AFRL-AFOSR-UK-TR-2019-0017

The eye as a window to working memory: how to improve multitasking decisions using eye data

Niels Taatgen
RIJKSUNIVERSITEIT GRONINGEN

01/21/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOE
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 14-10-2018	**2. REPORT TYPE** Final performance report	**3. DATES COVERED** *(From - To)* 15 July 2017 - 14 July 2018

4. TITLE AND SUBTITLE

The eye is the window to working memory.
How to improve multitasking decisions using eye data

5a. CONTRACT NUMBER

FA9550-17-1-0309

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Taatgen, Niels A.
Borst, Jelmer P.
Shaposhnik, Hagit

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Rijksuniversiteit Groningen
Broerstraat 5
9712 CP Groningen Netherlands

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

USAF, AFRL DUNS 143574726
AF OFFICE OF SCIENTIFIC RESEARCH
875 NORTH RANDOLPH STREET, RM 3112 ARLINGTON VA 22203-1954

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

No restrictions

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Interruption management systems have as a goal to defer interruptions to a moment of low workload for the user, minimizing the costs of interruptions, and reducing risks. In this research we have explored whether eye-data can serve as a basis for such a decision. Exploring multiple datasets and paradigms we found that such a system is feasible, assuming alterations in workload are not too rapid. The most successful classifier is based on the boosted tree algorithm, leading to a classification accuracy of 74%.

15. SUBJECT TERMS

Human performance, Human multitasking, Eye data, Machine Learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Niels Taatgen
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			
U	U	U		11	**19b. TELEPHONE NUMBER** *(Include area code)* +31 50 363 6435

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATE COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33315-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report. e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

The Eye is the Window to Working Memory

How to improve multitasking decisions using eye data

FA9550-17-1-0309

PI: Niels Taatgen

co-PI: Jelmer Borst

Period of performance: 15 July 2017 - 14 July 2018

Table of Contents

Summary

Interruption management systems have as a goal to defer interruptions to a moment of low workload for the user, minimizing the costs of interruptions, and reducing risks. In this research we have explored whether eye-data can serve as a basis for such a decision. Exploring multiple datasets and paradigms we found that such a system is feasible, assuming alterations in workload are not too rapid. The most successful classifier is based on the boosted tree algorithm, leading to a classification accuracy of 74%. In order to further increase accuracy, we need to personalize the classifier by adding online training.

Introduction

The goal of this project was to investigate which methods and data sources are the most promising to build an *interruption scheduling system*. An interruption scheduling system intends to help users to defer interruptions if they arrive at a moment where an interruption is undesirable. This may be because it is too dangerous to be interrupted (in driving, piloting and air traffic control), or because an interruption leads to serious interruption costs (in terms of time) in the main task. A desirable property of the system is that it is independent from the task. With this we mean that the algorithm does not require any knowledge of the task that is performed. This means that it does not need to be tailored to the particulars of the application that it is used with, and that the application itself does not need to be modified in order to be used in conjunction with the interruption system.

What is a good moment to switch between tasks? According to the literature, the best moment for an interruption is a moment of low workload (Borst, Taatgen & van Rijn, 2015; Iqbal & Bailey, 2005; Katidioti & Taatgen, 2014; Monk, Boehm-Davis & Trafton, 2004). The problem with this notion is that workload is typically considered as a variable that gradually increases and decreases in time, but is otherwise not well defined. In several experiments in which subjects had to alternate between multiple tasks, we found that the key disrupting factor is the extent to which both tasks need to store temporary information, in other words, whether they use working memory. If no more than one task needs working memory, alternation between tasks is often without extra costs, even if the tasks involved are complex. However, if two tasks both need working memory, alternating tasks is costly and prone to errors, even if the tasks involved are simple (Borst, Taatgen & van Rijn, 2010; Nijboer, Borst, van Rijn & Taatgen, 2016). This means we cannot just take the perceived workload of one task and add it to another to get a combined workload measure. Instead, the overlap in working memory requirements determines multitasking disruption. Therefore, in order to determine moments at which a person can be interrupted without incurring unreasonable costs, we have to detect moments where working memory is not engaged.

