges that need to be addressed before realizing this promise. First, as noted, there are at least three simultaneous processes underlying the one-dimensional noisy measurement (pupil size) we observe. To identify variation due to higher level cognitive processes, we must model the interactions between various causal

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391319

variables that affect the expression of pupil size (see [Binda and Murray 2015; Mathôt et al. 2014; Peysakhovich et al. 2017]). Second, the pupil variations associated with luminance and vergence are larger than their cognitive counterparts [Mathôt 2018]. Such low signal-to-noise ratio makes the robust separation of variations due to cognitive processes in uncontrolled conditions extremely difficult.

To enable the inference of cognitive state changes from pupillary variations in the real-world, we thus need to (1) determine the cognitive state change(s) of interest that could be inferred reliably from pupil variation in real-world scenarios (i.e., where luminance, viewing distance, and tasks are free to vary), and (2) develop the ability to use real-time changes in pupil size to detect such cognitive state changes.

Towards this goal, we make two contributions. First, we perform visual search experiments involving different types of search at room scale, under normal indoor lighting conditions where participants could freely make head and eye movements and where the event of interest (target discovery) was not under direct experimental control. Importantly, we used qualitatively different types of visual search tasks. These include searching for unspecified targets, searching for a known target object, identifying differences and also tracing a path. By including such a variety of visual search tasks, we provide a basis for generalizing models of the desired state change (successful target detection) across tasks having different levels of difficulty and context. Second, we pose the cognitive state change detection problem as an event detection problem from continuous time series going beyond characterizing mean event related pupil dilation responses. We study the performance of discriminant machine learning techniques such as Support Vector Machine (SVM) that use descriptive features of pupillary variations over a local time window to detect relevant cognitive state changes. Our detection approach is amenable to incorporation of pupillometry in HCI applications such as adaptive user interfaces [Lindlbauer et al. 2019]. The detection performance highlights challenges while providing a starting point for future research. To the best of our knowledge, our contributions are the first attempt at understanding and demonstrating the challenges of deploying pupillometry in the real world for detecting cognitive state changes.

Our visual search experiment design and SVM-based data analysis approach also complement extensive research on pupillary responses from vision and cognitive science communities. Specifically, we do not attempt to characterize the dynamics of the luminance, vergence and cognitive effects. The combinatorial complexity of obtaining data to allow for such a generative model would be hard to scale. Instead, we focus on a discriminant machine learning approach. We are motivated to detect the cognitive state change of interest (target recognition) amidst such confounding cues, which would always be present in naturalistic settings. Next, we briefly review existing experimental design and data analysis approaches.

2 RELATED WORK

2.1 Task and experiment design

Pupillary responses to target detection have been demonstrated in a range of tasks with different stimuli and control conditions, using both active and passive participatory responses.

2.2 Stimulus and control conditions

Pupil size has been shown to vary systematically with target detection in a range of tasks including 2-AFC threshold tasks, rapid serial visual presentation (RSVP) tasks and oddball tasks [Hakeram and Sutton 1966; Qian et al. 2009]. Some of the earliest studies demonstrated dilation events in pupil responses corresponding to threshold detection and discrimination events [Hakeram and Sutton 1966]. More recent studies have shown dilatory responses to RSVP stimuli [Qian et al. 2009]. The majority of studies to date use fixed luminance and viewing distance to control for variation in pupil size due to these variables [Beatty and Lucero-Wagoner 2000; Eckstein et al. 2017; Hess and Polt 1960; Laeng et al. 2012; Mathôt 2018; van der Wel and van Steenbergen 2018]. Researchers have also used varying levels of control on luminance and viewing distance to determine if variation in pupil size due to cognitive factors can be separated from the effect of these variables. For instance, Kahneman and Beatty demonstrated a robust signal for task difficulty independent of pupil variation due to viewing distance [Kahneman and Beatty 1966]. More recent research has studied pupillary responses to visual search in less restrictive settings using virtual reality (VR) [Jangraw et al. 2014]. The latter demonstrated that objects of interest could be identified through a combination of EEG and pupillary responses as participants navigated through a virtual environment.

2.3 Participant response

Pupil responses to target detection have also been studied with both active and passive participant response conditions [Privitera et al. 2010]. By active, we refer to the cases where the participants were asked to report an event/stimulus. By passive, we refer to the cases where they were asked to detect but not report the event as well as oddball tasks where an unexpected stimulus would be presented. Critically, studies conducted using traditional laboratory protocols such as 2-AFC or RSVP have control over the timing of the event to be detected so the event can be correlated with the pupil response.

Our experiment represents a scenario more like we would expect in real-world applications by not forcing the timing of the decision/detection event and allowing free viewing of a target under natural room lighting. In particular, the timing of the stimuli assumed to induce cognitive state changes were not under the direct control of the experimenter. The cognitive state changes corresponding to target detection were marked by a button press. The participants were asked to press a response button as soon as they believed to have found a target. In additions to target detection, the motor command needed to report the cognitive state change could lead to the pupil size variation. In experiments where participants reported perceptual changes for bi-stable stimuli, explicit motor response accounted for 70% of the pupil dilation, where a button press itself may include attention, decision, motor preparation and execution [Hupé et al. 2009]. Separating these components is an open challenge. It is likely, therefore, that within our protocol any pupil response we observe is due to some combination of motor command/preparation and the cognitive state change that leads to that motor plan along with influences from other factors such as luminance. However, as mentioned earlier, our goal is to be able to detect cognitive state changes in presence of such real-world noise.

2.4 Data analysis

Using pupil size data to infer cognitive state change is challenging in part because pupil size varies due to a variety of factors. To overcome the low signal-to-noise ratio, most of the previous research efforts averaged event-related pupil response across trials and participants. Before computing this mean event-related response, the data from different trials must be aligned to the event of interest. Several strategies have been used for this including: stimulus onset [Privitera et al. 2010], ocular events such as fixations [Jangraw et al. 2014] and user's response events such as a button press [de Gee et al. 2014]. Each of these strategies has yielded reliable event-related pupillary responses and allowed for the key insight that pupil size varies with higher level cognition.

