Towards Robust Binary Communication Using Pupil Dilation by Mental Arithmetic

IMMO SCHUETZ, Justus-Liebig University Giessen, Germany and Chemnitz University of Technology, Germany

JULIA TROJANEK, Chemnitz University of Technology, Germany

WOLFGANG EINHÄUSER, Chemnitz University of Technology, Germany

The pupil reliably dilates with increases in cognitive load, e.g., when performing mental multiplication. This mechanism has previously been used for communication with Locked-In-Syndrome patients [37], providing a proof-of-principle without focusing on communication speed. We investigated a similar paradigm in a larger sample of healthy participants (N=24) using three different eye trackers, including a sub-$100 commercial solution. Using the previously suggested training-free approach, we reliably decoded binary (yes/no) responses with comparable performance in all devices. To address whether throughput can be increased without sacrificing accuracy, we explored different classification models. Using a support-vector machine (SVM), we reached useful decoding with as little as 5 s of data. Our results confirm that mental arithmetic-based pupil dilation is a viable communication solution and suggests that bitrates of high-end brain-computer interfaces (BCI) can be reached at substantially lower cost.

CCS Concepts: • **Human-centered computing** → **Accessibility technologies**; **Empirical studies in HCI**; *Interaction techniques*.

Additional Key Words and Phrases: pupil dilation, pupillometry, brain-computer interface, cognitive load, mental arithmetic

1 INTRODUCTION & RELATED WORK

In patients whose motor capabilities are severely limited, such as Amyotrophic Lateral Sclerosis patients, the ocular and pupillary muscles often remain under volitional control even when the disease has progressed far. This makes eye tracking and pupillometry a potentially powerful method to provide a brain-computer interface (BCI) for communication [25]. Unlike methods that operate directly on brain activity, like electroencephalography (EEG, e.g. [4]) or functional near infrared spectroscopy (fNIRS, e.g. [5]), eye tracking and pupillometry can be conducted with inexpensive off-the-shelf equipment and without direct contact to the user, allowing for a potentially wider range of use cases. Besides the well-known Pupillary Light Reflex (PLR), changes in pupil size reflect a multitude of task-related and cognitive influences (see [3, 17, 27] for reviews). Various approaches to achieve communication using volitional control of pupil size have been suggested. For example, modulations of the PLR by oscillatory changes in brightness can be used to decode one of multiple response alternatives simply by shifts of covert attention to an oscillating target [29, 30, 33]. Pupil size modulation by mental imagery of a bright or dark scene [24] was explored for binary communication [9], and the link between accommodation and pupil size ("near triad", e.g. [22]) was recently exploited in a similar way

[34]. Other recent work suggests that voluntary pupil control can be trained through biofeedback techniques based on affective mental imagery [13–15], and this method has also shown promise as an input channel for HCI applications [16, 39, 40]. Stoll and colleagues [37] proposed a communication method that exploits the robust and well-established relation between pupil dilation and cognitive load [2, 3, 20, 21], specifically load induced by mental arithmetic, which has already been described at the end of the nineteenth century [1, 6, 19, 23]: Participants performed a mental multiplication task in one out of two separate time intervals, corresponding to either a "yes" or "no" response, and responses were successfully decoded from the difference in pupil dilation between both intervals. Importantly, this method required neither training on the task nor specific adjustments of the analysis to the individual and was successfully applied in locked-in-syndrome (LIS) patients. Since the goal of Stoll et al. (2013) [37] was a proof-of-principle that pupil-based communication is possible in this patient group, where communication at virtually any speed is beneficial, these authors were not concerned with optimizing throughput. While [37] clearly demonstrated the viability of the mental arithmetic paradigm for communication, it used a medical grade wearable eye tracker [36, 38] and tested a relatively small sample of N=6 healthy control participants, mostly to optimize their parameter settings for later use in patients. Using more affordable eye tracking hardware would enable easy accessibility and allow for broader clinical or in-home use of a pupil-based BCI. In the present study, we evaluate a similar task as [37] in a sample of N=24 healthy participants and compare the same medical eye tracker (EyeSeeCam) to a low-cost commercial (EyeTribe) as well as a research-grade device (Eyelink 1000+). Moreover, we compare different classification strategies to the method used in [37], with the aim of improving decoding accuracy and communication throughput.

