Leveraging Pupil Dilation Measures for Understanding Users' Cognitive Load During Visualization Processing

Dereck Toker
Department of Computer Science
University of British Columbia, Canada
dtoker@cs.ubc.ca

Cristina Conati
Department of Computer Science
University of British Columbia, Canada
conati@cs.ubc.ca

ABSTRACT

In this paper we describe a preliminary investigation in using pupil dilation measurements to understand user visualization processing, with the long-term goal of building user-adaptive visualizations that can tailor the presentation of complex visual information to specific user needs and states. In particular, we look at how a selection of pupil dilation measurements are affected by adding several highlighting interventions designed to aid visualization processing to a bar graph.

CCS CONCEPTS

• **Human-centered computing** → *User Models; User centered design; Information visualization*;

KEYWORDS

Pupil Dilation; Eye Tracking;Adaptive Information Visualization

1 INTRODUCTION

The primary purpose of information visualizations is to assist users in exploring, managing, and understanding data. To date, most visualizations follow a one-size-fits all approach and do not take into account user differences. Several studies have shown, however, that individual differences such as perceptual speed and verbal/visual working memory can significantly impact performance and preferences during visualization processing [3, 4, 15, 17]. The long term goal of our research is to design user-adaptive visualizations that can support users based on their individual needs. As a first step toward this goal, Toker et al. [13, 16] analyzed users' gaze behaviour during visualization processing using eye tracking and identified several significant differences in attention patterns. In this paper, we extend this work by looking at pupil dilation measures. In particular, we analyze users' pupil dilation collected from a study involving visualization tasks with a bar graph and several alternative highlighting interventions designed to aid visualization processing. Results from this study pertaining to a variety of study factors on performance (completion time) already reported in [3]. Our aim is to combine these results with an analysis of pupil dilation to shed light on how the study factors (e.g., task type, interventions, and a variety of cognitive measures) impact visualization processing in terms of cognitive workload. Thus in this paper we present a preliminary analysis focusing on the effect of interventions on two measures of pupil dilation (mean and standard deviation of pupil size). We also outline our pupil dilation calibration methodology which examines alternative calibrating options when measuring baseline pupil size.

2 RELATED WORK

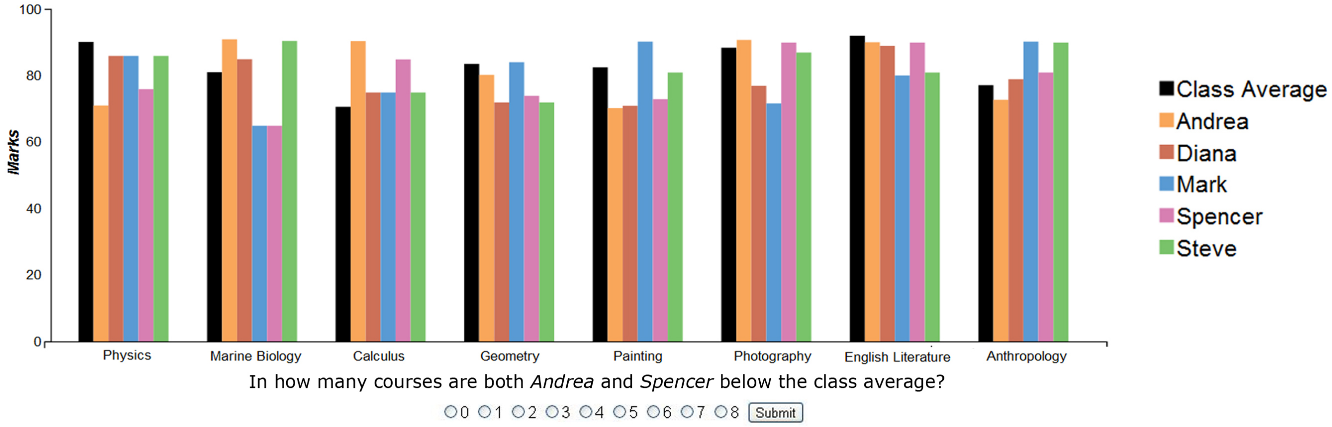
The use of eye tracking in information visualization systems has already been shown to be a strong candidate for predicting characteristics about the user in real-time. Both Steichen et al. [12] and Gingerich et al. [6] showed that a large set of aggregate eye-gaze features are a viable source to predict user differences (e.g., perceptual speed, visual working memory). In addition, Lallé et al. [9] and Toker et al. [14] have shown that including pupil dilation measures along with the set of aggregate eye-gaze features can lead to significantly better predictions of user differences (e.g., confusion, skill acquisition).