While previous research using mean responses convincingly demonstrates the relationship between pupil variation and cognitive processes, single trial level event detection is an advancement towards continuous time analysis. Research that has demonstrated capability on this problem has used pupil data in combination with other biosensors such as EEG [Brouwer et al. 2017; Jangraw et al. 2014; Qian et al. 2009]. These studies have demonstrated above chance level performance, indicating that single-trial event detection from biosensory data is possible. Our aim is to determine if pupil variation can be used in isolation to detect cognitive events. Methods that rely on pupil data in isolation open the possibility of using mobile eye-trackers to track changes in cognitive state.

To detect events in a continuous fashion, we present extensive analysis using a variety of features describing epochs obtained using rolling window operators. The features we used exhibit different degrees of sensitivity to temporal variations [Brouwer et al. 2017; McCloy et al. 2016]. We envision that our results can provide a baseline for cognitive state change detection using pupillary responses within the context of visual search in naturalistic conditions.

3 VISUAL SEARCH EXPERIMENT

The experiment was designed to examine the pupil size variation in visual search tasks with less control than existing laboratory studies. We used different types of search tasks to study the questions on generalization, i.e., is there a consistent event-related response across different types of search tasks and across individuals? Tasks included detecting camouflaged animals, finding a path in a maze, and spotting the difference between two similar looking images. We used a wearable eye-tracking system to measure eye position and pupil diameter. Participants performed a series of visual search tasks on printed posters under natural lighting. They performed the study while standing, were allowed to move freely, and were also given autonomy to report successful completion of the task.

3.1 Participants

Twenty-five subjects (eleven females, one non-binary) ranging between 18 and 55 years of age participated in the experiment. Previous research has indicated that gender and age may have an effect on pupil size [Sanchis-Gimeno et al. 2012; Winn et al. 1994]. Such differences in individual pupil sizes were dealt with by subtracting the mean pupil diameter per epoch during feature computation for our SVM-based detection models (see Sec. 5 for feature computation details). All the participants had normal or corrected-to-normal vision and were screened based on our ability to calibrate them with the wearable eye tracking system. They did not have any prior training related to the task.

3.2 Materials

We used Tobii Pro Glasses 2 [1] to measure gaze and pupillometric data at 100Hz and a hand-held response button for participants to indicate target detection. The response button was designed to provide an event-related response with high temporal precision using the Transistor-transistor logic (TTL) input channel on the Tobii Pro Glasses 2. Default Tobii settings [2] were used during the experiment. Stimuli were printed images mounted on a wall under natural lighting. Each image represented a type of visual search task. Specifically, we considered four types of visual search tasks: (1) spotting objects of a semantic category with no object related information, i.e., camouflaged animal search (2) spotting words in a letter matrix (3) identifying a path in a maze and (4) spotting the differences between a pair of similar images. Examples of images used in the experiment for different visual search tasks can be seen in Figure 1.

3.3 Procedure

After explaining the study, we obtained informed consent as approved by the Western Institutional Review Board. The experimenter then ran the eye tracker calibration routine with the participant to ensure high quality tracking data. Participants were instructed to stand between 100 and 200cm from the stimuli and indicate when they found a target by pressing the response button. This distance was selected to ensure that the participants did not get too close to the stimulus while maintaining have sufficient visibility of the small targets. The range was determined by piloting few participants and observing their normal viewing behavior. The room was lit with normal overhead lighting and the illumination was kept consistent across all the participants. The experimenter was in the room with the participant to ensure an approximate viewing distance and to deliver instructions before each trial. At the beginning of each trial, the experimenter provided a brief oral description of the search task associated with the particular image. For example, when the search task was locating animals, participants were instructed to press the response button each time they found an animal. Participants were not told the number of targets except in the word search where the number of words they have to find are known *a priori*. They were neither asked to keep a count nor provided with any feedback on the number of targets associated with each task/image. They had the autonomy to decide when they felt they had detected all the targets. There was no time pressure and participants were instructed to inform the experimenter when they believed they had completed the task associated with each image. To simplify experimental design, participants did not receive any feedback on performance or false detection We thus focus on the perceived target detections by the participants even though

[1] https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/

[2] Tobii uses an image-based method to measure pupil sizes: https://www.tobiipro.com/learn-and-support/learn/eye-tracking-essentials/is-pupil-size-calculations-possible-with-tobii-eye-trackers/. As per Tobii recommendations and in accordance with previous scientific literature, our analysis relies on pupil size variations instead of absolute pupil sizes.

they may be false positives. [3] At the end of the session, participants were asked to provide difficulty ratings on a scale of 1 to 100 for each of the tasks using NASA Task Load Index (TLX) scheme [Hart and Staveland 1988]. Ratings were collected on a computer tablet.

Because the search tasks varied in terms of difficulty and stimuli, we presented the tasks in the same order for each participant. The participants were also doing free-form exploration of the stimuli in each task and were not constrained as in a RSVP or walk-through task. Consequently, the saccadic exploration of the individual stimuli varied substantially across participants.

4 DATA ANALYSIS

To demonstrate that we can replicate findings from well-controlled laboratory conditions, we first examined the mean event-related responses. We also report the individual and task-specific variations in mean event-related responses as these have implications for model generalization. We then formulate the problem of detecting user state change from pupillometry data as a binary classification problem and leverage machine learning techniques to perform the classification (Section 5). The state change of interest corresponds to the participant's belief in finding a target in each of the visual search tasks shown in in Figure 1. As noted before, we focus on the state change as reflected by their self report (button press). For the classification experiments, we focus on event-related pupillary responses.

4.1 Raw data and preprocessing

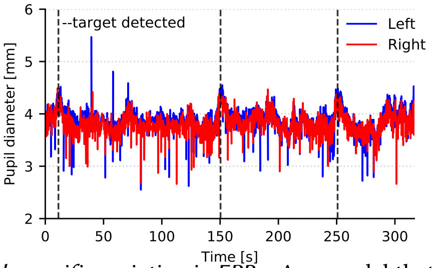
Prior to performing any analysis or classification experiments, we pre-process the data by removing outliers and filtering features that are unlikely to be related to the task. An example of raw pupil diameter as a function of time for one visual search task is shown in Figure 2 (top). The time at which the participants have indicated successful detection of a target by pressing a button are overlaid on the pupillary stream. Let these events be indicated by $\mathbb{1}_{detection}$. We use these events to align short windows of pupil data (also called as epochs) for computing mean event-related responses and as labels for our binary classification (Figure 2, top).

