Technical University of Denmark

DTU

Assessing Levels of Attention Using Low Cost Eye Tracking

Bækgaard, Per; Petersen, Michael Kai; Larsen, Jakob Eg

Link back to DTU Orbit

DTU Library
Technical Information Center of Denmark

Assessing Levels of Attention Using Low Cost Eye Tracking

Per Bækgaard[(✉)], Michael Kai Petersen, and Jakob Eg Larsen

Cognitive Systems, Department of Applied Mathematics and Computer Science,
Technical University of Denmark, Building 321, 2800 Kongens Lyngby, Denmark
{pgba,mkai,jaeg}@dtu.dk

Abstract. The emergence of mobile eye trackers embedded in next generation smartphones or VR displays will make it possible to trace not only what objects we look at but also the level of attention in a given situation. Exploring whether we can quantify the engagement of a user interacting with a laptop, we apply mobile eye tracking in an in-depth study over 2 weeks with nearly 10.000 observations to assess pupil size changes, related to attentional aspects of alertness, orientation and conflict resolution. Visually presenting conflicting cues and targets we hypothesize that it's feasible to measure the allocated effort when responding to confusing stimuli. Although such experiments are normally carried out in a lab, we have initial indications that we are able to differentiate between sustained alertness and complex decision making even with low cost eye tracking "in the wild". From a quantified self perspective of individual behavioural adaptation, the correlations between the pupil size and the task dependent reaction time and error rates may longer term provide a foundation for modifying smartphone content and interaction to the users perceived level of attention.

Keywords: Eye tracking · Attention network

1 Introduction

Low cost eye trackers which can be embedded in next generation smartphones will enable design of cognitive interfaces that adapt to the users perceived level of attention. Even when "in the wild", and no longer constrained to fixed lab setups, mobile eye tracking provides novel opportunities for continuous self-tracking of our ability to perform a variety of tasks across a number of different contexts.

Interacting with a smartphone screen requires attention which in turn involves different networks in the brain related to alertness, spatial orientation and conflict resolution [20]. These aspects can be separated by flanker-type of experiments with differently cued, sometimes conflicting, prompts. Dependent on whether the task involves fixating the eyes on an unexpected part of the screen, or resolving the direction of an arrow surrounded by distracting stimuli, different parts of the attention network will be activated, in turn resulting in varying reaction times [7].

The dilation and constriction of the pupil is not only triggered by changes in light and fixation but reflect fluctuations in arousal networks in the brain [13], which from a quantified self perspective may enable us to assess whether we are sufficiently concentrated when we interact with the screens of smartphones or laptops, carrying out our daily tasks. Likewise the pupil size increases when we face an unexpected uncertainty [1], physically apply force by flexing muscles, or motivationally have to decide on whether the outcome of a task justifies the required effort [23]. Thus, when we perform specific actions, the cognitive load involved can be estimated using eye tracking. The pupil dilates if the task requires a shift from a sustained tonic alertness and orientation to more complex decision making, in turn triggering a phasic component caused by the release of norepinephrine neurotransmitters in the brain [2,8], which may reflect both the increased energization as well as the unexpected uncertainty related to the task [1].

Whereas these results have typically been obtained under controlled lab conditions, we explore in the present study the feasibility of assessing a users level of attention "in the wild" using mobile eye tracking.

2   Method

2.1   Experimental Procedure

This longitudinal study was performed repeatedly over the course of two weeks in September-October 2015. Two male right-handed subjects, A and B, (of average age 56) each performed a session very similar to the Attention Network Test (ANT) [7] approximately twice every weekday, resulting in 16 resp. 17 complete datasets, totaling 9.504 individual reaction time tests. The experiment ran "in the wild" in typical office environments off a conventional MacBook Pro 13" (2013 model with Retina screen) that had an Eye Tribe Eye Tracker connected to it. The ANT used here is implemented in PsychoPy [18] and is available on github [4]. Simultaneously, eye tracking data is recorded at 60 Hz and timestamped for synchronization through the Eye Tracker API [21] via the PeyeTribe [3] interface.