Eye tracking has also been used to identify and understand differences in terms of how the visualization is processed by the user. Toker et al. [13, 16] found several differences in visualization processing based on users' cognitive traits. For instance, users with low perceptual speed generated significantly more fixations and transitioned more often to the legend component of the visualization when compared to users with high perceptual speed. A similar result was found linking a user's visual working memory to the task answer input component of the visualization (e.g., radio buttons). These findings are important instances of how eye tracking can be leveraged for designing adaptive support since they identify specific elements of where users are having difficulty. Our aim is to extend this work with a similar analysis using pupil dilation data because it has been reliably shown that pupil dilation relates to changes in cognitive load [2, 7].

Other research has also investigated pupil dilation within the context of user-adaptive systems. For instance, Iqbal et al. [8] evaluated cognitive workload during route planning and document editing tasks in order to identify opportune moments for interrupting the user. Prendinger et al. [11] monitor pupil dilation in order to predict user preferences when confronted with a choice of visually presented objects. Martínez-Gómez & Aizawa [10] tracked pupil dilation measures in order to infer reading comprehension, which can be used to model individual users' topic familiarity. In our paper, we examine pupil dilation

UMAP'17 Adjunct, July 09-12, 2017, Bratislava, Slovakia

ACM ISBN 978-1-4503-5067-9/17/07…$15.00
http://dx.doi.org/10.1145/3099023.3099059



In how many courses are both *Andrea* and *Spencer* below the class average?

Figure 1: Sample bar graph visualization and task administered in the user study.

in the context of information visualizations, to inform the design of adaptive interventions based on differences in cognitive workload.

3 USER STUDY

The dataset used in this paper comes from a study that investigated the effectiveness of four highlighting interventions designed to help the processing of bar graphs, as well as how this effectiveness is impacted by both task complexity and different user traits. The long-term goal of this study was to understand if/which of these interventions would be suitable for providing adaptive support under specific circumstances, although in the study they were not presented adaptively. The study was a single session, within-subjects design, lasting at most 90 minutes. 62 participants performed 80 tasks using bar graphs (Figure 1) with a fully automated interface while their gaze was tracked via a Tobii T120 eye tracker. Users performed two different types of tasks (40 of each), chosen from a standard set of primitive data analysis tasks in Amar et al. [1]. The first task was Retrieve Value, one of the simplest task types in [1], which in the study consisted of retrieving the value for a specific individual in the dataset and comparing it against the group average (e.g., "Is Michael's grade in *Chemistry* above the class average?"). The second, more complex task type, was Compute Derived Value, which in the study required users to perform a set of comparisons, and then compute an aggregate of the comparison outcomes (e.g., "In how many cities is the movie *Shark Swamp* above the average revenue and the movie *Love Letter* below it?"). All tasks involved visualizations with the same number of data points (48) and same number of bar groups (8).

Each intervention evaluated in the study (shown in Figure 2) was designed to highlight graph bars that were relevant to answer the current question by guiding a user's focus to a specific subset of the visualized data while still retaining the overall context of the data as a whole [5]. The *Bolding* intervention draws a thickened box around the relevant bars; *De-Emphasis* fades all non-relevant bars; *Average Reference Lines* draws a horizontal line from the top of the left-most bar (representing the average) to the last relevant bar; *Connected Arrows* involves a series of connected arrows pointing downwards to the relevant bars. Each participant performed each of the two task types with each of the 4 interventions as well as *No Intervention* as a baseline for comparison, in a fully randomized manner.