2 MENTAL ARITHMETIC TASK

Participants. A total of N=24 participants (10 female, 14 male; ages 20-33 years) took part in the study. Volunteers had normal or corrected-to-normal vision, no history of vision or neuromuscular disorders, provided written informed consent and were paid 6€/h for their participation.

Experimental Setup & Paradigm. Participants performed a mental arithmetic task while their eye movements and pupil size were recorded using one of three eye trackers: Eyelink 1000+ (EL; SR-Research, Ottawa, ON, Canada) at 1000 Hz, EyeSeeCam Sci (ESC; EyeSeeTec GmbH, Munich, Germany; [36, 38]) at 220 Hz, and EyeTribe (ET; TheEyeTribe, Copenhagen, Denmark) at 60 Hz sample rate. The order of devices and sessions was counterbalanced across participants. Only one eye tracker was active at any given time to avoid interference from different IR illuminators. Participants placed their chin on a chin rest (distance to screen: 80 cm) and fixated a central cross during trials to keep effects of pupil foreshortening constant [10, 18]. A trial began with the presentation of one of seven binary questions on an LCD monitor until key press. Questions were chosen to have an objectively correct answer based on situative knowledge (e.g., "Is today Monday?") or information collected from the participant (e.g., "Is your name Michael?") and were always presented twice, once worded with "yes" as the correct response and once with "no" (14 trials per participant). After the key press, a response (e.g., "yes") was presented for 10 s together with a multiplication task (first response interval, Fig. 1, left), followed by the alternative response ("no") and a different multiplication for 10 s (second response interval). The order of questions (trials) and response alternatives within a trial were counterbalanced. Participants were instructed to perform the mental arithmetic task in the interval showing the correct response, and perform no task besides fixation in the other interval.

Data Processing. The experiment was built in OpenSesame [31] and PyGaze [7]. Custom PyGaze code was written to record data from the EyeSeeCam device. Data from the left eye were analyzed, except for one participant where only
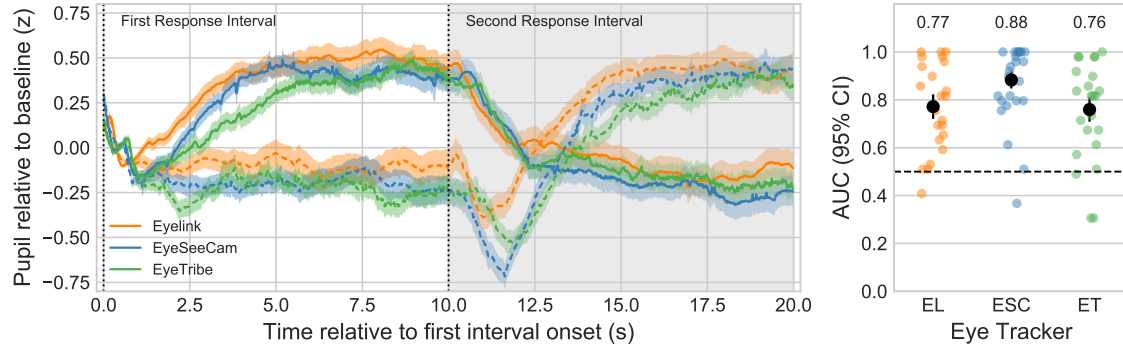
Fig. 1. *Left*: Average pupil size change (z-scored; shading: ± 1 SEM) when participants performed mental arithmetic in the first (solid lines) and second (dashed lines) response interval. Dotted vertical lines indicate response period onsets. *Right*: Small circles: Individual Area under the ROC curve (AUC) when decoding responses using the slope method used in [37]. Large circles and labels: AUC when decoding using all available responses for each device (using the method in [37]; error bars: 95% CI, dashed line: chance level).