Before the actual experimental procedure starts, a calibration of the Eye Tracker is performed. The experiment contains an initial trial run that the user may select to abort, after which 3 rounds of $2 \cdot 48$ conditioned reaction time tests follows (Fig. 1); each test is conditioned on one of 3 targets: *Incongruent, Neutral* or *Congruent* and on 4 cues: *No Cue, Center Cue, Double Cue* or *Spatial Cue.* At the start of each test, a fixation cross appears, and after a random delay of 0.4–1.6 s the user is presented to a cue (when present for the particular condition). 0.5 s later the target appears, either with incongruent, neutral or congruent flankers. The user is instructed to hit a button on the left or right side of the keyboard with his left or right hand depending on the direction of the central arrow of the target, which appeared above or below the initial centred fixation cross. Half the targets appear above and half below the fixation cross, and left/right pointing central arrows also appear evenly distributed.
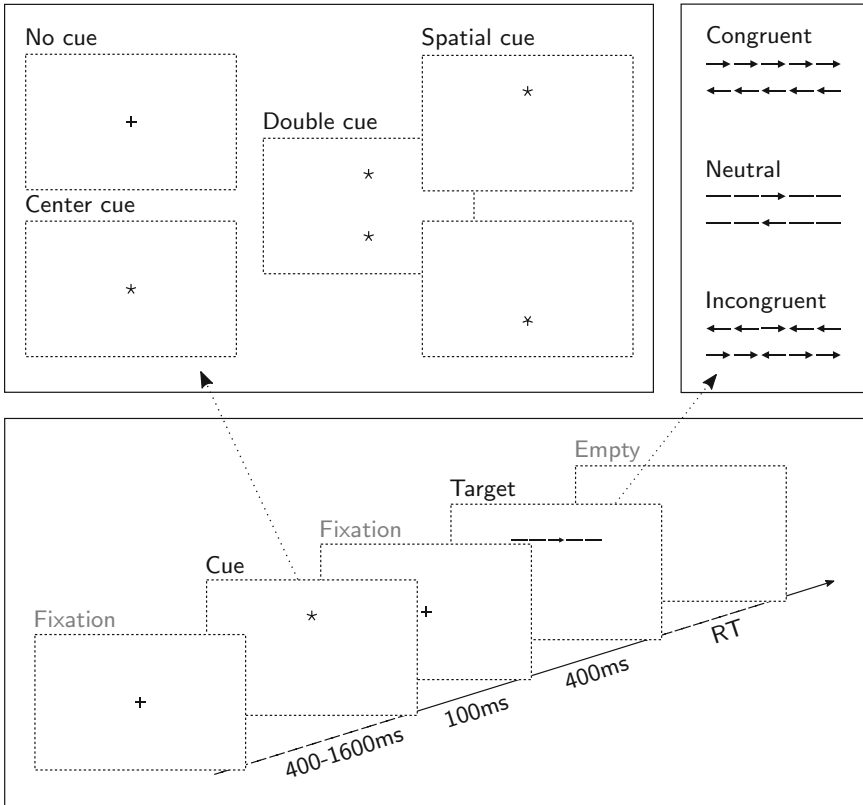
Fig. 1. This Attention Network Test procedure used here: Every 4 s, a cue (either of 4 conditions (Top, Left)) precedes a target (either of 3 congruency conditions (Top, Right)), to which the participant responds by pressing a key according to the central arrow. The reaction time differences between cue- and congruency conditions form the basis for calculating the latencies of the attention, orientation and conflict resolution networks.

The resulting reaction time "from target presentation to first registered key-press" is logged, together with the conditions of the individual tests, whether the user hit the correct left/right key or not, and a common timestamp. For further details on the ANT please see [7].

Each test takes approximately 4 s to perform. With 2·3 repetitions of all combinations of conditions, left/right arrows and above/below targets, this results in $6 \cdot 12 \cdot 2 \cdot 2 = 288$ single tests. The user has the option of a short break after each 96 performed tests. A typical session with calibration, experimental procedure and short breaks lasts approximately 25–30 min.