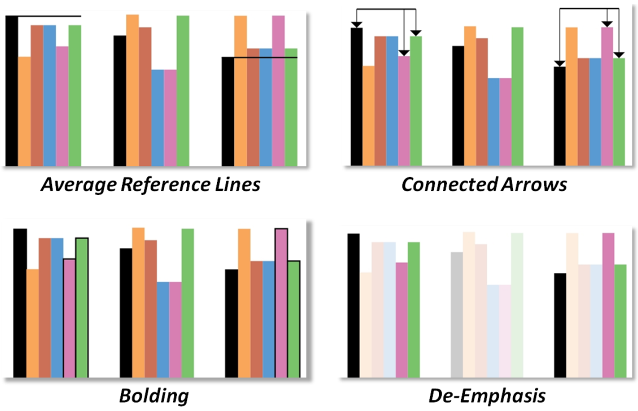

Average Reference Lines
Connected Arrows
Bolding
De-Emphasis

Figure 2: The four different highlighting interventions evaluated in the user study.

4 PROCESSING PUPIL DATA

As mentioned in the previous section, user gaze during the study was tracked using a Tobii T120 eye tracker. In addition to sampling information on gaze fixations and transitions, the eye tracker also records users' pupil diameter. In order to avoid possible confounds on pupil size due to lighting changes, the study was administered in a windowless room with uniform lighting. Because there are typically physiological differences in pupil size among individual users, it is also customary to collect a baseline pupil size from each user that can be used to later normalize the pupil measures.

In most work, the baseline is obtained by measuring a user's *rest pupil size*, obtained at the beginning of a study under relaxed conditions where there is little or low cognitive load. In our study, we considered two different ways to create these conditions. One, following a standard approach found in the literature, involves having participants stare at a blank screen for several seconds. However, we were concerned about potential issues with luminosity differences between a blank screen and what is shown on the screen during an actual task. Therefore we measured an alternative calibration baseline by displaying a mock bar graph visualization in order to produce similar lighting conditions to a real study task. We also removed the textual elements of the mock graph in order to minimize any added cognitive load. We distinguish between these two calibration measurements as: Blank/Graph[1]. Additionally, because the study was quite long and intensive (on average 90 min.), all participants were required to take a break halfway into the study. We took this opportunity to calibrate for pupil baselines twice in order to account for possible changes over the course of the study. Calibration measurements were therefore taken at the start of the session and again after the break, which are distinguished by: Start/Break.

In terms of calibration methodology, we are interested in knowing how similar/dissimilar the baseline pupil size measurements are in terms of luminosity differences between a blank screen versus a screen with a mock bar graph (Blank/Graph), as well as differences over time during the study (Start/Break). A Pearson correlation of the baseline pupil values for Blank vs. Graph produced an extremely strong correlation that was statistically significant (r = .921, n = 122, p < .001), indicating that these measures are almost identical. A Pearson correlation of the baseline pupil values for Start vs. Break also yielded a strong correlation that was statistically significant (r = .902, n = 122, p < .001), indicating that calibration across time intervals is also very consistent. In light of these findings, we selected the baseline pupil measurement obtained under the Blank/Start calibration condition for adjusting pupil measurements during the first half of the study, and the Blank/Break baseline for adjusting pupil values in the second half of the study.

5 RESULTS

Several measures related to pupil dilation have been used in the literature which include: mean pupil size, minimum pupil size, maximum pupil size, standard deviation of pupil size, as well as measures that track the speed and acceleration changes in pupil diameter (see [10] for a summary). For this paper's preliminary investigation, we focus on two of these pupil measures for analysis. First we select mean_pupilsize since it is a well-established measurement that appears in almost all work that investigates pupil size. Second we select a somewhat less common measure std.dev_pupilsize because previous work looking at gaze fixation related measures [13, 16] have found significant results relating to standard deviations which were computed based on gaze fixation angles. We then use the relevant baseline calibrations (see previous section) to normalize the pupil measures of each user by applying the percentage change in pupil size (**PCPS**) [8], which is defined as:

$$\frac{\text{measured\_pupilsize} - \text{baseline\_pupilsize}}{\text{baseline\_pupilsize}} \tag{1}$$

For both of the pupil measures mean_pupilsize and std.dev_pupilsize, we run a 5 (Intervention-Type) x 2 (Task-Type) ANOVA with Task-Order as a between subjects factor. Since we run two models, a Bonferroni correction of 2 is applied and p-values are reported post correction at the .05 level.