data from the right eye were available and thus used. Eye blinks including 100 ms of data before and after the blink were removed and resulting gaps were linearly interpolated. Each experimental session was separately z-standardized (zero mean, unit standard deviation) to harmonize pupil size measures reported by different devices and match individual variation in baseline pupil size. All data were then resampled to a sample rate of 60 Hz using linear interpolation to match the slowest eye tracking device (ET)[1]. Pupil traces of each trial and individual response interval were extracted relative to the onset of the first response interval, and subtractive baseline correction was applied based on the first second of data [28]. Pupil analyses were performed in Python (v. 3.7), ROC statistics in R (v. 4.0.3) with the pROC package [35]. Uncorrected p-values are reported throughout, but all results that are given as significant at a 0.05 or 0.001 α-level remain significant at this level after Bonferroni-Holm adjustment. Data and code are available online [2].

3 RESULTS

Eye Tracker Comparison. As a first step, we replicated the analysis of [37] without any adjustment of parameters or fitting. We computed the receiver operator characteristic (ROC) curve for separating the interval containing the correct response option from the interval with the incorrect response option based on the slope from t=1.5s to t=5s of each interval. We quantified the sensitivity for discriminating the response options by the area under the ROC curve (AUC). To test statistically whether an individual's AUC exceeded chance (0.5) and to compare AUCs between devices, we used DeLong's test [8, 35]. Numerically, 23/24 (EL, ESC) or 21/24 (ET) participants had AUCs above chance, with 12 (EL), 13 (ET) and 19 (ESC) reaching significant above-chance decoding on an individual level (Fig. 1, right). To compare the eye tracking devices, we first aggregated data across all participants and computed a single ROC curve per device (black circles in Fig. 1). Decoding was significantly better for ESC (AUC = 0.88) than for the other two devices (EL: AUC = 0.77; z = 3.50; p < .001 and ET: AUC = 0.76; z = 3.90; p < .001), which were not significantly different from each other (z = 0.33; p = .738). Interestingly, decoding performance on the aggregate data was in the same range as the average over individual decoding performances, indicating high consistency across participants.

---

[1]To rule out low sampling rate as a major influence on classification, we repeated the same analyses with all data upsampled to 1 kHz to match Eyelink, with qualitatively similar results. Both analyses can be found on the project's Github page.
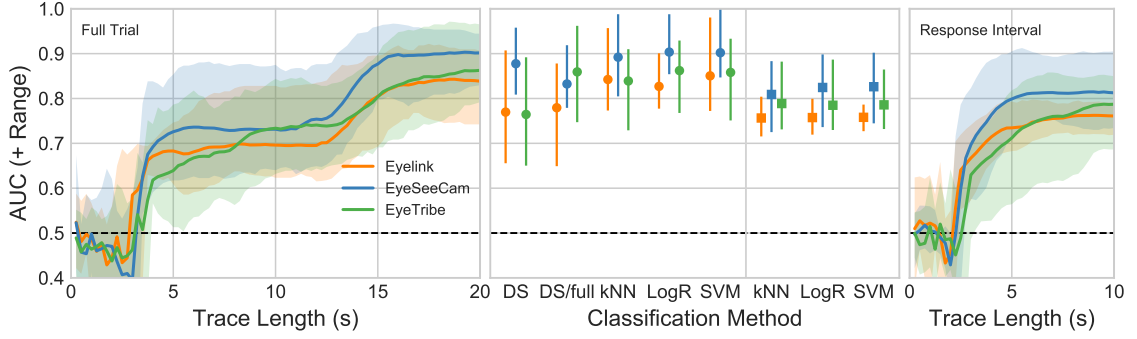[2]https://github.com/ischtz/emics2021-mentarith-bci

Fig. 2. Decoding Analysis. *Left:* AUC of the best-performing SVM model when trained and tested on time windows in steps of 250 ms (7-fold CV). *Middle:* Average cross-validation performance (AUC) of different algorithms in trial decoding (left; circles) and response interval decoding (right; squares). DS: Difference slope method - timing as in [37]; DS/full: DS method applied to full trial duration. *Right:* Best model (SVM) AUC when trained and tested on separate response intervals, 250 ms intervals. Error bars / shading: AUC range in 7-fold CV. Black dashed line: chance level.