To pre-process the data, we follow the three step procedure prescribed in [Kret and Sjak-Shie 2018]. First, the data samples outside the median absolute deviation of pupil dilation speeds are identified as outliers and removed. Because blinks are characterized by large inter-sample changes in pupil size, removing samples based on deviation of pupil dilation speeds also enables blink filtering [Kret and Sjak-Shie 2018]. Approximately 35% of the samples were rejected per subject (mean = 0.35, std = 0.07). The gaps between samples are then linearly interpolated. Lastly, a low-pass Butterworth filter with a 45Hz cutoff is applied. Note that this is unlike several studies that use a 4Hzcutoff [Kret and Sjak-Shie 2018]. We considered a higher cut-off to examine if any information related to cognitive activity is captured by higher frequencies. This is inspired by other prior work demonstrating that high frequency pupil fluctuations may provide a more reliable and valid indicator of cognitive state changes [Duchowski et al. 2018; Peysakhovich et al. 2015; Villalobos-Castaldi et al. 2016; Wong 2019].

Figure 2 shows the pupillary data before and after each of these three steps. Figure 3 shows the mean event-related pupil response across all individuals. The traces represent the pupillary activity after mean dilation score from the baseline period, which is the first 200ms of each epoch, is subtracted. Blue curve represents pupillary response around the decision events and black curve corresponds to mean pupil response for randomly selected epochs, referred to as background response. Epochs corresponding to decision events were aligned by the button press response time. The observed EPR is consistent with many studies on visual target detection (e.g., [Privitera et al. 2010]). From the mean EPR, we observe that there is brief constriction followed by sustained dilation ($\approx 1.5s$) before returning to baseline. In the mean dilation response reported by [Privitera et al. 2010], the constriction event seems to onset around $500 - 700ms$ following the stimulus onset which were used to align epochs. In comparison, we observe an onset approximately $200 - 300ms$ before the button response. This constriction event could be indicative of target detection.

4.2 Event-related Pupillary Response (EPR)

4.2.1 Individual variation in EPRs. As we examined individual EPRs, we noticed potentially important differences in the patterns observed across participants. For example, as others have noted [de Gee et al. 2014], we noticed that some people showed constriction rather than dilation at the time of target detection. Figure 4 illustrates baseline corrected (obtained by subtracting mean of the pupil dilation from first 200ms of epoch) EPRs averaged across tasks for a few participants. To gain further insight into these differences, we computed pairwise correlation distance between mean responses of participants (Figure 5). As highlighted by this figure, the mean response for a few participants are observed to be negatively correlated with others. These patterns suggest that individuals may have highly idiosyncratic response patterns and accounting for these may be important for any application. In particular, this would indicate that developing models, which generalize and perform well across participants is challenging.



4.2.2 Task-specific variation in EPRs. Any model that uses pupil data to infer cognitive events (target detection in our case) must take into account differences in the task during which that event occurs. That is, the model must be able to generalize across contexts and tasks. We thus examined task-specific EPRs. Mean EPRs for each of our seven tasks are shown in Figure 6. Mean response differ in terms of the rate, magnitude and direction across tasks. The response curves belonging to Maze (blue line in bottom) and Spot the difference #4 (red line, second from the bottom) seem to standout

[3] We also note that there may be a lot of "almost false positive" pupil dilations where a participant thought they saw a target or were about to respond but did not. However, it is challenging to get ground-truth on such events. We therefore leave this for future explorations.

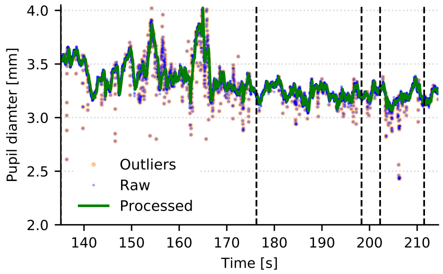


Figure 2: Top panel: Pupil size fluctuations during a visual search task are shown. Dotted lines indicate detection events. Bottom panel shows preprocessing steps of outlier removal and interpolation followed by low-pass filtering.

exhibiting higher amounts of constriction compared to the rest of the tasks. To determine if some of this variation could be accounted for by differences in task difficulty, we examined perceived difficulty of the task as measured by the NASA's Task Load Index (TLX). Figure 7 show results on each of the scales available in the NASA TLX. The Maze task (blue) was rated as lower in mental demand and effort than other other tasks and higher in Performance (corresponding to how satisfied people were with their performance). The higher levels of dilation we observe in tasks rated as more demanding and effortful is consistent with previous reports [Mathôt 2018]. However, the complex differences in the event related responses are likely due to task-specific demands that are not directly related to task difficulty. For example, the eye movement strategies that are effective in the maze task are different from effective strategies in the spot the difference task. Regardless of the cause, the inter-task variation we observe in addition to the inter-participant variation is indicative of the challenge we face in developing a model using pupil size to infer cognitive state changes that will generalize across tasks and individuals. We take first steps towards addressing this challenge by leveraging discriminant ML techniques in this work.

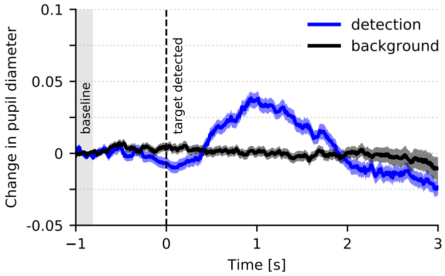


Figure 3: Event-related pupillary response over epochs aligned with button press. Zero corresponds to the time of button press. Black correspond to mean pupil response over epochs with no target detection. The shaded regions represent standard error computed by dividing standard deviation with the square root of number of events. We used all of the 798 detection event epochs in our dataset for computing the blue curve. There is significant variation across epochs.

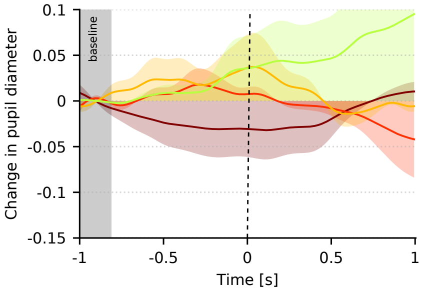


Figure 4: Mean pupil dilation response to target detection for 4 subjects across tasks. Shaded regions are standard errors.