To detect when working memory engagement is low – and thus when an opportune moment for an interruption occurs – we have found that an excellent source of information is the eye, in particular the size of the pupil (Katidioti, Borst, Bierens de Haan, Pepping, van Vugt & Taatgen, 2016). The pupil size varies with working memory load: when a task requires no working memory, the pupil is generally smaller or decreased in size compared to when the task requires working memory. Ideally, the system should only interrupt the user if there is no load on working memory.

The challenge of building such a system it has to monitor the eye online, and therefore has to deal with a signal that is quite a bit more noisy than what the data look like when they are averaged. In addition, individuals differ in their pupil response.

Methods and Results

To achieve the goals outlined in the introduction, we have investigated two datasets in which the pupil size varied with working memory load. Each of the datasets was entered into several machine learning algorithms in order to investigate how well the resulting classifier was able to discriminate a low-workload from a high-workload phase in the task.

Simulated Air Traffic Control

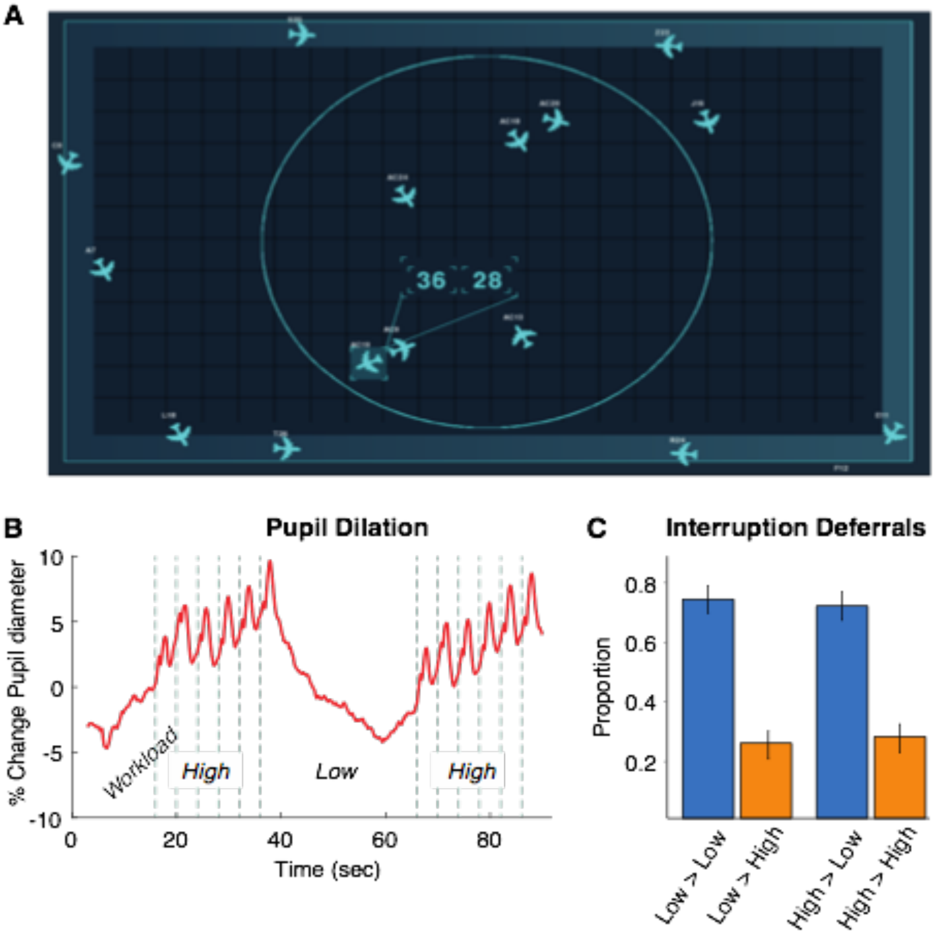

Figure 1. (A) Screenshot of the Air Traffic Control Task. Airplanes would post requests for altitute changes offering two choices (36 and 28 in the example). The subjects had to pick the altitude that had not been used before in a previous request. (B) Pupil dilation during task performance. Each vertical dotted line is an altitude request. The goal of the systems was to differentiate between the periods with and without requests (High-Low). (C) Performance of the final classifier: blue bars are successes, orange bars failures of the system.