2.2   Analysis

The reaction times for each experiment, for which the user responded correctly within 1.7 s, are grouped and averaged over each of the 3 congruency and 4 cue conditions, and the Attention Network Test timings can be calculated as follows:

$$t_{\text{alertness}} = \overline{t_{\text{no cue}}} - \overline{t_{\text{double cue}}}$$
$$t_{\text{orientation}} = \overline{t_{\text{center cue}}} - \overline{t_{\text{spatial cue}}}$$
$$t_{\text{conflict resolution}} = \overline{t_{\text{incongruent}}} - \overline{t_{\text{congruent}}}$$

where

$$\overline{t_{\text{cond}}} = \frac{1}{N} \sum_{i|i=\text{cond}}^{N} t_i$$

Linear pupil size and inter-pupil distance data can be somewhat "noisy" when recording in office conditions. After epoch'ing to corresponding cue times for the individual tests, invalid/missing data from blink-affected periods are removed, and a Hampel [9] filter is therefore applied, using a centered window of ± 83 ms (shorter than a typical blink) and a limit of 3σ, to remove remaining outliers. Data is then downsampled to 100 ms resolution using a windowed averaging filter, and scaled proportionally to the value at epoch start (cue presentation), so that the resulting pupil dilations represent relative change[1] vs the pupil size at cue presentation. This last part was done to compensate for varying environmental luminosity changes and, to some degree, to offset any effect from immediately preceding reaction time test(s) and to compensate for accidental head position drift.

Time-locked averaging is then done by grouping data from similar conditions within each experiment, from which the group-mean relative pupil dilations can be derived.

At the same time, the inter-pupil distance is calculated, to ensure that pupil size changes would not be the accidental result of moving the head slightly during the experiment. Additionally, a "baseline" experiment has been performed, recording eye tracking data in a condition where no action can be taken by the user and when no arrow-heads are visible on the targets but otherwise presented in similar conditions, in order to rule out that the recorded pupil dilations would be the result of (small) luminosity changes caused by the presented cue and targets, or a result of slightly changing accommodation between the focus points of the cue and the target.

The inter-pupil distance variation was found to be significantly smaller (typically much less than 0.2 %) than the recorded pupil dilations, and the "baseline" experiment could not account for the recorded pupil dilations from the real experimental procedure either; it just showed the expected random variations.

---

[1] The data received from the eye tracker is uncalibrated and cannot easily be referenced to a metric measurement.

The data processing has been done with iPython [19] using the numpy [22], matplotlib [11], pandas [15], scipy [16] and scikit-learn [17] toolboxes.

3   Results

3.1   Attention Network Test Timings

Table 1 shows the aggregate Overall Mean Reaction- and Attention Network timings for each subject A and B, with estimates of the variation over the week. The figures are not significantly different from what is found in [7]; the MeanRT reported here is slightly higher than an estimated 512 ms in the reference, whereas the alertness, orientation and conflict resolution are slightly lower or similar to the 47 ms, 51 ms and 84 ms reported.

Table 1. Average Reaction- and Attention Network-Times over all correctly replied experiments for the two week period for either user (the variation over the period is given as estimated ± Sample Standard Deviation of the aggregate values), in milliseconds.

| Subject	MeanRT	Alert	Orient	Conflict
A	577 (±54)	27 (±21)	22 (±18)	85 (±16)
B	559 (±55)	35 (±17)	49 (±15)	81 (±17)

There are, however, behavioural variations in reaction time throughout the weeks. Figure 2 shows the variation of the derived ANT timings throughout the experimental period, and the relative error rate for each experiment. The variation appear to be statistically significant, as can be estimated from the standard error of the mean (the shaded area), and may reflect underlying states of varying levels of attention, fatigue and motivation.

To sum up the behavioural results, A shows a somewhat increasing trend in error rate related to the objective task performance, whereas B shows a diminishing difference between the three estimated measures of conflict resolution, spatial orientation and alertness reaction time.

3.2   Pupil Dilations

The group-mean relative linear pupil dilations for each of the 3 congruency conditions are illustrated in Fig. 3.