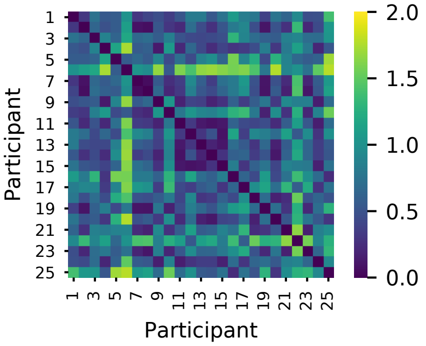


Figure 5: Inter-subject variability: Correlation distance between mean event-related responses of participants.

5 EVENT DETECTION USING BINARY CLASSIFICATION

We now consider the challenge of classifying observed changes in pupil size over short intervals (epochs of 1 or 2 seconds) as ones caused by a cognitive event. Each sample window of observation or epoch from the continuous pupillary stream could provide evidence for or against the hypothesis that the measured changes in pupil size is due to a cognitive state change. In our case, participant responses (button clicks) serve as labels for a cognitive state change corresponding to search target detection and are used for the development of classifiers using supervised machine learning methods. We thus examine how well discriminant machine learning methods perform on the task of classifying variation in pupil size within a short time window as one triggered by the subjective recognition that "the target has been detected" rather than other factors. We next describe our approach mathematically, before elaborating on the ML methods that we use.

5.1 Classification framework

Let $p(t)$ be pupil diameter of a subject as function of time t. An eye tracker operating at a sampling frequency f_s would provide us with a noisy sampling of $p(t)$, $p_{f_s}[n]$. To create epochs of $p_{f_s}[n]$, we define a local windowing operator $W_{N_1,N_2}[n]$ centered at time-stamp n.

$$p_{f_s}[n] = p(t) \times \sum_{-\infty}^{\infty} \delta(t - n(\frac{1}{f_s})) \tag{1}$$

$$W_{N1,N2}[n] = p_{fs}[n - N_1 : n + N_2] \tag{2}$$

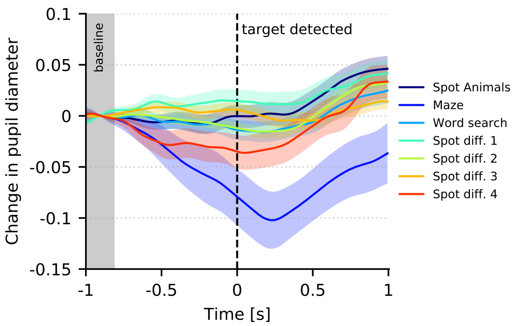


Figure 6: Mean pupil dilation response for each task across subjects. Spot the difference tasks involved finding differences between two similar images. Each search task had different levels of difficulty (see Figure 7).

We can use the epochs extracted by sliding windowing operators to train and test our classifiers. Traditionally, studies have explored windows of $1 - 3s$ duration centered on events of interest [de Gee et al. 2014; Jangraw et al. 2014; Privitera et al. 2010]. Inspired by this, we considered (1) a $2s$ symmetric window centered around participants' button press with $N_2 = N_1 = 100$, (2) a $1s$ window with $N_1 = 100, N_2 = 0$ and (3) a $1s$ window with $N_1 = 0, N_2 = 100$ for training the classifiers [4]. We consider performance of classifiers trained on epochs from each of these sliding windows separately [Peysakhovich et al. 2015] to determine whether the information in the pupil response to event detection occurs before or after explicit recognition of the event by participants. Results of these experiments are presented in second part of the results section. An instance of the symmetric windowing operator $W_{N_1,N_2}[n]$ with $(N_2 = N_1)$ is illustrated in Figure 8[A].

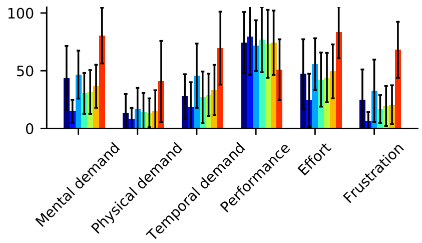


Figure 7: The perceived task load for each task along the six dimensions recommended by NASA TLX is shown. Error bars indicate the standard deviation across subjects and height indicates the mean score. Bar colors correspond to the different tasks within the experiment as in Figure 6.

Because we are considering possible application of the pupil data to real world scenarios, we wanted to test classification performance on all possible samples. That is, using the $W_{N_1,N_2}[n]$ windowing operator, we parse the time series into sequential overlapping epochs as shown in Figure 8[B]. Specifically, the center of the windowing operator n is shifted by 1 sample on the time axis ($.01s$ and the sample frequency is $100Hz$), This is the 'stride length' and can be considered as a hyperparameter during the classification experiments. We consider a stride length of 1 sample as this leads to the consideration of all possible windows.

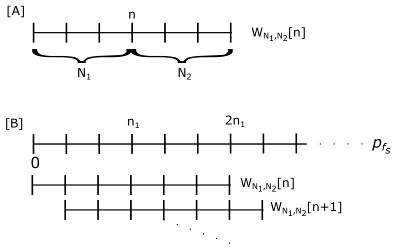


Figure 8: [A] A symmetric windowing operator of length $N_1(N_2 = N_1)$, centered at n is shown. [B] The epochs are extracted sequentially in a rolling window manner by sliding this operator over the sampled pupillary data p_{f_s}. We use a window of length $2s$ with the $100Hz$ sampling rate of the eye-tracker. Consequently, each epoch has 200 discrete samples.