Trial Decoding. We tested how well three standard classification algorithms (Logistic Regression, k-Nearest-Neighbor and SVM) could decode the correct response interval (first or second) in a trial. The aggregate data contained 336 (24×14) trials × 1,200 pupil samples (10 s per response period at 60 Hz), and all 1,200 samples were included as model features. We used a nested cross-validation (CV) strategy, with a 7-fold outer CV to split trials into a training (288 trials) and test set (48), and an 8-fold inner CV (252/36) for hyperparameter tuning using grid search. We compared the best models to the difference slope (*DS*, cf Fig. 2) method as described in [37]. For a fairer comparison with the trial-based classifiers, we also calculated the DS measure using the full 20 s trial trace for slope fitting (*DS/full*). Fig. 2 (circles) shows average CV performance for all models. Final models were then fit to all data using optimal hyperparameters (LogR: C = 0.0001; kNN: k = 50; SVM: C = 1, γ = 0.0001, rbf kernel) and overall ROC curves and AUC were computed per device and compared to the approach in [37]. All classifiers showed better decoding than DS for EL and ET data (all z > 3.29, all p < .001), whereas differences were not significant in ESC after correction (all z < 2.44, all p > .015). DS/full was outperformed by all models on EL and ESC (all z > 3.73, all p < .001), but not on ET (all z < 2.60, all p > .001). SVM outperformed most models across devices significantly (all z > 2.93, all p < .001) except for DS on ESC (z = 2.43, p = .015), DS/full on ET (z = 2.60, p = .009) and kNN on ET (z = 1.97, p = .049) and EL (z = 0.07, p = .941). To investigate how decoding performance varied with data availability, we further evaluated the SVM model for pupil traces of varying length, starting at first interval onset and increasing in 250 ms steps (Fig. 2, left) using 7-fold CV. Performance rapidly increased around 3 s after onset, then further improved when samples from the corresponding period of the second interval were included ($\geq$ 12.5 s).

Interval Decoding. We further explored task decoding of individual response intervals (calculate or rest) using pupil data from the individual 10 s response intervals (two intervals per trial; 672 intervals × 1,200 pupil samples in total). Models and CV strategy were identical to trial decoding. On this dataset, performance was lower than with full trials but still well above chance (Fig. 2, squares). SVM and kNN achieved comparable performance (all z < 2.36, p > .018), generally better than LogR (all z > 3.20, p < .002 except kNN/EL: z = 1.33, p = .182). As with trial traces, we selected SVM for follow-up trace length analysis (Fig. 2, right). Performance again steeply increased after the first 2.5 s of data, reaching asymptotic performance after 5 s.

4 DISCUSSION

We demonstrate that robust communication by controlling one's pupil through mental arithmetic can be achieved with a variety of devices, including a sub-$100 eye tracker. With a larger set of healthy users, we replicated that the method suggested by Stoll and colleagues [37] achieves good performance, in particular for the head-mounted device (ESC) also used in their study. Nonetheless, supervised learning methods further improve classification performance. Our data suggest that a single 5-s response interval (calculate or rest) would be sufficient for decoding, allowing an at least four-fold speed up compared to [37]. It is conceiveable that decoding can be further improved in patients and healthy users by applying time series-aware models, continuous cognitive load measures such as the Index of Cognitive Activity (ICA, [26]) and (Low/High-) Index of Pupillary Activity (IPA / LHIPA, [11, 12]), or neural network models, provided sufficient data are available. Importantly, our method does not require any adjustment to the individual user (aggregate data was used), though further improvements may be possible if more trials per user are recorded and classifiers are trained on an individual level. While the EyeTribe device used here is no longer available commercially, the results presented are based on a sampling rate of 60 Hz and are likely to generalize to similar low-end devices. Additionally, robust pupillometry based on webcam or smart phone camera data is now within reach (e.g., [32, 41]), which could further expand the possible fields of application for HCI approaches that exploit volitional pupil control based on mental arithmetic.

REFERENCES

[1] Sylvia Ahern and Jackson Beatty. 1979. Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science* 205, 4412 (1979), 1289–1292.