Pupil dilation responses are all epoch'ed to the cue (at time 0 ms) and target presentation (time 500 ms). A small and slow pupil dilation onset is seen <300 ms after cue presentation, followed by a larger response likely triggered by the target presentation, with an onset of approximately 700 ms and a peak approximately 1300 ms after target, with some variation between conditions, subject and eye.
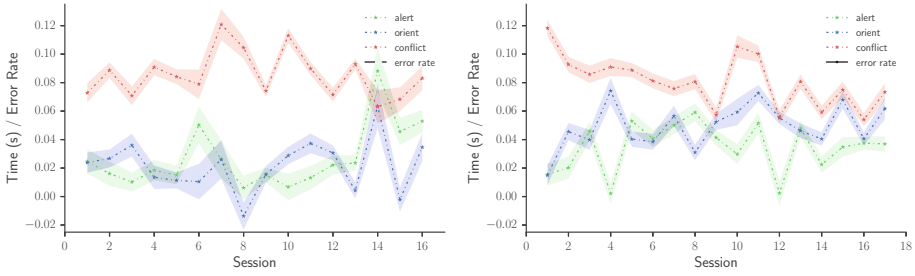
Fig. 2. Attention Network Timing over all sessions in the two week period. Conflict Resolution (Red) is slower than Alertness (Green) and Orientation (Blue). A (Left) shows an increasing error rate trend (Solid); Conflict Resolution for B gradually approaches the other latencies. Both A and B have large variations over time, pointing to varying levels of attention, fatigue and motivation. (Color figure online)
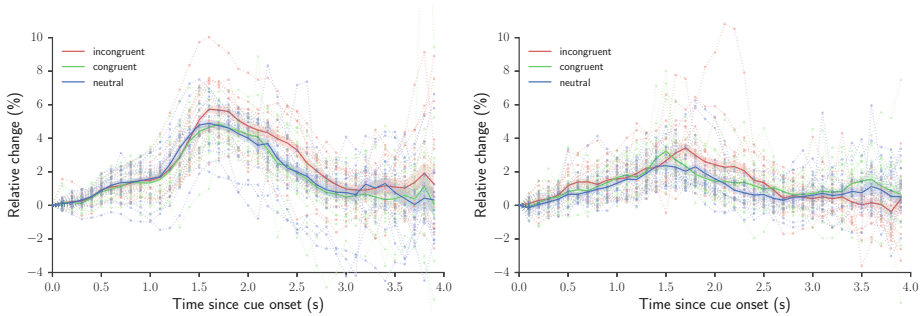

Fig. 3. Averaged left-eye pupil dilations for each session, coloured according to congruency (A (Left) and B). All-session average shown in bold, with the shaded area representing the standard error of the mean. The average incongruent (Red) pupil dilation is stronger than the others, indicating a higher cognitive load. (Color figure online)

Even though the experimental conditions are not directly comparable, [14] reported comparable peak latencies at 1400 ms after stimulus for a Stroop effect experiment. Our results are thus in line with these previous findings of pupil dilations, as well as with those reported in earlier processing load experiments [12] at approximately 900–1200 ms. The initial onset of the pupil dilation can occur even faster in some conditions [6,10] although generally onset and peak latencies appear to be within the 150–1400 ms.

The incongruent pupil dilation is larger than the more similar neutral and congruent dilations; there is however no such difference when comparing the 4 cue condition (not shown). The incongruent pupil dilation also has a tendency to appear slightly later (most easily visible for A), consistent with the longer reaction times for the inconsistent condition.

Figure 4 shows the (relative) pupil size BLUE vs the median value over a selected period that covers 48 reaction time tests, in this case for B, for two different experiments. Test-related pupil dilation responses, that occur every 4 s, are not immediately visible in this graph due to random noise and a relatively strong longer-periodic variation over 20–60 s[2]. The GREEN curve shows the relative variation of the inter-pupil distance, with variations an order of magnitude smaller than the pupil size changes.
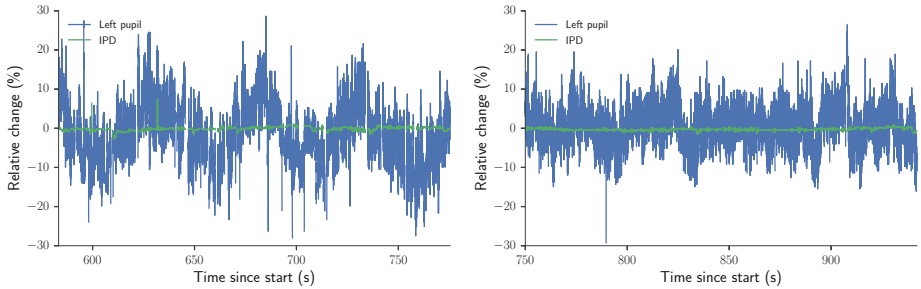
Fig. 4. Filtered pupil size plots; 48-test long sections of two experiments (B, left-eye). Relative inter-pupil distance (GREEN) indicates stable eye-to-screen distances. (Color figure online)