We now have to make a decision about how to represent the data within each epoch to train and test a classifier. This is an important decision as raw representation of the data may contain irrelevant features and would lead to the need for more data to train effective classifiers while too simple of a summary statistic (e.g. an average) would lead to loss of information that make it impossible to classify events. We define a function $G(w_{N_1,N_2}[n])$ to compute representative features over the windowed sample,

$$G(W_{N1,N2}[n]) = x_n \tag{3}$$

Due to the lack of well-established descriptors for this task, we considered three types of features computed on epochs of data. Our first feature type is a coarse level statistical descriptor. Every epoch is represented using a 6D feature vector consisting of minimum, maximum, mean, variance, skewness and kurtosis computed over the epoch. We refer to this descriptive statistics feature set as *DS*. These statistics lose the temporal structure within in window but describe simple properties that may allow for effective learning from small data samples (i.e., "low-shot learning"). For our second feature set, our aim was to capture more of the temporal structure within a window. For this we use local Z-score for each epoch [McCloy et al. 2016]. Specifically, each data-point in the epoch is represented using the Z-score (LZ) computed using the observations within the epoch. We considered this feature as a good candidate to better capture the temporal dynamics of the pupillary variations while being more robust to differences across subjects and across time within subjects. We also tested baseline (BL) adjusted raw pupil dilation as a feature where the mean of the pupil dilation from first $200ms$ of the epoch is subtracted from the samples of the epoch [5].

Finally, we define a function f, as a binary classifier that maps the features to either 1 or 0, where 1 corresponds to the detection and classification of cognitive event from pupil data.

[4] We assume that a standard motor response delay of $200 - 400ms$ could occur when the button is pressed. This lies within the window sizes considered for the analysis.

[5] We empirically chose $200ms$ of the epoch to compute baseline pupillary activity. One could also treat this as a hyperparameter and tune for it using standard grid-search approaches for hyperparameter optimization, if needed.

$$f(x_n) = \begin{cases} 1, & \text{if } n \in \mathbb{1}_{detection} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

5.2 ML methods for classification

We experimented with a variety of standard classification algorithms for modeling the function f (eq. 4) including SVMs, Random Forests, and Nearest Neighbor classifiers [Bishop 2006; Cristianini and Shawe-Taylor 2000]. The classifiers obtained using SVM algorithm [Cristianini and Shawe-Taylor 2000] outperformed the other standard classifiers. We therefore present the results with SVM in the results section for the three feature sets (descriptive statistics DS, local Z-score LZ, and baseline adjusted data BL). The SVM classifiers used Radial Basis Function (RBF)-based kernel. The kernel type and its parameters were chosen through grid-search for hyperparameter optimization [Lameski et al. 2015]. The SVM classifiers with RBF kernels outperformed linear SVM classifiers highlighting the non-linearity present in our data. Note that the number of samples was on average ten times greater than the feature dimensions in our experiments, reducing the chance of over-fitting. To further prevent over-fitting, we used cross-validation.

Following conventional practice, the classification experiments consisted of training and test phases. In the training phase for a binary classifier, it is important to have an equal number of 'event' and 'non-event' trials. This is because many classification learning algorithms have low predictive accuracy for the infrequent class [Japkowicz and Stephen 2002]. For our epoched data samples, event-related epochs are rare in comparison to non-detection epochs. Each of our search task had on an average 7 search targets. Consequently, the total number of possible positive event samples over 7 tasks and 25 subjects would be equal to 1225. However, because of the task difficulty and errors, we only obtained 798 positive samples through participant button presses in our experiments. To compensate for this class imbalance, we performed random undersampling of the background epochs followed by Synthetic Minority Oversampling Technique (SMOTE) [Chawla et al. 2002]. Once these classifiers were trained, we tested performance on the full sample of epochs as defined by the rolling window described in Figure 8[B]. That is, we tested classification performance on all possible epochs in the full data stream. This procedure is intended to model how we would use a classifier for real-world applications.

5.3 Generalization across subjects and tasks

Inspired by the individual and task-specific variations we observed in our data (see Figures 4, 6), we analyze and test the ability of the SVM classifiers to generalize across subjects and tasks. The ability to generalize effectively across tasks and individuals is essential for using these models in any real world application.

We split the dataset into two groups with 10 participants for training and 15 for testing (i.e., about 40% of the data was used for training) [6]. During the training phase, we used leave-one-out cross validation to avoid over-fitting. Specifically, for across-subjects generalization experiments, we train the SVM on 9 participants and cross-validated on the tenth. Similarly, to study generalization across tasks, we grouped the epochs belonging to the tasks across the subjects and held out data corresponding to individual tasks, i.e., epochs belonging to 6 of the 7 tasks from all the 10 participants are used for training and epochs from left-over task from all the participants are used for cross-validation.

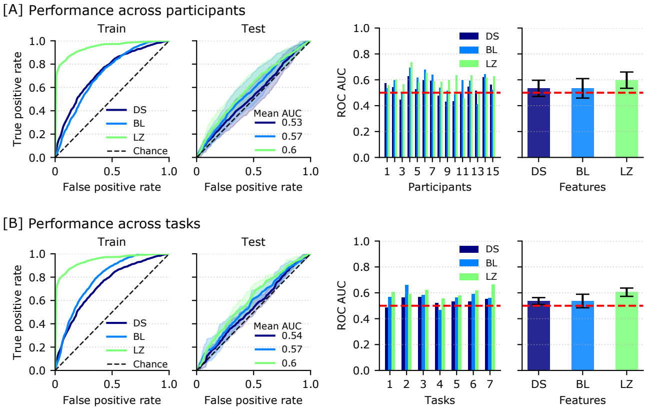


Figure 9: Detection performance of the SVM classifier using different features across participants and across tasks are shown in [A] and [B] respectively. The ROC curves are reported considering epochs of $2s$ duration ($N_1 = N_2 = 100$, where N_1, N_2 are defined as in eq. 2) for three different features – DS: Descriptive statistics, BL: Baseline corrected raw data, LZ: Local Z-score. The dotted red and black lines indicate chance level performance. The ROC during training correspond to the best performing classifier obtained with cross validation. In the test phase, the solid lines indicate the mean performance over participants and shaded regions indicate the corresponding standard deviation.

The resultant best performing classifier from the training phase was then tested on the epochs extracted using the sliding window. That is, we subjected the descriptive statistic DS, local Z-score LZ, and baseline adjusted data BL in all possible windows in the data stream to the binary classifier, f, that was learned using the SVM for the 15 test participants. As noted above, this is how we imagine such a classifier would be employed in any real-world application where we are trying to do continuous detection in a real-time application. Because we tested on all possible windows, we establish a lower bound in terms of expected performance during testing.