[2] Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin* 91, 2 (1982), 276.

[3] Jackson Beatty, Brennis Lucero-Wagoner, et al. 2000. The pupillary system. *Handbook of psychophysiology* 2, 142-162 (2000).

[4] Niels Birbaumer, Nimr Ghanayim, Thilo Hinterberger, Iver Iversen, Boris Kotchoubey, Andrea Kübler, Juri Perelmouter, Edward Taub, and Herta Flor. 1999. A spelling device for the paralysed. *Nature* 398, 6725 (1999), 297–298.

[5] Seyyed Bahram Borgheai, John McLinden, Alyssa Hillary Zisk, Sarah Ismail Hosni, Roohollah Jafari Deligani, Mohammadreza Abtahi, Kunal Mankodiya, and Yalda Shahriari. 2020. Enhancing communication for people in late-stage ALS using an fNIRS-based BCI system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28, 5 (2020), 1198–1207.

[6] Siyuan Chen and Julien Epps. 2014. Using task-induced pupil diameter and blink rate to infer cognitive load. *Human–Computer Interaction* 29, 4 (2014), 390–413.

[7] Edwin S Dalmaijer, Sebastiaan Mathôt, and Stefan Van der Stigchel. 2014. PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior research methods* 46, 4 (2014), 913–921.

[8] Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* (1988), 837–845.

[9] Victoria Anne Diedrichs. 2015. *Leveraging Pupillometry and Luminance-Based Mental Imagery for a Novel Mode of Communication.* Ph.D. Dissertation. Temple University. Libraries.

[10] Jan Drewes, Guillaume S Masson, and Anna Montagnini. 2012. Shifts in reported gaze position due to changes in pupil size: Ground truth and compensation. In *Proceedings of the symposium on eye tracking research and applications.* 209–212.

[11] Andrew T Duchowski, Krzysztof Krejtz, Nina A Gehrer, Tanya Bafna, and Per Bækgaard. 2020. The Low/High Index of Pupillary Activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–12.

[12] 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.* 1–13.

[13] Jan Ehlers, Nikola Bubalo, Markus Loose, and Anke Huckauf. 2015. Towards Voluntary Pupil Control-Training Affective Strategies?. In *International Conference on Physiological Computing Systems*, Vol. 2. SCITEPRESS, 5–12.

[14] Jan Ehlers, Christoph Strauch, Juliane Georgi, and Anke Huckauf. 2016. Pupil size changes as an active information channel for biofeedback applications. *Applied psychophysiology and biofeedback* 41, 3 (2016), 331–339.

[15] Jan Ehlers, Christoph Strauch, and Anke Huckauf. 2018. Training facilitates cognitive control on pupil dilation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications.* 1–5.

[16] Jan Ehlers, Christoph Strauch, and Anke Huckauf. 2018. A view to a click: pupil size changes as input command in eyes-only human-computer interaction. *International Journal of Human-Computer Studies* 119 (2018), 28–34.

[17] Wolfgang Einhäuser. 2017. The pupil as marker of cognitive processes. In *Computational and cognitive neuroscience of vision*. Springer, 141–169.

[18] Taylor R Hayes and Alexander A Petrov. 2016. Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods* 48, 2 (2016), 510–527.

[19] W. Heinrich. 1986. Die Aufmerksamkeit und die Funktion der Sinnesorgane: Erster Beitrag. *Zeitschrift für Psychologie und Physiologie der Sinnesorgane* 9 (1986), 342–388.

[20] Eckhard H Hess and James M Polt. 1964. Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 3611 (1964), 1190–1192.

[21] Daniel Kahneman and Jackson Beatty. 1966. Pupil diameter and load on memory. *Science* 154, 3756 (1966), 1583–1585.

[22] Sanjeev Kasthurirangan and Adrian Glasser. 2005. Characteristics of pupil responses during far-to-near and near-to-far accommodation. *Ophthalmic and Physiological Optics* 25, 4 (2005), 328–339.

[23] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 69–72.