The first dataset concerned a simulated Air Traffic Control task. In the experiment, subjects had to respond to requests made by airplanes. Subjects alternated between a high-load phase, in which they had to make decisions every four seconds, and a low-load phase, where airplanes would continue to move on the screen but in which no requests were made. Figure 1A shows a screenshot of the task. Figure 1B shows to average pupil size over the course of a single block, in which low- and high-workload phases were alternated. It is clear that the pupil size was larger on average in the high-workload phases, and increased within that phase as the working memory load increases with each decision.

The next step was to determine the best type of classifier and the best set of features to feed into that classifier. Eventually we chose 11 features, five based on the pupil size (e.g., median over the last four seconds, maximum), and six based on eye-movements (e.g., number of saccades, distance travelled, area covered). These features were fed into a logistic regression classifier, a random forest classifier (with 200 or 300 trees), and a gradient boosted tree classifier (also with 200 or 300 trees). We used leave-one-out cross validation to calculate the performance of the classifier, which means that we train the classifier on all but one subject, and then predict the load for the remaining subject (and so this for all subjects). This is the most conservative measure, because it means we do not know anything about the subject we are testing on.

The logistic regression classifier performed worst, with only 63.5% correct classifications. Next was the random forest, with a performance of 70.4%. The gradient boosted tree gave the best performance: 74.3% (the number of trees did not matter).

If we would have used this classifier in an interruption management system, it would have made decisions about when to allow an interruption, and when to defer it (and to what moment it would be deferred). We could still determine this based on the data. The results of this are shown in Figure 1C. It shows that 78% of the cases where the experiment is in a low workload phase, the system would have correctly allowed an interruption. However, in the remaining 22% it would have postponed the interruption to a high-load moment, which is, of course, a bad decision. On the other hand, if the experiment is in a high-workload phase, the system defers them to a low-load phase in 77% of the cases. In the remaining 23% it still allows the interruption.

This results shows that the system is far from perfect, but do lead to better interruptions. One thing we cannot be completely sure of is whether a high-workload periods is actually experienced as such by the subject. This means that the classifier may still be correct in allowing an interruption.

Email task

The second dataset was from an experiment in which subjects had to answer emails. To answer an email, they had to look up information in a simulated web browser. While they are searching, they have to maintain information in their working memory until the search is done, after which there is a brief moment without memory load. They then had to type the answer, during which they need to remember the information to put into the email. Figure 1 illustrates the pupil dilation in this task. As already indicated, there were clear high and low working memory load moments in this task (red and green labels), but depending on the previous state of the pupil it was not always easy to identify what the current load is (i.e. the 'Mail select' moment following 'Mail reply').