Effects of Intervention-Type

There was a main effect of Intervention-Type on both mean_pupilsize (p < .001, R^2 = .942) and std.dev_pupilsize (p < .001, R^2 = .022). Refer to Figure 3 and Figure 4 for the directionality of these findings.

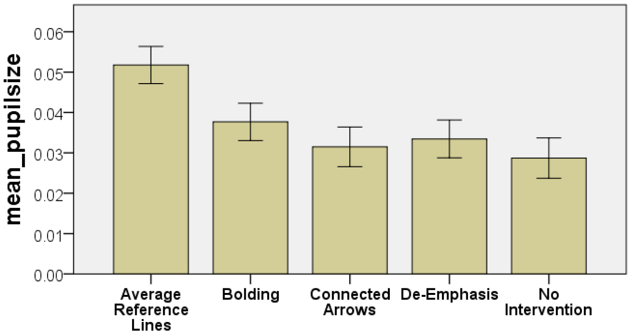


Figure 3: Main effect of Intervention-Type on users' mean pupil size.

Bonferroni-adjusted pairwise comparisons on mean_pupilsize (Figure 3) indicate that pupil size was significantly larger with the *Average Reference Lines* intervention than with all the other interventions. This is interesting because [3] reported, for the same study, that *Average Reference Lines* was the only intervention that did not significantly improve completion time when compared to tasks where *No Intervention* was provided. This suggests that the lack of performance improvement from *Average Reference Lines* could be explained in terms of increased cognitive load due to possible intrusiveness of this graphical object. It is also interesting to note that, whereas in [3] conditions with no interventions had similar performance as conditions with *Average Reference Lines*, *No intervention* has a significantly lower mean_pupilsize than *Average Reference Lines.* This suggests that slower completion time with no intervention is not the result of increased cognitive load, but rather it may be due to the lack of guidance provided by the more successful interventions.

As for std.dev_pupilsize (see Figure 4), pairwise comparisons indicate that std.dev_pupilsize during *Average Reference Lines* is significantly lower than with all other interventions except for *Bolding*. Given that *Average Reference Lines* also has the highest mean_pupilsize, the low std.dev_pupilsize tells us that users are likely maintaining a consistently high cognitive load throughout the whole task when they receive this intervention. In contrast, with other interventions there are only selected points with higher values of std.dev_pupilsize. Because in [3] these interventions were associated with improved performance, these

[1] Relative luminance of the *Graph* calibration screen was calculated to be 16% darker than the *Blank* calibration screen.

higher values may be associated with some notion of productive cognitive load (i.e., greater variability in pupil size is possibly an indicator of useful cognitive activity).

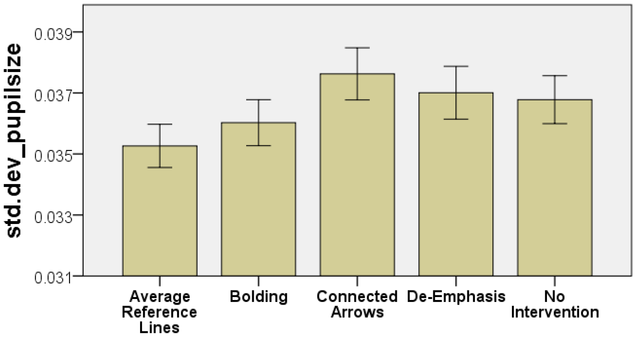


Figure 4: Main effect of Intervention-Type on standard deviation of users' pupil size.

6 CONCLUSIONS & FUTURE WORK

The long-term goal of our work is to build user-adaptive visualizations that can support the user based on their individual needs and states. In this paper, we examined how pupil dilation measurements can be leveraged to better understand information visualization processing.

We started by providing methodology towards controlling for possible confounds that can interfere with measuring rest pupil size in a user study, which is needed to correct for physiological differences between users. We evaluated our methodology by comparing two different calibration methods for obtaining a user's rest pupil size. First, we compared rest pupil sizes obtained on a blank screen versus a screen displaying a mock visualization since the screen brightness of our study tasks did not match the brightness of a blank calibration screen. We found a very strong significant correlation between both measurements, indicating that differences in rest pupil size between the two screens of differing brightness was consistent across users. Next, we compared rest pupil sizes obtained at the beginning of the study and during the middle of the study because the duration of the study was over an hour long. We also found a very strong significant correlation between both measurements, indicating that little difference exists between the two calibration times. Thus for our study, taking only one measurement of rest pupil size at the beginning of the study would have likely been adequate. Still, other researchers thinking of using pupil measurements in their studies ought to consider using the full set of calibration methods we presented here, in order to see if our findings will hold under other study conditions.