[24] Bruno Laeng and Unni Sulutvedt. 2014. The eye pupil adjusts to imaginary light. *Psychological science* 25, 1 (2014), 188–197.

[25] Päivi Majaranta. 2011. *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies*. IGI Global.

[26] Sandra P Marshall. 2002. The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th conference on Human Factors and Power Plants*. IEEE, 7–7.

[27] Sebastiaan Mathôt. 2018. Pupillometry: Psychology, physiology, and function. *Journal of Cognition* 1, 1 (2018).

[28] Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior research methods* 50, 1 (2018), 94–106.

[29] Sebastiaan Mathôt, Jean-Baptiste Melmi, Lotje Van der Linden, and Stefan Van der Stigchel. 2015. The mind-writing pupil: near-perfect decoding of visual attention with pupillometry. *J Vis* 15, 12 (Sep 2015), 176. https://doi.org/10.1167/15.12.176

[30] Sebastiaan Mathôt, Jean-Baptiste Melmi, Lotje van der Linden, and Stefan Van der Stigchel. 2016. The Mind-Writing Pupil: A Human-Computer Interface Based on Decoding of Covert Attention through Pupillometry. *PLoS One* 11, 2 (2016), e0148805. https://doi.org/10.1371/journal.pone.0148805

[31] Sebastiaan Mathôt, Daniel Schreij, and Jan Theeuwes. 2012. OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior research methods* 44, 2 (2012), 314–324.

[32] Raffaele M Mazziotti, Fabio Carrara, Aurelia Viglione, Leonardo Lupori, Luca Lo Verde, Alessandro Benedetto, Giulia Ricci, Giulia Sagona, Giuseppe Amato, and Tommaso Pizzorusso. 2021. MEYE: Web-app for translational and real-time pupillometry. *bioRxiv* (2021).

[33] Marnix Naber, George A. Alvarez, and Ken Nakayama. 2013. Tracking the allocation of attention using human pupillary oscillations. *Front Psychol* 4 (2013), 919. https://doi.org/10.3389/fpsyg.2013.00919

[34] Francesco Ponzio, Andres Eduardo Lorenzo Villalobos, Luca Mesin, Claudio de'Sperati, and Silvestro Roatta. 2019. A human-computer interface based on the "voluntary" pupil accommodative response. *International Journal of Human-Computer Studies* 126 (2019), 53–63.

[35] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* 12, 1 (2011), 1–8.

[36] Erich Schneider, Thomas Villgrattner, Johannes Vockeroth, Klaus Bartl, Stefan Kohlbecher, Stanislavs Bardins, Heinz Ulbrich, and Thomas Brandt. 2009. EyeSeeCam: an eye movement-driven head camera for the examination of natural visual exploration. *Ann N Y Acad Sci* 1164 (May 2009), 461–467. https://doi.org/10.1111/j.1749-6632.2009.03858.x

[37] Josef Stoll, Camille Chatelle, Olivia Carter, Christof Koch, Steven Laureys, and Wolfgang Einhäuser. 2013. Pupil responses allow communication in locked-in syndrome patients. *Curr Biol* 23, 15 (Aug 2013), R647–R648. https://doi.org/10.1016/j.cub.2013.06.011

[38] Josef Stoll, Stefan Kohlbecher, Svenja Marx, Erich Schneider, and Wolfgang Einhäuser. 2011. Mobile three dimensional gaze tracking. *Studies in health technology and informatics* 163 (2011), 616–622.

[39] Christoph Strauch, Jan Ehlers, and Anke Huckauf. 2017. Pupil-assisted target selection (pats). In *Ifip conference on human-computer interaction*. Springer, 297–312.

[40] Christoph Strauch, Lukas Greiter, and Anke Huckauf. 2017. Towards pupil-assisted target selection in natural settings: Introducing an on-screen keyboard. In *IFIP Conference on Human-Computer Interaction*. Springer, 534–543.

[41] Chatchai Wangwiwattana, Xinyi Ding, and Eric C Larson. 2018. Pupilnet, measuring task evoked pupillary response using commodity rgb tablet cameras: Comparison to mobile, infrared gaze trackers for inferring cognitive load. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–26.