Figure 5 shows the area under the pupil dilation curve between 1.5–2.5 s after cue (1.0–2.0 s after target) for each experiment, serving as a very rough indicator of the relative cognitive load caused by the tests. From these, also a $\delta(\text{incon})$ can be calculated by subtracting the congruent value from the incongruent.
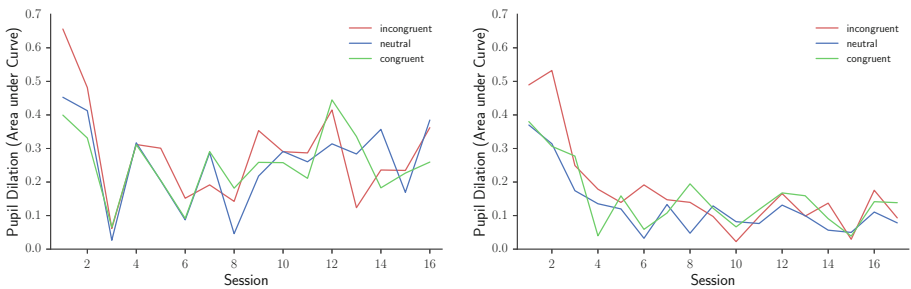
Fig. 5. Area under left-eye pupil dilation curves $[1.5, 2.5]$ s for each session, indicative of cognitive load, grouped after congruency. Both A (LEFT) and B show initial training effects; only A however shows an increasing trend in cognitive load for the remaining sessions. (Color figure online)

---

[2] A frequency domain analysis of the signal shows, however, a distinct peak at 0.25 Hz, as expected.

It is seen that both A and B have larger pupil dilation responses for the initial two experiments, after which the level is lower. For B it remains at lower levels, indicating a training effect. For A, the pattern is less clear, with possibly an increased load towards the end of the two week period.

3.3   Predicting Congruency Condition from Pupil Dilations

In order to verify how well previous pupil dilations allow predicting the class of congruency condition, a subset of the 3 within-experiment 96–average pupil dilation responses from each subject were ordered in each of the 6 possible permutations of the 3 congruency conditions. A neural-network type classifier was then trained to identify which of the 3 averaged pupil dilations were the incongruent.
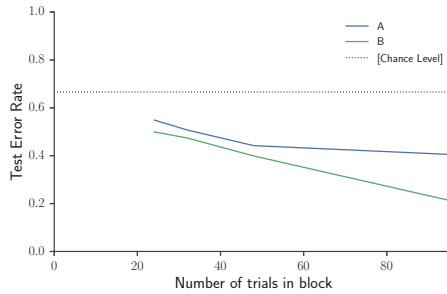
Fig. 6. Test error rates (0.9/0.1 train/test split) predicting averaged 3 s incongruent pupil dilations after cue vs number of averaged experimental tests. At 48 averaged experimental tests, the test error rate at 50 % is clearly below chance (66.6 %, dotted). (Color figure online)

Figure 6 shows the resulting test error rate vs. the number of averaged experimental tests, dividing the 96 equal-condition responses of each experiment into groups of 96, 48, 32 or 24 tests, and using a test/train split of 0.9/0.1. The performance is clearly above chance level (66.6 %), and approaches 80 % accuracy for B vs 60 % for A. Even at groups of 24 averaged experimental tests, the classifier operate above chance level, with continuing improved performance for larger groups for B, however only marginally improving performance for A.

3.4   Correlating Response Times and Pupil Reactions

Table 2 show the Pearson Correlation Coefficients for all combinations of Attention Network- and Reaction-Times, Pupil Dilation metrics and Time-of-Day for each subject, as it varies over the two week period. As the data sets are small (16 and 17 sets), caution is needed when judging the significance levels (p-values).

With some variation between subjects, pupil dilation responses appear correlated.

Table 2. Pearsons correlation coefficients between key metrics for A (Top) and B. A shows negative correlation between mean reaction time and error rate ("speed-accuracy tradeoff"). B (opposed to A) shows correlation between pupil dilations and error rate, possibly indicating a different response to varying levels of fatigue or motivation; additionally alertness (and partly orientation) may inversely correlate to pupil dilations. Both show expected correlations between pupil dilation metrics.