6 RESULTS

The training and testing performance in terms of the Receiver Operating Characteristic (ROC) curves for the three features, DS, BL and LZ, are shown in Figure 9. Classifier performance is shown across participants in the top row [A] and across tasks in the bottom panels [B]. ROCs for train and test phases are shown in the first two columns. The classifiers with all three feature sets: descriptive statistics (DS), baseline corrected raw data (BL) and local Z-score (LZ) performed above chance with an average AUC of 0.56 during testing. The ROC analysis further shows that the classifier trained with the local Z-score generalized better (an area under the ROC

[6] Note that an 80/20 split is more typical in the literature. However, using more data of the individuals during training in this manner resulted in poor test performance in our experiments (See Sec. 6.1)

curve (AUC) of .6[7]) than the ones trained with descriptive statistics or raw baseline features. The average AUC under ROC obtained using local Z-score is also better than chance as shown in the rightmost bar charts in Figure 9[A, B]. To give some idea of the differences in how well the classifier performs with different held out individuals or tasks, per subject and per task AUCs are presented in third column panels of Figure 9[A, B] respectively. The overall detection accuracy of the classifiers is poor when tested in a rolling window manner (Figure 9[A, B], column 2).

6.1 Implications

This poor performance of classifiers is likely due to multiple reasons. First, the cognitive event of target detection may be co-occurring with local light changes and motor responses owing to our real-world experimental conditions, making the classification challenging. Second, the poor performance could also be attributed to the inter-participant variability we observed in Figure 5. Such inter-participant variation in the pattern of pupil variation to event detection is consistent with previous reports [de Gee et al. 2014]. Approximately 20% of our participants exhibited negative correlation with others; that is, their pupils constricted where others' dilated. Such variability may prevent the classification models from generalizing, inspite of availability of more training data. When we repeated our experiments with an 80/20 split, we obtained a mean test AUC of 0.55 and 0.56 across participants and across tasks respectively using the LZ feature. Using more number of individuals during training thus did not lead to better test performance. This suggests that the use of pupil size variation to infer cognitive state changes may benefit from personalized models of pupil responses that are capable of handling individual variations. Test performance was somewhat better across tasks as compared to across individuals (Figure 9[A, B], column 2) suggesting there is more commonality across tasks than across individuals.

Overall, we found that the local Z score (LZ) features are less sensitive to the pupil size of the individual compared to the descriptive statistics (DS) or baseline corrected pupil response (BL) and performs better. This suggests the need for careful feature design for better performance. The results also indicate the possibility for better generalization across tasks compared to generalization across subjects and further imply the need for personalized models.

Our classification experiments and results should be considered as a baseline. We demonstrate what well-established classification methods could achieve with real-world pupil data in applied settings. We hope that our results would encourage ML researchers to think about methods applicable to challenging pupillometry data for future applications.

6.2 What is the best rolling window over pupil data for cognitive state change detection?

In addition to examining generalization performance of the classifiers across individuals and tasks, we have also examined three types of rolling windows. That is, in addition to the classifiers developed from the $2s$ window encompassing the button press ($W_{N_1=100,N_2=100}[n]$) with results shown in Figure 9, we trained the SVM with $1s$ windows prior to the button press ($W_{N_1=100,N_2=0}[n]$) and a $1s$ after the button press ($W_{N_1=100,N_2=0}[n]$). If the classifier trained using features in pre-button press windows outperforms or is on par with the others, it would indicate that pupil response reflects a change in cognitive state before or at the time the participant is making a motor plan. We followed the same cross-validation procedure described above during the training phase and tested the classifiers on all possible windows as in the previous analysis (Figure 9). However, as the local Z-score was the best performing feature, we trained these classifiers only using this feature for the experiment.

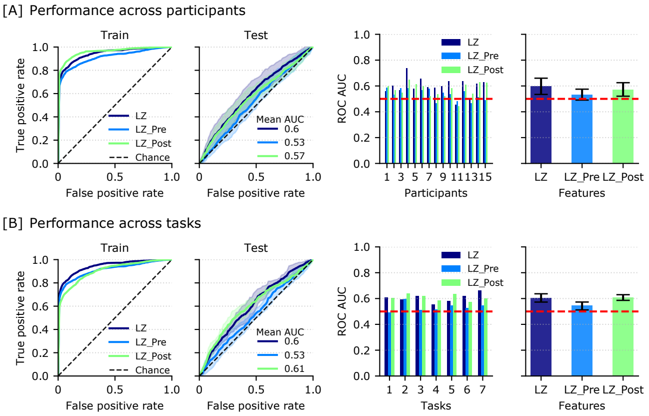


Figure 10: Detection performance of the SVM classifier using windows of observation right before and immediately after button press events across participants and across tasks are shown in [A] and [B] respectively. The ROC curves are reported for three windows, LZ: local Z-score, LZ_{Pre}: local Z-score considering only short window before button press, LZ_{Post}: local Z-score considering short window immediately after button press. The dotted red and black lines indicate chance level performance. The ROC during training correspond to the best performing classifier obtained with cross validation. In the test phase, the solid lines indicate the mean performance over participants and shaded regions indicate the corresponding standard deviation.

In a manner similar to Figure 9, classifier performance in terms of ROC and AUCs of models with different windows of observation are presented in Figure 10. Performance across participants in the top row [A] and across tasks in the bottom panels [B]. ROCs for train and test phases are shown in the first two columns. The three curves correspond to three different windows: local Z-score computing using pre-button press windows (LZ_{Pre}), local Z-score using post-button press windows (LZ_{Post}) and a combination of both (LZ_{Post}).

The results obtained from the test phase indicate that the features from post-button press windows lead to a better classification performance compared to pre-button press windows. This implies that there could be a stronger signature from the pupil after explicit response from a participant in the form of a button press. One would assume that an encompassing window of 2s would perform on par with either of the pre- or post-button press windows alone due to more samples. However, if either of the windows are

[7] AUC can be viewed as an estimate of the probability that the classifier ranks a randomly chosen positive example higher than a negative example and is equivalent to a Wilcoxon-Mann-Whitney statistic [Cortes and Mohri 2005]. For a binary classification problem, a chance model will thus have an AUC of 0.5.

not informative, there could even be a drop in the overall performance. In our case we find that pre- and post- windows provide complementary information. Consequently, 2s windows lead to better performance accuracy on average in both across subjects and across tasks experiments.