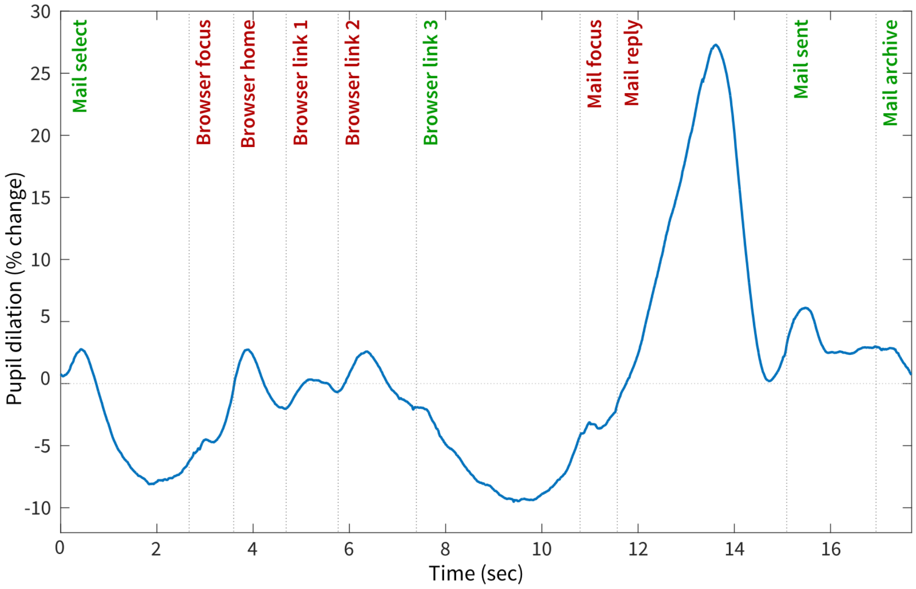

Figure 2. Pupil size during the email task (Katidioti, Borst, Bierens de Haan, Pepping, van Vugt & Taatgen, 2016).

To determine the best classifier, we used the same approach as in the simulated air traffic control, except that we trained a Support Vector Machine instead of a random forest classifier. Unfortunately, the classification was not very accurate, with the Support Vector Machine and logistic regression tying for the best accuracy at 59%.

The problem in this task is that high- and low workload periods alternate relatively quickly. This problematic because a change in workload is only visible in the pupil after approximately one second.

Discussion and Conclusions

The results show that it is possible to improve the timing of interruptions on the basis of the pupil size. A constraint is that low- and high workload periods should not alternate too quickly. In many practical situations this will fortunately be the case. We also see that the system is yet far from perfect. We may expect some improvement with further calibrations of the technique, but for a significant improvement we will need to personalize the classifier. This requires a system that uses a general classifier as a baseline, but calibrates this on the basis of user feedback (e.g., if the user manually defers an interruption, this is a learning signal that the interruption decision was wrong).

Another factor to consider is that we may have underestimated the accuracy of the classifier because we assume on the basis of the task characteristics that workload is either high or low. This may not be accurate for all circumstances and all subjects. Therefore a period that we think has a high workload may really be low workload for some subjects, meaning that an interruption may be valid at that time.

To conclude, managing interruptions on the basis of eye-data is a promising technique that will require further investigation to develop into a system that can be used in practical applications.

Resulting publications

Shaposhnik, H., Borst, J.P., & Taatgen, N.A. (2018). Predicting the optimal time for interruption using pupillary data and classification. In: Proceedings of the annual conference of the cognitive science society 2018 (pp. 2479-2484). Austin, TX: Cognitive Science Society.

Tuijl, T. van (2018). Keeping an eye on interruption: using pupil dilation to predict suitable moments for task-switching. Unpublished Master Thesis.

Shaposhnik, H., Borst, J.P., & Taatgen, N.A. (2018). A task-independent online interruption management system based on eye-tracking data. Unpublished manuscript.

References

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The Problem State: A Cognitive Bottleneck in Multitasking. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *36*(2), 363–382.

Borst, J. P., Taatgen, N. A., & Rijn, H. Van. (2015). What Makes Interruptions Disruptive ? A Process-Model Account of the Effects of the Problem State Bottleneck on Task Interruption and Resumption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI 2015*. New York: ACM.

Iqbal, S. T., & Bailey, B. P. (2005). Investigating the effectiveness of mental workload as a predictor of opportune moments for interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI 2005* (pp. 1489–1492). incollection, New York: ACM Press.

Katidioti, I., Borst, J. P., Haan, D. J. B. De, Pepping, T., Vugt, M. K. Van, & Taatgen, N. A. (2016). Interrupted by your Pupil : An Interruption Management System based on Pupil Dilation. *International Journal of Human–Computer Interaction*, *32*(10), 791–801.

Katidioti, I., Borst, J. P., & Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of Experimental Psychology: Applied*, *20*(4), 380–396.

Monk, C. A., Boehm-Davis, D. A., & Trafton, J. G. (2004). Recovering from interruptions: Implications for driver distraction research. *Human Factors*, *46*, 650–663.

Nijboer, M., Borst, J., van Rijn, H., & Taatgen, N. (2016). Contrasting single and multi-component working-memory systems in dual tasking. *Cognitive Psychology*, *86*, 1–26.