	Att.-Net/Reaction Time			Pupil Dilation					
	Orient	Conflict	μ(RT)	Incon	Neutral	Con	δ(Incon)	ToD	Errors
Att.-Net/Reaction Time									
Alert	0.112	−0.047	−0.189	−0.013	−0.131	−0.011	−0.008	0.061	−0.051
Orient		**−0.548**†	−0.468*	0.274	0.269	−0.020	0.402	0.132	0.270
Conflict			0.474*	−0.081	−0.149	0.035	−0.147	0.330	−0.416
μ(RT)				0.002	0.049	−0.069	0.068	0.237	**−0.635**†
Pupil Dilation									
Incon					**0.767**‡	**0.701**‡	**0.737**‡	0.062	−0.098
Neutral						**0.752**‡	0.362	0.222	0.109
Con							0.034	0.000	−0.018
δ(Incon)								0.087	−0.121
ToD									0.066

Two-tailed significance less than *7.5%, †**5%** and ‡**0.25%** marked.

	Att.-Net/Reaction Time			Pupil Dilation					
	Orient	Conflict	μ(RT)	Incon	Neutral	Con	δ(Incon)	ToD	Errors
Att.-Net/Reaction Time									
Alert	0.015	−0.107	0.438	**−0.499**†	**−0.534**†	−0.231	**−0.576**†	0.062	−0.358
Orient		−0.094	0.352	−0.474*	−0.407	**−0.559**†	−0.155	0.056	−0.386
Conflict			0.289	0.431	0.439	0.362	0.309	0.411	0.301
μ(RT)				−0.220	−0.286	−0.173	−0.173	0.481*	−0.400
Pupil Dilation									
Incon					**0.894**‡	**0.817**‡	**0.746**‡	−0.026	**0.725**‡
Neutral						**0.831**‡	**0.549**†	−0.184	**0.701**‡
Con							0.224	−0.020	**0.626**†
δ(Incon)								−0.021	**0.501**†
ToD									−0.215

Two-tailed significance less than *7.5%, †**5%** and ‡**0.25%** marked.

Subject A shows correlation between orientation and conflict resolution timings, which is however not seen at all for B. A also may have some correlation between mean reaction time and orientation resp conflict resolution timings, which are however again not quite as present with B.

Subject B shows correlation between alertness timing and both incongruent, neutral and δ(incon) pupil dilations, as well as correlation between orientation timing and congruent pupil dilations. These are not present for A, however. Also, there are indications of a correlation between the time of day and the mean reaction time; the experiments done on B were spread out over larger sections of the day than for A, which might explain why this is not seen for A.

[7] reported correlations between the conflict resolution timing and the mean reaction time over a large group of people. As such, the conditions are not similar to the within-person variation, but it might be worth pointing out that a similar correlation is partly present for A and cannot be ruled out for B.

4   Discussion

Using low cost portable eye tracking to measure the variations in pupil size, we have initial indications that we were able to differentiate and predict whether users were engaged in more complex decision making or merely maintaining a general alertness when interacting with a laptop, over nearly 10.000 tests. A parallel single-experiment study [5] repeating the experimental setup with nearly 10.000 additional tests over 18 more subjects, have confirmed that similar significant pupil response differences characterize the contrasts between incongruent versus neutral or congruent task conditions.

In the present study, we found a significant difference based on the left eye pupil size for the conflict resolution task in contrast to the attentional network components of alertness and re-orientation, but not between these two latter tasks. These results may reflect findings in other studies indicating that the phasic component in attention is predominantly triggered by tasks requiring a decision, whereas the tonic alertness may suffice for solving less demanding tasks like responding to visual cues or re-orienting attention to an unexpected part of the screen [2] as seen in the "baseline" experiment, where no decision needs to be made and no motor cortex activation takes place.

From a quantified self perspective of individual behaviour, using mobile eye tracking to assess levels of engagement, the relations between pupil size (a possible quantification of the cognitive load), and error rate/reaction time (a quantification of the objective task performance), indicate individual differences among the subjects' behavioural adaptation to the attentional tasks. Participant A is apparently coping with the cognitive load by trading off speed and accuracy to optimize performance, as indicated by the lack of correlation between pupil size and either of the performance related measures. However, for Participant B the correlation between pupil size and accuracy may suggest a behavior characterized by applying more effort to the task if the number of errors increase.