We note that, while previous studies were designed to disentangle possible contributions of latent cognitive constructs (perception, attention, prediction, decision, motor preparation etc) to pupil changes [Hupé et al. 2009], our study aimed to determine if well-established ML methods could be used to classify changes in pupil size to any aspect of cognitive state change in a more naturalistic task. Performance of the pre-button press classifier is likely attributable to both cognitive state change related to target detection and preparation of a motor command. The full 2 second window would also include sending of the motor command and proprioceptive feedback from the button press. We return to the challenge of attributing variation in pupil size to changes in specific hypothesized cognitive constructs in the discussion section.

7 DISCUSSION & CONCLUSION

Our results demonstrate that, (1) on average, we can measure a change in pupil size associated with a cognitive state change in an active task with less experimental control than many previous studies and (2) in this context, well-established ML classifiers show at best an AUC of .6 for classifying a single sample as caused by a cognitive event. We will discuss these result in relation to existing literature. We will then also briefly discuss the additional challenge of mapping pupil changes to hypothesized cognitive state changes (e.g., arousal, valence, attention, effort).

First, while the pattern of pupil size change, on average, replicate previous laboratory studies, the amplitude of the dilation response we observed were smaller than what has been observed in other studies. For example, Privitera et al. measured pupillary response to the target detection in active and passive tasks [Privitera et al. 2010]. In their case, the amplitude of the response was larger in the active task. The amplitude of the response we observed was similar to what they observed in their passive task. This could be due to our more naturalistic viewing conditions, greater variation in the timing of the response button press associated with the cognitive event or it could be that our task simply evoked a lower magnitude cognitive response. Lower amplitude response are a concern, of course, when it comes to the challenge of associating a pupillary change with a cognitive event, especially given other factors that cause variation in pupil size.

Next, we considered classification performance at the single trial level. As we noted in the introduction, previous studies demonstrating above chance classification performance for single trials, used pupil data in combination with other biosensors [Brouwer et al. 2017; Jangraw et al. 2014; Qian et al. 2009]. Our results thus provide a new baseline for continuous cognitive state change inference using only pupil size variation within the context of visual search in naturalistic conditions. Given that pupil size variations in such real-world conditions may be driven by motor response and local light changes along with the cognitive state changes, our results set realistic expectations of what it means to detect cognitive state changes for practical applications. We also tested two types of generalization of the classifier, inter-subject and inter-task. The generally poor performance may be a consequence of the large amount of inter-participant variability in the pupillary response (see Figure 5) as discussed in detail in the results section. In the future, we hope to test more advanced feature representations such as wavelets and models with better statistical treatment, inspired by recent progress in deep-learning.

Finally, to determine when is the best signal available from pupil data relative to the participants' button presses, we compared the classifier performance using signal features from pre-button press, post-button press, and pre- and post- encompassing windows. While each performed above chance, the classifiers obtained using the features from post-button press windows outperformed the ones that used pre-button press windows. This result could reflect a strong signature from the dilation response that followed a period of constriction. This dilation has also been shown to be effected by task difficulty, see Figure 6. This brings us to our final point of discussion, which is mapping between pupil size variation and hypothesized cognitive constructs.

While the literature has identified mapping between pupil variation and many hypothetical cognitive processes (arousal, attention, decision processes, cognitive load etc.), the experimental work often presumes validity of hypothetical cognitive processes. As the relationships between many of these hypothetical constructs are not well established, we focused on a simplified model of cognitive processes that includes task engagement, disengagement[8], and evaluation. That is, we model the human as an agent with a goal that requires time and computational resources to achieve and the ability to continuously evaluate outcomes of its actions. In practice, to move beyond such a simplified model to the one that can use pupil data to infer categorically distinct cognitive states will require rich contextual information as the one-dimension of data we have is highly ambiguous. Nonetheless, our results demonstrate that there is information in pupil variation and advanced ML methods hold the promise of using this signal to infer cognitive state variation in relatively naturalistic real world scenarios.

ACKNOWLEDGMENTS

We would like to thank Yuyang Zou for helping us with image stimuli generation. We would also like to thank Mingfei Yan for her support with the experiment design and Deborah Bohem-Davis for her inputs on the manuscript. We are currently unable to share the code and data from our experiments due to our institution policies. We are working on making this possible in the near future. Please email us if you are interested in working on our dataset.

REFERENCES

J Beatty and B Lucero-Wagoner. 2000. The pupillary system. In *Handbook of psychophysiology*, John T Cacioppo, Louis G Tassinary, and Gary G Berntson (Eds.). Cambridge University Press, Cambridge, UK, Chapter 6, 142–162.

Paola Binda and Scott O. Murray. 2015. Spatial attention increases the pupillary response to light changes. *Journal of Vision* 15, 2 (02 2015), 1–1. https://doi.org/10.1167/15.2.1

Christopher M Bishop. 2006. *Pattern recognition and machine learning.* springer.

[8] disengagement refers to a cognitive state transition that can occur through (1) recognition of goal achievement, (2) goal abandonment or (3) disruption (e.g., a time out or a an unpredicted event in the environment). In our visual search case, the person's goal is to find the target, each eye movement (fixation) is a sample allowing an update of progress toward the goal and target recognition or a timeout is the outcome/disruption.

Anne-Marie Brouwer, Maarten A. Hogervorst, Bob Oudejans, Anthony J. Ries, and Jonathan Touryan. 2017. EEG and Eye Tracking Signatures of Target Encoding during Structured Visual Search. *Frontiers in Human Neuroscience* 11 (2017), 264. https://doi.org/10.3389/fnhum.2017.00264

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357.

Wikimedia Commons. 2017. File:Ptarmigan and five chicks.JPG — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:Ptarmigan_and_five_chicks.JPG [Online; accessed 15-July-2019].

Wikimedia Commons. 2018. File:Spot the difference.png — Wikimedia Commons, the free media repository. https://commons.wikimedia.org/w/index.php?title=File:Spot_the_difference.png [Online; accessed 15-July-2019].

Corinna Cortes and Mehryar Mohri. 2005. Confidence intervals for the area under the ROC curve. In *Advances in neural information processing systems*. 305–312.