Next, we examined the effect that several highlighting interventions had on pupil dilation. In particular, we found that *Average Reference Lines* was the only intervention for which mean pupil size was significantly larger. *Average Reference Lines* was also the only intervention that did not improve user performance, suggesting that the lack of improvement in performance is due to the high cognitive load induced by this intervention. We offer two possible implications for user-adaptive visualizations based on this finding. First, monitoring pupil size could be beneficial towards designing, testing, or validating interventions, since instances of high cognitive load alone could be used to filter out unsuitable interventions (as opposed to relying on task performance). Second, pupil size could be leveraged as a real-time indicator of cognitive load to detect if users are having difficulty. Adaptations could then be triggered to support instances of high cognitive load during visualization processing. In fact, similar approaches have already been used in other areas of HCI, where cognitive load is tracked to determine suitable times to interrupt the user (e.g., [8]).

Lastly, more work will be needed to see if our findings will transfer to other visualizations, tasks, or interventions. Our hope is that members of the user modeling community interested in using pupil dilation methods in their research can help further corroborate our results.

REFERENCES

[1] Amar, R. et al. 2005. Low-Level Components of Analytic Activity in Information Visualization. *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization* (2005), 15–21.

[2] Beatty, J. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psych. Bulletin.* 91, 2 (1982), 276–292.

[3] Carenini, G. et al. 2014. Highlighting Interventions and User Differences: Informing Adaptive Information Visualization Support. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2014), 1835–1844.

[4] Conati, C. and Maclaren, H. 2008. Exploring the role of individual differences in information visualization. *Proceedings of the working conference on Advanced visual interfaces* (2008), 199–206.

[5] Few, S. 2009. *Now you see it: simple visualization techniques for quantitative analysis.* Analytics Press.

[6] Gingerich, M.J. and Conati, C. 2015. Constructing Models of User and Task Characteristics from Eye Gaze Data for User-Adaptive Information Highlighting. *Twenty-Ninth AAAI Conference on Artificial Intelligence* (Feb. 2015).

[7] Hess, E.H. and Polt, J.M. 1964. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science.* 143, 3611, 1190–1192.

[8] Iqbal, S.T. et al. 2005. Towards an index of opportunity: understanding changes in mental workload during task execution. (2005), 311.

[9] Lallé, S. et al. 2016. Predicting confusion in information visualization from eye tracking and interaction data. *Proceedings on the 25th International Joint Conference on Artificial Intelligence* (2016), 2529–2535.

[10] Martínez-Gómez, P. and Aizawa, A. 2014. Recognition of understanding level and language skill using measurements of reading behavior. (2014), 95–104.

[11] Prendinger, H. et al. 2009. Attentive interfaces for users with disabilities: eye gaze for intention and uncertainty estimation. *Universal Access in the Information Society.* 8, 4 (Nov. 2009), 339–354.

[12] Steichen, B. et al. 2013. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. *Proceedings of the 2013 international conference on Intelligent user interfaces* (New York, NY, USA, 2013), 317–328.

[13] Toker, D. et al. 2013. Individual user characteristics and information visualization: connecting the dots through eye tracking. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2013), 295–304.

[14] Toker, D. et al. 2017. Pupillometry and Head Distance to the Screen to Predict Skill Acquisition During Information Visualization Tasks. (2017), 221–231.

[15] Toker, D. et al. 2012. Towards adaptive information visualization: on the influence of user characteristics. *Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization* (Berlin, Heidelberg, 2012), 274–285.

[16] Toker, D. and Conati, C. 2014. Eye tracking to understand user differences in visualization processing with highlighting interventions. *Proceedings of the 22nd international conference on User Modeling, Adaptation, and Personalization* (Aalborg, Denmark, 2014).

[17] Velez, M.C. et al. 2005. Understanding visualization through spatial ability differences. *Proceedings of the IEEE Conference on Visualization* (Minneapolis, MN, USA, 2005), 511–518.