As we have in this study only used the pupil size as a measure of attention, even without considering the spatial density of fixations or the speed of saccadic eye movements that could entail further information, we suggest that mobile eye tracking may not only enable us to assess the effort required when undertaking a variety of tasks in an everyday context, but could also longer term provide a foundation for continuously adapting the content and interaction with smartphones and laptops based on our perceived level of attention.

References

1. Ang, Y.S., Manohar, S., Apps, M.A.J.: Commentary: noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. Front. Behav. Neurosci. **9**, 310 (2015). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4644795/pdf/fnbeh-09-00310.pdf, http://journal.frontiersin.org/Article/10.3389/fnbeh.2015.00310/abstract

2. Aston-Jones, G., Cohen, J.D.: An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. Ann. Rev. Neurosci. **28**(1), 403–450 (2005). http://www.annualreviews.org/doi/abs/10.1146/annurev.neuro.28.061604.135709

3. Bækgaard, P.: Simple python interface to the Eye Tribe eye tracker (2015). https://github.com/baekgaard/peyetribe/

4. Bækgaard, P.: Attention Network Test implemented in PsychoPy (2016). https://github.com/baekgaard/ant

5. Baekgaard, P., Petersen, M.K., Larsen, J.E.: Differentiating attentional network components using mobile eye tracking (in preparation)

6. Beatty, J.: Task-evoked pupillary responses, processing load, and the structure of processing resources (1982)

7. Fan, J., McCandliss, B.D., Sommer, T., Raz, A., Posner, M.I.: Testing the efficiency and independence of attentional networks. J. Cogn. Neurosci. **14**(3), 340–347 (2002). http://www.mitpressjournals.org//abs/10.1162/089892902317361886

8. Gabay, S., Pertzov, Y., Henik, A.: Orienting of attention, pupil size, and the norepinephrine system. Attention Percept. Psychophysics **73**(1), 123–129 (2011). http://www.ncbi.nlm.nih.gov/pubmed/21258914

9. Hampel, F.R.: The influence curve and its role in robust estimation. J. Am. Stat. Assoc. **69**(346), 383–393 (1974). http://www.tandfonline.com//abs/10.1080/01621459.1974.10482962

10. Holmqvist, K.: Eye Tracking: A Comprehensive Guide to Methods and Measures. Oxford University Press, Oxford (2011)

11. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**(3), 99–104 (2007)

12. Hyönä, J., Tommola, J., Alaja, A.M.: Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. Q. J. Exp. Psychology Sect. A **48**(3), 598–612 (1995). http://www.tandfonline.com//abs/10.1080/14640749508401407

13. Joshi, S., Li, Y., Kalwani, R.M., Gold, J.I.: Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. Neuron **89**(1), 221–234 (2016)

14. Laeng, B., Ørbo, M., Holmlund, T., Miozzo, M.: Pupillary stroop effects. Cogn. Process. **12**(1), 13–21 (2011)

15. McKinney, W.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference 1697900(Scipy), pp. 51-56 (2010). http://conference.scipy.org/proceedings/scipy2010/mckinney.html

16. Oliphant, T.E.: SciPy: open source scientific tools for python. Comput. Sci. Eng. **9**, 10–20 (2007). http://www.scipy.org/

17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2012). http://dl.acm.org/citation.cfm?id=2078195, http://arxiv.org/abs/1201.0490

18. Peirce, J.W.: PsychoPy-psychophysics software in python. J. Neurosci. Methods **162**(1–2), 8–13 (2007). http://dx.org/10.1016/j.jneumeth.2006.11.017

19. Pérez, F., Granger, B.E.: IPython: a system for interactive scientific computing. Comput. Sci. Eng. **9**(3), 21–29 (2007). http://ipython.org

20. Posner, M.I.: Attentional networks and consciousness. Front. Psychol. **3**, 1–4 (2012). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3298960/

21. The Eye Tribe: The Eye Tribe API Reference. http://dev.theeyetribe.com/api/

22. Van Der Walt, S., Colbert, S.C., Varoquaux, G.: The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22–30 (2011)

23. Varazzani, C., San-Galli, A., Gilardeau, S., Bouret, S.: Noradrenaline and dopamine neurons in the reward/effort trade-off: a direct electrophysiological comparison in behaving monkeys. J. Neurosci. **35**(20), 7866–7877 (2015). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4644795/pdf/fnbeh-09-00310.pdf, http://www.jneurosci.org/cgi//10.1523/JNEUROSCI.0454-15.2015