Joseph Coyne and Ciara Sibley. 2016. Investigating the Use of Two Low Cost Eye Tracking Systems for Detecting Pupillary Response to Changes in Mental Workload. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60, 1 (Sep 2016), 37–41.

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA.

Jan Willem de Gee, Tomas Knapen, and Tobias H. Donner. 2014. Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences* 111, 5 (2014), E618–E625. https://doi.org/10.1073/pnas.1317557111 arXiv:https://www.pnas.org/content/111/5/E618.full.pdf

Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. 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 (CHI '18)*. ACM, New York, NY, USA, Article 282, 282:1–282:13 pages.

Maria K. Eckstein, Belén Guerra-Carrillo, Alison T. Miller Singley, and Silvia A. Bunge. 2017. Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience* 25 (2017), 69 – 91. Sensitive periods across development.

Gad Hakeram and Samuel Sutton. 1966. Pupillary Response at Visual Threshold. *Nature* 212, 5061 (1966), 485–486. https://doi.org/10.1038/212485a0

Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139 – 183. https://doi.org/10.1016/S0166-4115(08)62386-9

Eckhard H. Hess and James M. Polt. 1960. Pupil Size as Related to Interest Value of Visual Stimuli. *Science* 132, 3423 (1960), 349–350.

Jean-Michel Hupé, Cédric Lamirel, and Jean Lorenceau. 2009. Pupil dynamics during bistable motion perception. *Journal of vision* 9, 7 (2009), 10–10.

David C Jangraw, Jun Wang, Brent J Lance, Shih-Fu Chang, and Paul Sajda. 2014. Neurally and ocularly informed graph-based models for searching 3D environments. *Journal of Neural Engineering* 11, 4 (jun 2014), 046003. https://doi.org/10.1088/1741-2560/11/4/046003

Nathalie Japkowicz and Shaju Stephen. 2002. The Class Imbalance Problem: A Systematic Study. *Intell. Data Anal.* 6, 5 (Oct. 2002), 429–449.

Daniel Kahneman and Jackson Beatty. 1966. Pupil Diameter and Load on Memory. *Science* 154, 3756 (1966), 1583–1585.

Christoph W. Korn and Dominik R. Bach. 2016. A solid frame for the window on cognition: Modeling event-related pupil responses. *Journal of Vision* 16, 3 (02 2016), 28–28. https://doi.org/10.1167/16.3.28 arXiv:https://jov.arvojournals.org/arvo/content_public/journal/jov/934914/i1534-7362-16-3-28.pdf

M.E. Kret and E.E. Sjak-Shie. 2018. Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods* (07 2018).

Bruno Laeng, Sylvain Sirois, and Gustaf Gredebäck. 2012. Pupillometry: A Window to the preconsicous. *Perspectives on Psychological Science* 7, 1 (2012), 18–27.

Petre Lameski, Eftim Zdravevski, Riste Mingov, and Andrea Kulakov. 2015. SVM Parameter Tuning with Grid Search and Its Impact on Reduction of Model Overfitting. In *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*, Yiyu Yao, Qinghua Hu, Hong Yu, and Jerzy W. Grzymala-Busse (Eds.). Springer International Publishing, Cham, 464–474.

David Lindlbauer, Anna Maria Feit, and Otmar Hilliges. 2019. Context-Aware Online Adaptation of Mixed Reality Interfaces. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 147–160.

Sebastiaan Mathôt. 2018. Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition* 1, 1 (feb 2018), 1289–23.

Sebastiaan Mathôt, Edwin Dalmaijer, Jonathan Grainger, and Stefan Van der Stigchel. 2014. The pupillary light response reflects exogenous attention and inhibition of return. *Journal of Vision* 14, 14 (12 2014), 7–7.

Daniel R. McCloy, Eric D. Larson, Bonnie Lau, and Adrian K. C. Lee. 2016. Temporal alignment of pupillary response with stimulus events via deconvolution. *The Journal of the Acoustical Society of America* 139, 3 (Mar 2016), EL57–EL62. https://doi.org/10.1121/1.4943787 PMC5392052[pmcid].

Vsevolod Peysakhovich, Mickaël Causse, Sébastien Scannella, and Frédéric Dehais. 2015. Frequency analysis of a task-evoked pupillary response: Luminance-independent measure of mental effort. *International Journal of Psychophysiology* 97, 1 (2015), 30 – 37.

Vsevolod Peysakhovich, Francois Vachon, and Frédéric Dehais. 2017. The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology* 112 (2017), 40 – 45.

Claudio M. Privitera, Laura W. Renninger, Thom Carney, Stanley Klein, and Mario Aguilar. 2010. Pupil dilation during visual target detection. *Journal of Vision* 10, 10 (08 2010), 3–3.

M. Qian, M. Aguilar, K. N. Zachery, C. Privitera, S. Klein, T. Carney, and L. W. Nolte. 2009. Decision-Level Fusion of EEG and Pupil Features for Single-Trial Visual Detection Analysis. *IEEE Transactions on Biomedical Engineering* 56, 7 (July 2009), 1929–1937. https://doi.org/10.1109/TBME.2009.2016670

Juan A Sanchis-Gimeno, Daniel Sanchez-Zuriaga, and Francisco Martinez-Soriano. 2012. White-to-white corneal diameter, pupil diameter, central corneal thickness and thinnest corneal thickness values of emmetropic subjects. *Surgical and radiologic anatomy* 34, 2 (2012), 167–170.

Pauline van der Wel and Henk van Steenbergen. 2018. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review* 25, 6 (Dec 2018), 2005–2015.

Fabiola M. Villalobos-Castaldi, José Ruiz-Pinales, Nicolás C. Kemper-Valverde, Mercedes Flores-Flores, Laura G. Ramírez-Sánchez, and Metztli G. Ortiz-Hernández. 2016. Spontaneous Pupillary Oscillation Signal Analysis Applying Hilbert Huang Transform. In *BIOSIGNALS*.

Barry Winn, David Whitaker, David B Elliott, and Nicholas J Phillips. 1994. Factors affecting light-adapted pupil size in normal human subjects. *Investigative ophthalmology & visual science* 35, 3 (1994), 1132–1137.

Hoe Kin Wong. 2019. *Instantaneous and Robust Pupil-Based Cognitive Load Measurement for Eyewear Computing*. PhD dissertation. UNSW Sydney.